Conference Dedication
Lotfi A. Zadeh, Ph.D.
February 4, 1921 – September 6, 2017
Foreword from the Program Chairs

The Society of Design and Process Science (SDPS) is proud to have hosted its 22\textsuperscript{nd} conference, on November 5 - 9, 2017, in the magic city of Birmingham, Alabama. The conference has a tradition of attracting exceptional talent and innovative work from academics, Nobel Laureates, industry leaders. This year was no exception.

We are proud to report that the conference had more than one hundred and fifty registrations, with attendees representing such diverse regions as the US, Germany, Brazil, England, Japan, Norway, India, Turkey, Italy, Korea, Taiwan, Singapore, and China. On behalf of the SDPS Board, Program Committee, session organizers, external reviewers, and our dedicated volunteers, we send a warm message of gratitude to all who have attended and helped us make this year's conference a great success.

As an international nonprofit all-volunteer, society of engineers and industry leaders, SDPS encourages Transformative Research and Education through Transdisciplinary Means\textsuperscript{TM}. This year the theme adopted by the conference was on the emerging trends and technologies in convergence solutions. Convergence is the integration of insights and approaches from historically distinct scientific and technological disciplines. The National Academy of Sciences states that Convergence comes as a result of the sharing of methods and ideas by chemists, physicists, computer scientists, engineers, mathematicians, and life scientists across multiple fields and industries. We thank all the paper authors, workshop participants, and keynote speakers for their contributions towards this important and significant area of research.
We are proud to have had many individuals lend their support to this year's conference. Among the list of notables, we recognize Ms. Kay Ivey (Governor of the State of Alabama), Mr. William Bell (Mayor of the City of Birmingham), Mr. Tony Petelos (Jefferson County Manager), and Dr. Frank Wilczek (MIT and Nobel Prize in Physics, 2004). Dr. Wilczek is also the winner of the SDPS 2017 Transformative medal.

It is with a profound sense of loss that we commemorate Dr. Lotfi A. Zadeh on his passing earlier this year. Dr. Zadeh was a great supporter of SDPS and we will deeply miss his guidance and contributions. We will also miss the contributions of SDPS Fellow and former program chair, great friend Hartmut Ehrig.

As the program chairpersons, we would like to express our deepest gratitude to the SDPS community for another successful year. More specifically, conference managers, Michael Lipscomb and Steven Fernandez, Registration chair David Robbins, Submission chair Deniz Kocak, and documentation chair Ting Zhang carried the heavy load behind the scenes. We are thankful for their efforts.

We look forward to seeing you next year.

Leon Jololian, PhD
Donna Malvey, PhD
David Robbins, PhD
Dear Members of the Society for Design and Process Science and colleagues:

Thank you very much for honoring me with the Transformative Achievement Award. In accepting it I will be joining a group of people I very much admire.

I’m sorry that I can’t be with you in person, but I’d like to share some thoughts I’ve had recently about the importance of imagination and wishful thinking in the creative process. They seem relevant to your theme of Convergence, and to share in its spirit.

In recent years, there have been several spectacular success stories in machine learning. Checkers, chess, and now the deep and ancient game Go have been conquered, in the sense that computers are now the best players. What does it teach us?

Perhaps ironically, one big takeaway is the value of a quintessentially human characteristic: imagination. At first that conclusion might sound strange. Do computers really have imagination? But, on deeper reflection, I think its compelling. To choose their moves, the computers spin out and examine many alternative possibilities, before settling on the most promising. And what is imagination, but the ability to consider what is not, but might be?

Many creative people, including many physicists, have testified to the value of imagination in their work. In one of his most famous quotes, Albert Einstein said: “Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world.”

The great quantum physicist Paul Dirac, when asked about how he made his epochal discoveries, replied: “I like to play about with equations, just looking for beautiful mathematical relations which maybe don’t have any physical meaning at all. Sometimes they do.”

And Richard Feynman said: “The game I play is a very interesting one. Its imagination, in a tight straitjacket.” The straitjacket, he goes on to explain, is that if your goal is to discover new things about physical reality, you have to respect known facts, accumulated over centuries of research.

Perhaps the most impressive testimonial to scientific imagination came from James Clerk Maxwell, the leading theoretician of electromagnetism, in his tribute to Michael Faraday, its leading experimenter. The self-taught Faraday had to rely on his visual imagination, because his mathematical training was sketchy. This led him to new ways of thinking, as
Maxwell described: Faraday, in his minds eye, saw lines of force traversing all space where the mathematicians saw centres of force attracting at a distance: Faraday saw a medium where they saw nothing but distance; Faraday sought the seat of the phenomena in real actions going on in a medium.

Later, Maxwell later converted Faraday's imaginative visions into wonderful new kinds of equations, which we still use today.

But imagination is only half the story. The process of imagining possibilities must be married to methods for evaluating their success.

In games, the final goal like winning or solving the puzzle is specified clearly. But in interesting games its not possible to see a clear path to that goal straight away. To make progress you must form more limited, tractable, intermediate goals, and aim for those. In other words, youve got to decide what to wish for.

Wishful thinking, in this sense, is an essential part of problem-solving. This becomes even more true when we move outside the context of games, where often there are no set rules that supply the definition of ultimate success.

Its very plausible, then, that an important step toward achievement, whether for machines or human beings, is to cultivate smart, systematic wishful thinking. New Year’s resolutions, business plans, and visionary “to do” lists embody that strategy.

The issue of goals arises most keenly at the highest levels of learning and problem-solving. Trying to make a scientific breakthrough, write a novel or compose a symphony, or create some other great work of art can be hard, frustrating work. Success is not guaranteed, and the tangible, economic rewards are usually modest. What keeps people going?

Some motivations are obvious. The dopamine rush that accompanies successful problem-solving can be its own reward. And successful problem-solving can earn the esteem of others another thing people find rewarding in itself. But that isnt the whole story.

For many of the greatest physicists, religion was a big motivator. Galileo, Newton, Faraday, and Maxwell were all deeply believing, if not entirely orthodox, Christians. Heres how John Maynard Keynes, the economist, described Isaac Newton, in his famous lecture “Newton, the Man”: He looked on the whole universe and all that is in it as a riddle, as a secret which could be read by applying pure thought to certain evidence, certain mystic clues which God had lain about the world to allow a sort of philosophers treasure hunt to the esoteric brotherhood. He believed that these clues were to be found partly in the evidence of the heavens and in the constitution of elements (and that is what gives the false suggestion of his being an experimental natural philosopher), but also partly in certain papers and traditions handed down by the brethren in an unbroken chain back to the original cryptic revelation in Babylonia. He regarded the universe as a cryptogram set by the Almighty.

And here is Maxwell again, delighting in his discoveries: The vast interplanetary and interstellar regions will no longer be regarded as waste places in the universe, which the Creator has not seen fit to fill with the symbols of the manifold order of His kingdom. We shall find them to be already full of this wonderful medium; so full, that no human power
can remove it from the smallest portion of space, or produce the slightest flaw in its infinite continuity.

These scientists were determined, as Stephen Hawking put it, to “know the mind of God.” That inspired them to work prodigiously hard.

Though Albert Einstein was not religious in any conventional sense, he, too, was driven by a passion to know. In his Autobiographical Notes, he recounts his fascination as a child when his father showed him a compass: “This experience made a deep and lasting impression upon me. Something deeply hidden had to be behind things.” And elsewhere he wrote movingly of his struggle to reach the general theory of relativity: “The years of searching in the dark for a truth that one feels but cannot express, the intense desire and the alternations of confidence and misgiving until one breaks through to clarity and understanding.”

Although I am no longer a believer, I am grateful for the training I received in Roman Catholicism, which taught me to see the world as having grandeur and hidden meaning. That vision helped inspire my wishful searching for grandeur and meaning in the physical world, which became habitual.

Finally, let me add that in deciding what to wish for, a feeling for beauty is an invaluable asset. Exposure to beautiful objects and sounds—art and music as well as beautiful ideas—can develop that sense.

It’s a wonderful thing and not, I think an accident, that we humans naturally take delight in imagination and wishful thinking. They help us to solve problems, and thrive. Teachers and managers should encourage them, and even the most practical, hard-headed engineers and scientists should cultivate them.

Sincerely,

Frank Wilczek
Starting in the Middle Ages in Europe, many villages had a “village commons” or a “village green.” The village commons is an open area of a village (often land that is unsuitable for farming or farmed land made available after crops have been harvested) where villagers contribute land, labor, money and other resources to maintain this area. Traditionally, the village commons was often a common grassland at the center of the village and was used for grazing, watering stock, etc. In addition, the village commons also served as a hub for meetings and social events, such as dances and ceremonies. During such meetings, villagers often exchanged information on effective practices, as well as ideas to enable resilient collaboration and community building. Villagers also used these meetings to contribute and barter tools and services. The village commons holds the space for the convergence of diverse ideas and practices into serviceable new solutions for the common good.

SDPS has been serving as a village common during the last 20 years, holding the space for the convergence of diverse disciplines. While it is heartwarming to witness the newly found awareness of huge effect of convergences such as between artificial intelligence and big data, it is also somewhat frightening to contemplate the potential consequences of such a convergence. We must remember that the intention of a village commons is for the common good. I hope that SDPS will actively encourage the contributions that are unique to the human race, especially ethical considerations in the continuing evolution of transdisciplinary convergence for transcendence.
Table of Contents

SESSION 1: INFORMATION THEORIES, SIMULATION AND TESTING TECHNOLOGIES

PROCESS DECOMPOSITION USING AXIOMATIC DESIGN THEORY ................................................... 1
Cengiz Toğay, Gül Tokdemir, Ali H. Doğru

METRICS FOR SOFTWARE RELIABILITY: AN INITIAL SYSTEMATIC MAPPING STUDY ..................... 8
Elisabetta Ronchieri, Marco Canaparo

A FORMAL PRODUCT-LINE ENGINEERING APPROACH FOR SCHEDULERS .............................. 15
Güner Orhan, Mehmet Akşit, Arend Rensink

SESSION 2: INTERNET OF THINGS

COMPONENT INTEGRATION THROUGH CONNECTOR SUPPORTED PROCESS MODELS ...................... 31
Anil Cetinkaya, M. Cagri Kaya, Ali H. Dogru

A VARIABILITY PERSPECTIVE FOR IOT HETEROGENEITY IN SMART CITIES .............................. 38
Selin Unal, Ali H. Dogru

CASE STUDY ON IMPROVING DATA ACCESSIBILITY AND REDUCING INTERFACE COMPLEXITY IN THE I2B2 WEB CLIENT .................................................................................................................. 45
Sai Saran Chintha, Leon Jololian

SESSION 3: DATA DRIVEN DEEP LEARNING IN MEDICINE

RADIAL FIBER ATROPHY: A NEW METRIC FOR TENSOR-BASED MORPHOMETRY .......................... 54
Jon R. Marstrander, Thomas Anthony, Veronica L. Powell, R. Glenn Brook, Mitchel D. Horton, Frank M. Skidmore

LUNG SEGMENTATION USING A CONVOLUTIONAL NEURAL NETWORK ........................................ 61
David Odaibo, Solomon Hatcher, Zheng Zhang, Chandra Yarlagadda, Frank Skidmore

WIRELESS AD HOC NETWORK MODELING AND ANALYSIS USING MAXIMUM INDEPENDENT SET ........................................................................................................................................... 65
Saleh Allehaibi, Serkan Güldal, Murat M. Tanik

SESSION 4: SEMANTIC TECHNOLOGIES AND APPLICATIONS

SOFTWARE ARCHITECTURES FOR SMART APPLICATIONS WHICH MERGE ONTOLOGICAL REASONING WITH BIG DATA ANALYTICS ................................................................. 75
Radmila Juric, Inhwa Kim

LEARNING SPACE MODEL WHICH ADDRESSES DIFFERENCES IN LEARNING ............................. 83
Eiman Almami, Radmila Juric, Mohammed Zaki Ahmed

DEVELOPING AN ONTOLOGICAL MODEL FOR ADDRESSING LIFESTYLE CHANGES FOR REVERSIBILITY OF DIABETES 2 .................................................................................................. 93
Mehdi Tarabi, Radmila Juric, Steinar Larsen
SESSION 5: BIG DATA AND DATA SCIENCE (IN A POST-BIG DATA ERA)

SOUND ANALYSIS AND MACHINE LEARNING IN NONINVASIVE CLASSIFICATION OF NEUROLOGICAL CONDITIONS ................................................................. 103
Khaled Elzaafarany, Gyanendra Kumar, Moustafa H. Aly, Arie Nakhmani

A MODEL OF WIKIPEDIA KNOWLEDGE COMMUNITIES: LEARNING FROM USER BEHAVIOR .............. 108
Chester Parrott, Doris Carver

THE FRAMEWORK FOR COMPARING BIG DATA TECHNOLOGIES IN PROCESSING LIVE AND USER GENERATED DATA .......................................................... 118
Joseph Desire Fotso, Radmila Juric

SESSION 6: DESIGN IN TRANSDISCIPLINARY RESEARCH - I

TRANSDISCIPLINARY APPROACH TO CONVERGENCE-BASED ENGINEERING MODELING WITH COMMUNICATION THEORY ............................................. 124
Darrell Fielder, Karthik Lingasubramanian, Murat Tanik

DESIGN AND EVALUATION OF STROKE REHABILITATION GAMES BASED ON SERIOUS GAME AND CLINICAL KNOWLEDGE ................................................. 132
Chien-Lung Hsu, Sheng-Hui Lu, Tzu-Wei Lin

A GENERALIZED APPROACH FOR PMT PERFORMANCE PREDICTION .................................................. 137
Fan Xiong, Ed Marquette

DESIGNING AND IMPLEMENTING AN EDUCATIONAL GAME FOR ACHIEVING LEARNING GOALS IN OPTOMETRY .......................................................... 142
August Aalde, Radmila Juric, Hanne-Mari Schiøtz Thorud, Vibeke Sundling

SESSION 7: INTERPROFESSIONAL EDUCATION AND VIRTUAL REALITY

COMPARING USABILITY OF BUSINESS RULE VALIDATION PROCESSES IN HEALTH INFORMATION SYSTEMS ................................................................. 150
Ramaraju Rudraraju, Nick Timkovich, Ryan Cantor

A MULTI-TIERED DISTRIBUTION STRATEGY FOR IMMERSIVE VIRTUAL REALITY CONTENT IN BLENDED DELIVERY INTERPROFESSIONAL HEALTHCARE EDUCATION ........................................... 158
David Robbins, John Galdo, Sarah Waller

USING VIRTUAL REALITY TO TEACH INTERPROFESSIONAL EDUCATION ........................................ 161
John A. Galdo, Ashley Parish, Rebecca Warr, David Robbins

ADDING ADVANCED FEATURES TO A DATABASE COURSEWARE .................................................... 165
Mario Guimaraes

INTERNATIONAL CONVERGENCE ON THE TRANSDISCIPLINARY RESOURCE CORRELATION PLATFORM ............................................................. 169
U. John Tanik, Sang C.Suh, Abdullah Arslan, Kaoning Ju, Hemanth Bandi
SESSION 8: SIGNAL AND IMAGE PROCESSING

AN EMPIRICAL STUDY ON THE MEASURABILITY OF MEDITATION .............................................. 181
Hong Lin, Steven Lawrence Fernandes, V. Rajinikanth

A CLOSER LOOK AT VIDEO BASED SIDE CHANNEL ATTACKS ON THE SMARTPHONE
USERS’ PIN .............................................................................................................................................. 189
Diksha Shukla, Vir V. Phoha

INVESTIGATION OF RAPID HARDWARE STUDIES ON ANALOG CIRCUITS FOR
DESIGNING FILTER SYSTEMS ................................................................................................................ 195
Ozgur Aksu

SESSION 9: AUTONOMOUS SYSTEMS AND MACHINE LEARNING: CONCEPTS, TECHNIQUES AND APPLICATIONS - I

TRAFFIC SIGNS AUTOMATIC RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK ........... 202
Rina Komatsu, Tad Gonsalves

FEYNMAN’S DRUMS: A STEM EXERCISE ON PERMUTATIONS ...................................................... 208
Michael Lipscomb, William Monroe, Karthik Lingasubramanian

MACHINE LEARNING FOR CAPACITIVE PROXIMITY SENSORS .................................................. 214
Badis Madani, Hosam Alagi, Björn Hein, Aurilla Aurelie Arntzen

A HIGHLY EFFECTIVE TRANSFER LEARNING FRAMEWORK OF DCNNs APPLIED TO
AUTOMATIC POLYP DETECTION IN COLONOSCOPY IMAGES ..................................................... 220
Qinghui Liu, Antonio Ramos, Sergio L. Netto

SESSION 10: AUTONOMOUS SYSTEMS AND MACHINE LEARNING: CONCEPTS, TECHNIQUES AND APPLICATIONS - II

HUMAN-LIKE AUTONOMOUS NAVIGATION FOR COLLISION AVOIDANCE IN DYNAMIC
ENvironments ............................................................................................................................................ 228
Daichi Kawamura, Tad Gonsalves

CYBER PHYSICAL SPACE FOR DOCUMENTATION MANAGEMENT WITH BEACONS ....................... 234
Ole Christian Moholth, Daniel Idris, Afshin Alavi, Tor Egil Øverberg, Radmila Juric

AUTONOMOUS SECURITY SYSTEM SPECIFICATION WITHIN OIL AND GAS INDUSTRY
FOR OFFSHORE VESSEL: REQUIREMENT AND CONCEPT ............................................................. 244
Hugues Broyde, Kristin Falk, Aurilla Aurelie Arntzen Bechina
PROCESS DECOMPOSITION USING AXIOMATIC DESIGN THEORY
Cengiz Togay¹, Gül Tokdemir², Ali H. Dogru³
1 Uludağ University, 2 Çankaya University, 3 Middle East Technical University
{ctogay@uludag.edu.tr, gtokdemir@cankaya.edu.tr, dogru@ceng.metu.edu.tr}

ABSTRACT
This research proposes process decomposition as the initial step in system specification for the cases where software intensive systems are developed by composition. Axiomatic Design Theory (ADT) has been utilized in a systematic decomposition of requirements and design related models simultaneously. We consider process as the top-level flow representation used in the development of software as inspired by the choreography and the orchestration concepts used in Service Oriented Architecture. Since algorithmic level code is provided in the form of web services or software components, a specification of their invocation order becomes the key activity in the development of software. Commonly expressed as a process model, the invocation order becomes the glue for large systems, that need to be modeled within a methodological decomposition that is supported in this research by ADT. Eventually this research offers a method to systematically decompose the requirements, the dynamic model, and a static design model altogether. Hence it bridges the gap in a different approach that utilizes the Unified Modeling Language for object oriented development, between the Use Case Diagrams and the rest, although being applied in the Component Oriented world.

INTRODUCTION
There have been various approaches to developing big software. Many of those approaches targeted code writing. However, with the improvement of development techniques, the exponential characteristics of the demand may be supplied with compositional approaches. It is a known fact that considerable amounts of code is generated automatically today. There are domain specific code generation techniques such as those widely used in the construction of graphical user interfaces, as well as the Model Driven Engineering (MDE) (Pérez-Medina et al., 2007) where automated code generation is the main source of the resultant code for core software development. Also, compositional approaches are improving, offering ways to integrate existing components or web services into the final application. Our proposal here is to systematically utilize the technologies provided for Service Oriented Architecture (SOA) or Component Based/Oriented Software Development or Engineering (CBSD/COSE) (Kaur and Mann, 2010, Dogru and Tanik, 2003).

SOA is actually a very broad definition of the foundations for arbitrating the demand and offer for services. In real world, it is supported with a plethora of techniques that could be effectively utilized for “development by integration”. Some of those techniques include the top-level flow modelling of the activities in process models and search for the web services. These tools imply an effective use of a two-level architecture for constructing distributed systems: The process model and the web services.

COSE on the other hand, has provided a structural view that excluded the invocation order which is being worked on currently. Here, systems are decomposed from their whole to lower-level units, eventually to existing components. Higher-level units can be thought of logical specifications for components. Towards the lower levels, if logical specifications match the real capabilities of the existing components the specification by decomposition activity is hence successful. Invocation order can be superimposed on this basically structural specification, in terms of a collaboration diagram by including ordered message connections. However, there is no reason for not using a process model to effectively model the invocation orders.

In this research, a process model is taken as the first model to be developed to define and even interpretively execute the final system. SOA or COSE technologies can be incorporated, eventually allowing the connection of web services or components. Before any of these, the process model needs to be constructed as the important central entity defining the application. The approach is top-down and the flow control is basically central. Decomposition of the central process model is the first and the important step: the developers will be more efficient if they have knowledge about the domain, especially the components and web services at their disposal. A “mature domain” is one that is populated with an established set of components (Togay et al., 2008). Here on, the term components is used for either components or web services. Also a mature domain is supported with engineers familiar with those components. Consequently, the developers should gain expertise in decomposing a given requirements set in terms of processes that control a given set of components. This research supports such decomposition activity leveraging the ADT systematics.
ADT provides the mechanisms to decompose a related set of models simultaneously, for the specification of engineering systems and establishing modularity in terms of the coupling among the model constituents that are components and process modules in our case. Our previous work included the incorporation of ADT especially in the “structure” dimension of software development (Dogru et al., 2011). This research employs ADT in the dynamic aspects of the specification that is the process model that orders in time, the activations.

We also made use of the process patterns (Aalst et al., 2003) concept that bases our decomposition on well-defined process modules. These patterns categorize the various flow usages such as forking a serial flow of actions to two parallel flows and then joining them. Various synchronization options in forking and joining define many flow patterns that can be used as fundamental blocks. Like the case for many technologies, it would be very beneficial to define patterns for business domains. The patterns defined so far look more syntactic and with generic use. However, they are still great help in organizing a process decomposition effort.

In the rest of this article, background information about existing concepts and technologies is summarized. A description of the suggested methodology is provided. An example development is demonstrated before the conclusions.

Comparison to Object Oriented Development

Widely used Unified Modeling Language (UML) based development suggests some methodological path although it is more a notation. A typical modeling starts with Use Case diagrams that are in the functional dimension, modeling the top level system functions. The next step corresponds to the decomposition of these functions, which is now in the process dimension since ordering is also involved. Many organizations prefer activity diagrams while some use sequence diagrams in the detailing of a Use Case. This detailing is an ordered set of activities or messages. Consequently, process decomposition is realized without a conscious intention that involves separation of concerns. Activity diagram choice ends up with completely decomposing one system function without regarding the structural units (objects). On the other hand, sequence diagram (in general an interaction diagram) alternative is closer to our proposal considering the objects while ordering the messages among them. However, this approach is more declarative for the immediate structure, implying a bottom-up approach for the structure dimension: objects declared for defining messages are later generalized into classes – inheritance and composition relations for higher-level structural definitions come later. The Use Case model is usually evaluated as insulated from the rest of the model enhancement activities.

In our approach, all models manifest their dimensions for separation of concerns. This is one result of engaging a compositional method rather than writing code. The influence of any step of development conducted in any model is instantly given a chance to reflect on any other model before it is too late. In this regard, over-specification is guarded against.

Process Modeling

Process models provide a graphical technique to order the activities that are connected with a common goal in mind. The order can be serial, parallel, conditional, repetitive and in a nested structure utilizing any of these, at different parts of the model. Process models are used in engineering a business, to define how a service or production will be conducted. Also they are used in studying the requirements in software development. Also lately they appear as an alternative to programming, providing a graphical and executable specification. In SOA, they are incorporated in two levels where choreography is in general not executable due to the variety of constituents and orchestration is more executable that integrates web services. Orchestration is also supported with executable engines that interpret a graphical model. Business Process Execution Language (BPEL) is an example to the wide used and effectively executable process models, whereas Business Process Modeling Notation (BPMN) is used widely in expressing more general processes, which was more difficult to execute in its earlier days. Some related terminology can be listed as workflows and enactment where workflow is another model for executable software that is based on a process model. There usually exists a workflow engine to support process models in a specified domain and developing an application in that domain is now basically defining a process model. Enactment is the conducting of the activities as ordered in a process model, which could be partially computerized and partially human executed.

BPMN has been a de-facto standard in both the business world and the software world due to its rich element set. We are also exploiting its syntax in this work. Providing basic structures for serial activities, conditional or parallel branching to a set of other flows, also a combination of conditional and unconditional branching and then joining mechanisms are the fundamental flow capabilities. The atomic unit in a process model is an activity, used in such flow combinations. However, there are other units such as documents, messages, timers etc. Also, the activities can be organized into locations referred to lanes or pools, based on the roles and organizations that are responsible for the activities.

Service Oriented Architecture

As a current approach to building complex systems, this architecture is defined to support loosely coupled distributed systems. Actually it is an evolution of the software structures, in the line of objects, components, and finally web services combined with a software market view for composition. The architecture suggests mechanisms to organize service developers, service users (system integrators) and broker entities that facilitate the meeting of the demand and offer. This abstract architecture is somehow
supported with a plethora of well-established technologies, tools and protocols if utilized carefully, can offer effective solutions to compositional approaches. Figure 1 shows the possible utilization of those technologies in a two-level architecture for fast deployment of applications in mature domains.

![Two-level decomposition architecture based on SOA technologies](image)

**Fig 1. Two-level decomposition architecture based on SOA technologies**

The suggested architecture in Figure 1 is not well published. However, since the early days of experimenting with web services, their usage has been opened for access for many web services. With the development of executable process model that are capable of such web services, this architecture is now practically available. This architecture can also be interpreted as our view where the logical levels of component modeling are supported with a process model for executability.

**Component Oriented Software Engineering**

This is an early attempt that is in line with software architecture centric approaches. Component technologies appeared first, as enabling foundation for compositional approaches. Even popular desktop operating systems are supporting them fully. Then followed component based approaches that aid in the definition, finding, adaptation, and integration of components. Even graphical model building tools for UML claim to be component-based since they allow the insertion of components in a fundamentally Object Oriented approach. Component Orientation (Dogru and Tanik, 2003) however suggests the consideration of only one basic structure that is components. This is from requirements to execution, from abstractions to code. No other concepts such as objects are involved in the models. This is due to a complete orientation towards compositional methods. UML based approaches are however, for defining the code to be written although they have been very successful in this direction.

**Axiomatic Design Theory**

Axiomatic Design Theory (ADT) proposed by the Nam P. Suh is a systematic methodology to decompose systems in a top-down fashion that assists designers to structure design problems (Suh, 2001). As an interdisciplinary methodology, ADT is applied to various disciplines such as software designs, object oriented software design, requirements management, project planning, and Component-Oriented software design (Togay et al., 2008).

ADT has four concepts namely domains (Customer Needs (CNs), Functional Requirements (FRs), Design Parameters (DPs), and Process Variables (PVs)) as depicted in Figure 2, hierarchies, zigzaggig, and axioms (Suh, 2001). “What” and “How” questions express the relation between the domains. Functional needs are represented as FRs. The FRs are satisfied by the system with DPs that can be methods, services, etc. One FR can be satisfied by one or more DPs. PVs implement the DPs such as components or services in terms of the software. Therefore, in software discipline, customer domain represents “attributes desired in the software”, functional domain represents FRs specified for products, physical domain represents processes, methods, abstractions (component, interface, package), and process domain represents real components, and services.

Zigzagging allows a parallel decomposition of all four domains. Process starts with the customer domain and ends with the process domain. There are two fundamental axioms namely Independence axiom and Information axiom. While independence axiom maintains the independence of the functional requirements, information axiom minimizes the information content of the design.

![Domains](image)

**Fig 2. Domains (adapted from (Suh, 2001))**

**Process Decomposition**

Process model size is one of the factors that affect comprehension of process models (Reijers and Mendling, 2011). Larger process models are more prone to errors (Mendling et al., 2007) and hence less manageable. Decomposition of large process models into simpler sub-models would contribute to decrease in complexity (Dumas et al., 2003) as it increases understandability (Mendling et al., 2007) and contribute to easier communication and maintainability (Milani et al., 2016). As decomposition minimizes coupling and increases cohesion it also enables reusability of the sub-models (Bass et al., 1998). There are several studies proposing different approaches to decompose process models.

Top-down decomposition studies suggest methods based on heuristics like breakpoints, data objects, role, shared processes, repetition and structuredness based on the classification of Milani et al. (2016). A process can be decomposed based on specific interest points called milestones during the process lifespan which can be points
where process properties are measurable (Milani et al., 2016, Dijkman et al., 2014). Decomposition can also be performed based on how data objects are shared between the activities (Conforti et al., 2014; de Leoni et al., 2014). Processes can be detached by means of the stakeholder views (Turetken and Demirors, 2011, Malinova et al., 2013). Another basis considered for decomposition is shared processes which combine process parts that are used several times within different parts of the process (Weber et al., 2011). Some studies (Weber et al., 2011, León et al., 2013) anticipate behavioral aspects of process parts like sequential, parallel or cyclical and combine ones with similar behavior. Similarly, “single entry single exit” process parts can be identified in a block-structuring approach and through graph clustering. Related nodes can be combined as a sub-process (Reijers et al., 2011, Huang et al., 2014). Caetano et al. (2010) uses the principle of separation of concerns that are involved in process activities by specifying roles played by activities for process decomposition (Caetano, et al., 2010).

Bottom-up approaches propose merging algorithms based on the graphs like EPCs and BPMN (La Rosa et al., 2013, Gottschalk et al., 2008). Process mining approaches can be employed to learn process models from event logs as well (Aalst et al., 2005). Huang et al. (2016) applied a graph mining algorithm to find out frequent subgraphs under the same process topic for merging processes (Huang et al., 2016). The bottom-up approaches fail to provide consistent rules for decomposition and to identify atomic activities (Caetano, et al., 2010).

Johannsen et al. (2014) has applied Wand and Weber’s decomposition conditions namely minimality, determinism, losslessness, minimum coupling, and strong cohesion for BPMN resulting in decomposition guidelines. They also claim that for each modeling language guidelines and conditions should be redefined.

Although these approaches provide valuable insights into the decomposition practice, they are far from providing a generalized guideline that specifies accepted properties for a consistent decomposition (Johannsen and Leist, 2012) independent of the modeling context.

**METHODOLOGY**

This work actually introduces a new methodology for decomposing process models in the development approaches that also require component structures to be defined. The suggested development would not require the usage of UML for example, or models produced using it. The resulting process model and component structure will constitute the executable code. Developed system will emerge as a result of composing existing components.

Our approach is based on the definition of the order of activities that will be executed through invoking function calls. The functions could be methods of web services or components. The calls are usually conducted by messages. Ordering of the activities are specified in terms of a process model. An outline for the whole development process can be summarized as:

1. Define top-level process model
   a. Decompose Process in agreement with components
2. Connect Activities to components
   a. Conduct necessary adaptations
3. Evaluate the system and iterate for modifications

The critical activity demonstrates itself as the construction of the process model (through decomposition). ADT supports this activity so that the process model does not emerge as one that is not compatible with other considerations. Here our considerations are the requirements and the components. Therefore, the three “domains” for the ADT surfaces as the Functional Requirements (FR) and Design Parameters (DP) besides the Process Flow (PF) that is introduced in this study.

Fundamentally the developer’s task is to decompose the process while attributing the defined member processes with requirements modules and components. This activity is the decomposition of three models simultaneously. The decomposition corresponds to more declarative models. However, the component set is desired to be existent and requirements are expected to be existing and complete. As a result, the decomposition of the FR domain and the DP domain may be easier and not declarative, such decompositions may end up with categorizing the existing terminology in categories defined during decomposition. If components are not existing, the decomposition could still be helpful, coinciding with the COSEML decomposition: Higher level expectations are inserted in the models declaratively and as lower-level decompositions are provided, existing components finally get inserted in the specification. A mapping between the declared units and existing units takes place. Such mappings are referred to as “represents” relations in COSEML. This is a two-way representation, otherwise the developers prefer to use the term “implements” used in a directed association.

Similar considerations can be valid for the set of requirements. Actually, requirements should come in some decomposition naturally. The standards do not suggest more than a few levels of abstraction (decomposition) for them anyway.

**The dynamic dimension**

Considering the PF domain in its right place in time is pivotal. The original domains could be decomposed in zigzag actions, preferably in an order to exhaust a breadth first “zig” from FR to PV. Now besides this defined order in abstraction priority among the static views, we need to insert a zig or zag in its correct time for a dynamic domain. In other words, starting with requirements, when should we consider the execution time ordering for a concept is the question. Our methodology suggests the placement of the ordering as the top-level system property, therefore it should be allocated right after the requirements. A full cycle of
decomposition travels should therefore reflect the following line:

\[ \text{FR} : \text{PF} : \text{DP} : \text{PV} \]

In a graphical view that corresponds more to the ADT tradition, Figure 3 depicts the travels among the model domains. We interpret the ADT domains of DP and PV for methods and components.

![Fig. 3. The model domains of ADT incorporating the dynamic dimension in Process Flow.](image)

At the early stages of the decomposition, a system function will correspond to a process chunk that will be implemented by a set of components through their methods. This will be more or less the picture after the first chain of zigs. The zags will provide a very important enhancement. Considering a set of proposed methods, what could be the ordering among them for achieving the system function? The answer is the final design for that specific system function level requirement therefore it will need more articulation: the decomposition is not over yet. Instead of guessing the whole sequence of invocations, a higher-level draft could serve useful. This higher level process notion will define the mid-level nodes in the process model hence creating the decomposition for the process. Of course, another zag will take us back to the requirements and we can be more informed to decompose the requirements for the next level, now being equipped with the ordering information of its constituents. Hence, one zag will bring one level of decomposition for both structural units and for ordering information. Once all the zig zags are finished, both the process containing the ordering information and the structure that is the components and their methods, will be defined.

A small example may prove useful at this instant where the key contribution is explained. Assume the user authentication requirement. Taken as a top level FR, a zig will take us to the process world where we define the ordered set of activities to establish the goal: enter name, enter password, verify password could be the 3 tasks in the decomposition of the process model. So these three activities are our earliest defined PFs. Another zig to the DP domain, we can imagine methods such as GetName() and GetPassword(). Finally, in the PV world components that processes user typed input: TextProcessor, and another component that connects to the central database through running queries: QueryProcessor can be assumed. The methods can be assigned to those components:

- TextProcessor::GetName(),
- TextProcessor::GetPassword(), and
- QueryProcessor::VerifyPassword(name:string, password:string).

This example provided a one-step decomposition for simplicity. It involved the first step in all the domains: the requirement was decomposed into three requirements, the process was decomposed into three tasks (in more complex processes they would be considered as sub-processes), and two components, and three methods were defined that corresponds to decomposing the code.

The tasks decomposed in the PF domain are similar to the methods in the PV domain. Here, the components in the PV domain provide an implementation medium offered in a structural organization. However, further articulation in the information content related ADT methods do not need to be repeated for the PF domain due to their representing the same functionalities with the methods in the DP domain. Also, coupling in design has been investigated as a static concern in ADT therefore no additional effort is required to apply coupling investigation to the PF elements. Future work may explore a similar relation among process tasks for a more modular design on the dynamic aspects.

An Example Decomposition

As an example, a proposal preparation problem is represented in Figures 4, 5 and 6. FR1 defines the functional customer requirement and during the ZigZag process it is enhanced by sub FRs as depicted in Figure 4. Process decomposition, hence the newly introduced PF domain for ADT extension is represented in BPMN as depicted in Figure 5. PF is also detailed during the ZigZag process. There are two functions namely discount computation and proposal creation. Each function is also represented as a design parameter (DP) in the FR-DP Matrix. These two can be implemented as methods namely getDiscount and createPDF. The GetDiscount function, based on the customerID, selects the user from a database and calculates the discount. The createPDF method creates a PDF file from a proposal template and uses parameters such as customer name etc. These methods are depicted in Figure 6 and they can be part of one or more components.

![Fig. 4. FR-DP Matrix](image)
CONCLUSIONS

Our early work has demonstrated the feasibility of decomposing the flow model for an application as a dynamic model along with the structural models. Mapping across different kinds of models brings in a different kind of thinking to developers. Ordering of invocations and specifying structural units and their connections are considered together, in a systematic manner. Although switching between these two different worlds that a dynamic and static models does not induce corresponding related partial solutions to the developers’ minds directly, we found this approach as a powerful guidance in providing feedback. It feels like extra work in the beginning but soon an extra dimension of feedback for the verification of intermediate steps in development is appreciated.

One extra benefit is the opportunity to assess traceability information between the dynamic and static models. Currently, such relations are exploited mostly between the pairs among requirements, design, or code. Testing structures are also a good candidate for traceability especially from requirements. ADT actually suggests mapping matrices between the models of the neighboring domains (such as FR and DP in original ADT, whereas in this work, FR and PF). Therefore, extra information structures are achieved for traceability between the requirements, the dynamic model, and the components domain. These “Design Matrices” are suggested in ADT to be articulated in providing modularity in the designs in terms of reducing coupling. This property is left for future work for an exciting investigation of modularity in the dynamic dimension that is dependent on the requirements.

This work is in the level of concept proving. It demonstrates the feasibility of considering the decomposition of different views simultaneously for top-down development approaches. As future work, tool support and industrial level experimentation will prove useful.

REFERENCES


ABSTRACT
Software Reliability represents a topic of paramount importance in many sectors of industry and society. Besides, it is characterized by a highly degree of dynamism as shown in literature. Published papers have tackled the study of this research area from different points of view, leading to considerable effort to get a complete overview of existing solutions. This paper focuses on metrics in software reliability in order to: investigate state-of-the-art, synthesize available evidence of their usage, and identify the most popular ones. We have carried out an initial Systematic Mapping Study that analyzes and structures the literature on software reliability metrics, obtaining a total of 386 studies. On the basis of 128 selected primary papers found, the results obtained show an increasing diversity of work. They represent a valid starting point for selecting the most useful metrics for an assessment of reliability in a software system.

INTRODUCTION
Software Reliability is essential in many critical and non-critical applications belonging to sectors of industry and society, such as nuclear security system and spacecraft ground control systems (Lyu, 1996). Reliability represents an important measurable factor of software quality as defined by various quality model standards, such as McCall (McCall, Richards, & Walters, 1977). FURPS standing for Functionality, Usability, Reliability, Performance and Supportability (Grady, 1992) and Dromey (Dromey, 1996). This factor can be assessed by using appropriate metrics, models and tools that are highly specialized in the phase of the software life cycle during which they are applicable. There is no single metric or model or tool that is used in all situations. Hence, it is important to spread their knowledge between engineers and users to make more precise decisions, choose the proper one and track the code status.

Reliability in Software Engineering have concerned the definitions of models since the early 1970s. Many papers have dealt with software reliability models (Lyu, 1996; Littlewood, 1980; Musa, John, & Okumoto, 1982); however, computer scientists have not reached an agreement about which models are the most suitable for assessing the complexity of Software Reliability. This situation penalizes the choice of an appropriate solution to a reliability related problem due to the knowledge required.

A Systematic Mapping Study (SMS) is a methodology that provides a means to systematically analyze a research topic (Budgen, Turner, Brereton, & Kitchenham, 2008; Petersen, Feldt, Mujtaba, & Mattsson, 2008). It provides an overview of a research area by identifying the quantity and type of research and results that are available. SMS is an example of secondary study since it is based on other publication papers called primary studies, which include a model, a technique or a case study. This paper presents a SMS in order to identify and categorize a set of primary studies that deal with the various aspects of reliability metrics in software system.

The reminder of this paper is organized as follows. The second Section presents the related works. The third Section describes the used methodology that is characterized by five phases, each of which is presented in the following dedicated Sections. The eighth Section describes the threats to validity. Finally, the last Section draws some conclusions and provides recommendations for further research on this topic.

RELATED WORK
The literature on Software Reliability provides an overwhelming number of studies dealing with both general and specific issues. Amongst them, we have identified some that can be considered as having similar ideas to our systematic mapping study.

The topic of Reliability is, for the vast majority, focused on Models, which have the purpose to represent the main characteristics of software code in order to determine when to stop testing and to release software. Software Reliability Modeling is based on statistical inference employed to analyze failed data. These kinds of models are called Software Reliability Growth Models and constitute the most conventional way of carrying-out reliability analysis. Other type of analysis is the one that leverages on the software’s internal structure, and the ones based on Bayesian Belief Networks or Test-Based methods. There exist a high number of papers that conducts review or survey of these models (Febro, Calero, & Moraga, 2014).

In (Saley, & Sasikumaran Sreedharan, 2014) the authors have investigated the different approaches used by non-parametric models to predict software reliability. In this work non parametric models use artificial intelligence techniques instead of the traditional statistical methods. Moreover, Saley and Sreedharan focused on different ways of applying Artificial Neural Networks to software reliability prediction.

In (Mohammad Irfan Mohammad Saidi, Mohd Adham Isa, Dayang, Jawawi, & Ong, 2015) the authors analyzed several techniques in software reliability model (SRM) selection, trying to assess their efficiency. They found five different publications: all of them established a certain number of criterions useful to categorize SRMs in terms of
their ability to reproduce and to predict the behavior of the software; some authors also formulated an algorithm based on their suggested selection criterion. Mohammad Irfan Mohammad Saidi et al. conclude that there is no definite approach to determine the criteria for SRM selection.

Chu (Chu, Yue, Martinez-Guridi, & Lehner, 2010) categorized Software Reliability Models into four classes. The first one is based on Software Reliability Growth Methods: none of them is better than the others since all of them leverage empirical formulas that are not applicable to all situations. The second class employs the Bayesian Belief Network methods, which are able to aggregate not apparently related software information and to include parameter uncertainty as a part of the modeling. The third category uses test-based methods in which the authors highlight some drawbacks: large number of tests required, testing environment not representing the actual one, and not consideration of errors in requirements and specification in tests. Finally, last class is based on correlation/regression analysis using past software experience. The employment of this method is challenging because of the unavailability of detailed information on the past software development projects.

In (Lyu, 2007) the author analyses both historical and current software reliability engineering techniques. From the past publications he deduced that there are three main reliability modeling approaches exploiting, respectively: the error seeding and tagging, the data domain, and the time domain. The last, which is considered to be the most popular one, consists of performing curve fitting of observed time-based failure data by a pre-specified model formula, such that the model can be parametrized with statistical techniques. By using extrapolation techniques, the model can provide estimation of current or a prediction of future reliability. As regards current trends, Lyu claims that Software Reliability Growth Model (SRGM) is probably one of the most popular techniques in literature, since there are hundreds of publications that deal with it in literature. In SRGM, there are two measurements related to reliability: the first one is the number of failures in a time period; the second one is the time between failures. More recently, other forms of metrics for testing efforts have been introduced.

In (IEEE, 2008) software models are divided into 2 main categories: software reliability prediction models and software reliability growth models. The first group is composed of models that can be used before testing, they can predict, for example, defect density and planning future rate. The second group is made up of models that can be used during and after testing, such as general exponential models and logarithmic Poisson execution time model.

We noticed that none of the above papers mention explicitly software reliability metrics and only one (February, Calero, & Moraga, 2014) employs the systematic mapping study for carrying-out a complete classification of software reliability models. For these reasons we believed that it was necessary to employ the systematic mapping study technique in the context of software reliability metrics.

METHODOLOGY DESCRIPTION

The method used in this research is a Systematic Mapping Study (Budgen, Turner, Brereton, & Kitchenham, 2008; Petersen, Feldt, Mujtaba, & Mattsson, 2008). A Mapping Study (MS) classifies empirical study data by considering the extent with which they answer one research question. Budgen and Petersen’s papers describe a set of good practices and procedures for undertaking MS in the software engineering context.

A MS entails the analysis of primary studies that address predefined research questions, which can be used to support or refuse particular research hypothesis. According to Budgen the primary reason for performing a MS is making an unbiased assessment of as many studies as possible, highlighting gaps and clusters in the set of primary studies. Furthermore, they can suggest topics and areas in which to perform other SMS.

To conduct this SMS, we have started from Kitchenham and Charters’ (Kitchenham & Charters, 2007) procedure, whose phases are mainly planning review, conducting review and reporting review. Figure 1 shows our adaptation from the aforementioned procedure, where we have put at phase level studies selection and data extraction in order to simplify the activities performed at each phase and to improve the overall understanding.

**Fig. 1 SMS procedure**

**PHASES**

**PROTOCOL DEFINITION**

**CONDUCTING THE SEARCH**

**STUDIES SELECTION**

**DATA EXTRACTION, ANALYSIS AND CLASSIFICATION**

**MAP BUILDING**

**ACTIVITIES**

- Research question definition
- Selection of sources
- Selection criteria definition
- Sources and search strings usage
- Application of the inclusion/exclusion criteria
- Filtering and Classification (title, abstract, text)
A protocol is a plan that defines the basic MS procedures, whereas a map building contains results. In the following sections we explain each phase and apply it to our work in order to identify, structure and classify those primary studies that use metrics for software reliability. This SMS was conducted on June 2017 to cover published between 1980 and 2017 and involved two researchers.

**PHASE 1 - PROTOCOL DEFINITION**

The purpose of this phase is to define the plan that will be used to conduct the SMS. At this stage the following activities are typically conducted as documented in Kitchenham and Charters (Kitchenham & Charters, 2007):

- defining research questions that need to be designed with regard to the objective that this MS has to attain;
- selecting sources in order to find primary studies,
- defining selection criteria to include relevant studies that answer research questions and exclude the others.

Concerning the activity of research questions, we are interested in those that contribute to understanding, characterizing and summarizing evidence and research issues on the topic of our interest. We have defined them on the basis of similar themes addressed by previous research. Answering to the following questions, summarized in Table 1, provides a panoramic view of current practices both in industrial and academic activities:

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1</td>
<td>Which publication sources are the main targets for the software reliability metrics research?</td>
</tr>
<tr>
<td>RQ2</td>
<td>How has the frequency of software reliability metrics research changed over time?</td>
</tr>
<tr>
<td>RQ3</td>
<td>Which metrics are adopted for assessing reliability?</td>
</tr>
<tr>
<td>RQ4</td>
<td>Which metrics are most often associated with software reliability?</td>
</tr>
<tr>
<td>RQ5</td>
<td>What are the analysis techniques applied to interpret measurements of the software reliability factor?</td>
</tr>
<tr>
<td>RQ6</td>
<td>Which research topics have been addressed over the past 10 years in metrics for software reliability</td>
</tr>
<tr>
<td>RQ7</td>
<td>Which study has been conducted with metrics for software reliability?</td>
</tr>
</tbody>
</table>

**Table 1 Research questions**

RQ1: Which publication sources are the main targets for the software reliability metrics research? This question aims at determining the sources amongst journals, conferences, symposium and others that support this research.

RQ2: How has the frequency of software reliability metrics research changed over time? This question provides information about the interest in this research over time.

RQ3: Which metrics are adopted for assessing software reliability? This question is intended to identify the metrics adopted to monitor and control software reliability.

RQ4: Which metrics are most often associated with software reliability? This question is intended to identify the set of metrics that characterize the reliability factor in software.

RQ5: What are the analysis techniques applied to interpret measurements of the software reliability factor? This question aims at identifying the analysis techniques that are the most used during the software development life cycle.

RQ6: Which research topics have been addressed over the past 10 years in metrics for software reliability? This question aims at identifying the state of software engineering research in this topic.

RQ7: Which study has been conducted with metrics for software reliability? This question aims at identifying the study performed with metrics related to software reliability.

For the selection of sources, it is essential to consider availability, accessibility and quality criteria. Therefore, we have firstly adopted scientific databases (see Table 2) that are recognized as valid means (Breteron, Kitchenham, Buddegen, Turner, & Khalil, 2007; Dyba, Dingsoyr, & Hanssen, (2007), 2007) whenever SMS is conducted, such as IEEE Xplore Digital Library, ACM Digital Library, ScienceDirect and Scopus. IEEE Xplore Digital Library delivers full text access to the highest quality technical literature in engineering and technology. ACM Digital Library is a research, discovery and networking platform containing full-text collection of publications. ScienceDirect searches for peer-reviewed literature. Scopus is the largest abstract and citation database of peer-reviewed literature.

**Table 2 Scientific databases**

<table>
<thead>
<tr>
<th>Database</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE Xplore</td>
<td><a href="http://www.ieeeexplore.ieee.org/">www.ieeeexplore.ieee.org/</a></td>
</tr>
<tr>
<td>Digital Library</td>
<td></td>
</tr>
<tr>
<td>ACM Digital</td>
<td><a href="http://www.journal.acm.org/">www.journal.acm.org/</a></td>
</tr>
<tr>
<td>Library</td>
<td></td>
</tr>
<tr>
<td>ScienceDirect</td>
<td><a href="http://www.sciencedirect.com/">www.sciencedirect.com/</a></td>
</tr>
<tr>
<td>Scopus</td>
<td><a href="http://www.scopus.com">www.scopus.com</a></td>
</tr>
</tbody>
</table>

Secondly, we have identified a set of keywords for defining a list of search strings to be used in the selection of primary studies: software, reliability, metrics and measurement.

Regarding the activity of selection criteria, it is important to formalize inclusion criteria and exclusion criteria that are easy to apply without any further interpretation. According to this, we have adopted the criteria shown in Table 3:
PHASE 2 – CONDUCTING THE SEARCH

At this phase the search is conducted in order to identify primary studies that reflect the established plan. This activity leverages automated search engines to which the defined search strings are applied. These engines require particular settings due to their specific way to work and our wish to reduce the number of duplications and rejections.

In the following paragraphs, we have specified the settings performed for each search engine in order to collect papers suitable with our objective. Furthermore, we have specified the search string used to obtain papers and the publication year range for each search engine; the publisher for IEEE Xplore Digital Library and ACM Digital Library search engines; the source type for IEEE Xplore Digital Library and Scopus search engines; the subject area for ScienceDirect and Scopus search engines. We have decided to include these details of information in order to allow other researchers to perform the same search.

For IEEE Xplore Digital Library:
- the search string is ((("Document Title":software reliability) OR "Abstract":software reliability) AND "Abstract":metrics) AND measurement
- the publisher is only IEEE
- the publication year range is 1980 – 2017
- the source type includes conference publications, journals and magazine.

For ACM Digital Library:
- the search string is {recordAbstract:(+software +reliability +metrics +measurement) AND acmdlPublisherName:(not IEEE)}
- the publishers do not include IEEE
- the publication year range is 1980 – 2017

For ScienceDirect:

PHASE 3 – STUDIES SELECTION

At this point of methodology, the selection criteria are applied to identify the relevant primary studies that answer our research questions. Therefore, all papers collected in the previous phase are examined at different levels starting with the title, abstract and keywords. Whenever abstract does not provide useful information to identify the content of the paper, it is recommended to read the whole paper.

This selection mechanism is characterized by two types of activities: the former consists of reviewing papers according to the defined criteria and it is typically performed by one researcher, the latter comprises the assessment of the reviewed papers and it was conducted by both the authors of this work.

During the reviewing activity, the included papers are labeled in order to facilitate the following analysis. Then, these first valid papers are evaluated in order to identify any errors performed by one single researcher and to find an agreement for their inclusion or exclusion. Table 4 shows details for the included and excluded papers in each search database. We have obtained a total of 128 papers (~33% of the total papers): 85 from IEEE Xplore Digital Library, 11 from ACM Digital Library in which IEEE papers have been excluded, 4 from ScienceDirect and 28 from Scopus. We have excluded a total of 258 papers (~67% of the total papers). We have noted that the number of duplicities was quite low due to the search settings. Furthermore, the excluded papers mainly respond to the criteria E2, E3, E4 and E5 detailed in Table 3.

Table 3 Selection criteria

<table>
<thead>
<tr>
<th>Type</th>
<th>SC</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inclusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SC</td>
<td>The paper relates to software reliability metrics.</td>
</tr>
<tr>
<td>12</td>
<td>SC</td>
<td>The paper evaluates a subset of metrics for software reliability.</td>
</tr>
<tr>
<td>13</td>
<td>SC</td>
<td>The paper discusses aspects of software reliability.</td>
</tr>
<tr>
<td>14</td>
<td>SC</td>
<td>The paper is published in a journal or magazine or conference proceedings.</td>
</tr>
<tr>
<td>15</td>
<td>SC</td>
<td>The paper has citations.</td>
</tr>
<tr>
<td>Exclusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>SC</td>
<td>The paper is not available.</td>
</tr>
<tr>
<td>E2</td>
<td>SC</td>
<td>The paper is not accessible.</td>
</tr>
<tr>
<td>E3</td>
<td>SC</td>
<td>The paper is not written in English.</td>
</tr>
<tr>
<td>E4</td>
<td>SC</td>
<td>The paper has 0 citation.</td>
</tr>
<tr>
<td>E5</td>
<td>SC</td>
<td>Only abstract is available.</td>
</tr>
</tbody>
</table>
Table 4 The search summaries

<table>
<thead>
<tr>
<th>Database</th>
<th>Included</th>
<th>Excluded</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE Xplore</td>
<td>85</td>
<td>86</td>
<td>171</td>
</tr>
<tr>
<td>Digital Library</td>
<td>11</td>
<td>40</td>
<td>51</td>
</tr>
<tr>
<td>ACM Digital</td>
<td>4</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>Library</td>
<td>28</td>
<td>121</td>
<td>149</td>
</tr>
<tr>
<td>ScienceDirect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scopus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>128</td>
<td>258</td>
<td>386</td>
</tr>
</tbody>
</table>

PHASE 4 – DATA EXTRACTION AND ANALYSIS

This phase is characterized by the reading of the title and of the abstract of the papers with the purpose of finding terms and concepts that may summarize the contribution of one paper.

The classification categories are thus obtained by using the keywords extracted from the meaningful terms found in the title, abstract and author’s keywords provided by the authors. Therefore, the reviewers rely on what the authors claimed about their work. Once gathered all the keywords they are clustered into relevant categories according to the relevant questions, e.g. the words “reliability” and “metrics” are keywords that are used to classify the studies.

The classification scheme can be developed either from scratch i.e. just relying on the set of included primary studies, or on the base of previously proposed taxonomies for the topic under study. We have chosen to build our classification schema from scratch and it will be discussed in a forthcoming full paper.

PHASE 5 – MAP BUILDING

The last phase focuses on the relevant papers which have to be classified according to the classification scheme already designed. In this phase the actual data extraction takes place, the classification scheme can still change together with the data extraction: the reviewers can add new categories or merge or split the existing ones.

By analyzing the final set of primary studies it emerges that the search engine that has provided the highest number of relevant studies is IEEE Xplore Digital Library as shown in Table 4.

For what regards the answers to the research questions, this paper include those related to RQ1 and RQ2; the others will be included in a forthcoming full paper with related discussions.

With regard to RQ1, the primary papers appear in various journals, conferences and symposiums with predominance in conference proceedings as shown in Figure 2 (the list of Journals obtained in the search for primary studies is available in Appendix A). Amongst the selected studies we have noted that the paper with the highest number of citations (that is 269) is titled ‘Predicting defects using network analysis on dependency graphs’ and it was presented at the International Conference on Software Engineering (ICSE) in 2005.

As regards RQ2, which is the number of papers published over time, we have found that the trend of activity in this topic area is not easy to be identified, except for the number of publications decreasing since 2009. However, 16 is the number of papers published in 2017 but not included in the primary studies due to the lack of citations. Figure 3 shows the publication frequency from 1980 until 2016.

THREATS TO VALIDITY

Although this research has been conducted by abiding to the rules of the systematic mapping study, there may be some threats to its validity. Firstly, the limited access to sources may have led to a bias in the publications’ research and, consequently, in their selection. However, the databases used
are the most popular in the field of software engineering; therefore, we believe they can cover the area of software reliability well. Thus, we are reasonably confident that we have not missed any important publication.

Another threat derives from the different terms that may have been used in the topic of interest. For this reason, we could have missed some papers, since the search string we used contained the keywords “Software Reliability Metrics Measurement”. However, the above four keywords are the most common in literature and the likeliness of one contributions that do not include them is minimal.

Some papers may not have been found because of the quality of the search engines used and the way researchers write their abstract and their keywords. For this reason, we employed four different search engines, amongst the most popular ones for this kind of study. Thus, we believe that we managed to reduce the error as much as possible.

The set of questions we defined might not have covered the whole area regarding metrics and reliability. Although we consider this as a feasible threat, we had different meetings in order to write and refine the questions. This way, even if we have not selected the most appropriate questions, we tried to address the most common open issues in the field.

The decision of excluding papers without citations and of including only publications on journal may have led to miss significant studies. However, our purposes were taking in account and classifying well known papers.

Finally, we excluded studies not published in the English language. For this reason, we may have left important studies published in other languages. However, English is the most wide-spread language used for scientific papers, thus the potential bias due to this choice may be considered minimal.

CONCLUSIONS

The main contribution of this paper is to provide an initial Systematic Mapping Study on software reliability, with particular attention to code metrics. We tried to achieve this goal by employing a currently standard technique for Systematic Mapping Study.

We defined one search query and exploited the results obtained by 4 different search engines.

We obtained a total of 386 works, 128 of which were considered suitable for answering our research questions. All the studies were classified according to publication attributes, the nature of the work, topic and content features.

The forthcoming work will contain the classification schema and a complete and detailed section addressing the different research questions. In future work, we will extend this work by using a systematic literature review that can involve more specific topics.

ACKNOWLEDGMENT

This work is supported by the National Institute of Nuclear Physics (i.e., Istituto Nazionale di Fisica Nucleare – INFN).

APPENDIX A - LIST OF JOURNALS

<table>
<thead>
<tr>
<th>Journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE Journal on Selected Areas in Communications (J-SAC)</td>
</tr>
<tr>
<td>IEEE Software</td>
</tr>
<tr>
<td>IEEE Transactions on Neural Networks and Learning Systems</td>
</tr>
<tr>
<td>IEEE Transactions on Services Computing</td>
</tr>
<tr>
<td>IEEE Transactions on Software Engineering</td>
</tr>
<tr>
<td>IEEE Transactions on Systems, Man, and Cybernetics</td>
</tr>
<tr>
<td>- Part A: Systems and Humans</td>
</tr>
<tr>
<td>IEEE Transactions on Reliability</td>
</tr>
<tr>
<td>IEICE Transactions on Information and Systems</td>
</tr>
<tr>
<td>Computer</td>
</tr>
<tr>
<td>Empirical Software Engineering</td>
</tr>
<tr>
<td>Information and Software Technology</td>
</tr>
<tr>
<td>International Journal of Engineering and Technology</td>
</tr>
<tr>
<td>International Journal of Web Services Research</td>
</tr>
<tr>
<td>Journal of Software Maintenance and Evolution</td>
</tr>
<tr>
<td>Journal of Systems and Software</td>
</tr>
<tr>
<td>Journal of Systems Architecture</td>
</tr>
<tr>
<td>Microprocessors and Microsystems</td>
</tr>
<tr>
<td>Software engineering journal</td>
</tr>
<tr>
<td>Software Quality Journal</td>
</tr>
<tr>
<td>Nuclear Engineering and design</td>
</tr>
<tr>
<td>Tsinghua Science and Technology</td>
</tr>
</tbody>
</table>

REFERENCES


ABSTRACT

Scheduling techniques have been applied to a large category of software systems, such as, processor scheduling in operating systems, car scheduling in elevator systems, facility scheduling at airports, antenna scheduling in radar systems, scheduling of events, control signals and data in cyber-physical systems, etc. Designing and implementing software systems that incorporate scheduling techniques are not trivial, due to complexity, large size and safety-critical nature of such systems. To manage complexity, one may adopt domain-specific abstraction techniques. Adopting Software Product Line Engineering (SPLE) approaches can reduce the cost of developing large size of scheduling software. Built-in analysis tools may be utilized to assure that the scheduling process can be realized correctly. This paper introduces a comprehensive SPLE approach to foster reuse in implementing scheduler product families. As a part of the SPLE approach, a feature model, a wizard-based product configuration tool and a set of integrated verification tools have been designed and implemented. To the best of our knowledge, this is the first paper in the literature that presents an SPLE approach in the scheduling domain which covers a large category of scheduling techniques, with built-in product configuration, implementation and verification tools.

INTRODUCTION

Scheduling is a decision-making process in which the resources are allocated to the activities over time (Pinedo 2008). Historically, scheduling processes have been applied to a large category of software systems (Buttazzo 2011), for example, processor scheduling in operating systems (Silberschatz et al. 1998), car scheduling in elevator systems (Nikovski and Brand 2003), work-force scheduling in project management (Kolisch and Hartmann 1999), facility scheduling at airports (Saraf and Slater 2006), antenna scheduling in radar systems (Hochwald et al. 2004) and assembly line balancing (scheduling) in factories (Becker and Scholl 2006). Due to the emergence of new application domains, adoption of schedulers in software tends to increase. For example, scheduling of events, control signals and data in cyber-physical systems (Tabuada 2007), sensor networks (Xiao et al. 2006) and big-data architectures (Gautam et al. 2015) play a crucial role in the overall functioning of these systems. In general, companies, which market such systems, have to design and implement family of products instead of a single product.

Firstly, designing and implementing robust scheduling systems, however, are not trivial. The complexity of scheduling processes depends on the requirements. In simple cases, it may reduce down a standard sorting algorithm. Unfortunately, in the applications listed above, due to many operational constraints in the requirements, software which implements scheduling processes can be very complex. For example, the multi-car elevator scheduling problem is known to be NP-hard (Koehler and Ottiger 2002). Secondly, the detail in the requirements may result in a very large scheduling system. For example, software systems that implement facility scheduling at airports are very large systems, since they contain many interrelated scheduling parameters such as planes, runways, gates, staff members, passengers, etc. Finally, products that adopt scheduling techniques are in general safety-critical systems. As such determining the schedulability of activities (Tindell and Clark 1994) is important.

To overcome these challenges, the following measures can be adopted. Firstly, complexity can be addressed by effective usage of abstraction techniques. Secondly, the cost of developing large-size scheduling software can be reduced by adopting software reuse techniques. To provide a systematic reuse in the development processes, to deal with the variations in the products, and to support development of family of products, adopting Software Product Line Engineering (SPLE) approaches can be also very beneficial. Finally, built-in schedulability analysis tools may be adopted to assure that the scheduling process can be realized properly.

Unfortunately, to the best of our knowledge, there has not been a comprehensive work in the literature to address these challenges in an integrated way. There is a lack of generic domain-specific modeling techniques, libraries and application frameworks that can be used in building scheduling applications. The engineers who develop software in such application areas need to deal with the large complexity of such systems and due to lack of high-level reuse, they may develop software from scratch again and again. Accordingly, to the best of our knowledge, there has not been any substantial work in the application of SPLE to scheduling techniques. Moreover, most schedulability analysis tools are built from general-purpose verification tools and they are naturally not integrated with scheduling software to be developed.
The novel contribution of this paper can be summarized as follows. Firstly, a feature model (Kang et al. 1990) and built-in verification tools are introduced to systematically deal with the complexity of designing scheduling applications. To this aim, a model is derived by carrying out an extensive literature study on the background material. The model is expressed formally using the specification language of Clafer (Antkiewicz et al. 2013). Secondly, a comprehensive SPLE approach is described and applied to foster reuse in implementing scheduler product families. Finally, by adopting various model-checking tools, schedulability analysis techniques are implemented and integrated in the SPLE process. Currently, our SPLE asset repository contains 18071 and 927904 lines of code written in Python and C/C++ programming languages, respectively. This paper mainly focuses on the definition, evaluation and verification of feature models (early phases of SPLE) but not in the code generation and implementation phases.

This paper is organized as follows. The related work is given in the next section. The section Objectives summarizes the aimed objectives of this paper. The section Background gives the background necessary to read the paper. Our SPLE approach is introduced in the following section. After this section, the feature models are derived from the domain knowledge. In the section Experiments, a number of canonical examples are presented to demonstrate the expressivity of the feature model and in the section Evaluation, an assessment of our approach is presented. The last section concludes the paper and summarizes the future work.

RELATED WORK

There has been a considerable number of publications which reports on the practical application of PLE approaches (Linden et al. 2007). To the best of our knowledge, the PLE methods and techniques have not been applied in the domain of scheduling, except the work published by Hatun et al. (Hatun et al. 2011). This has been a preliminary work in defining a feature model for schedulers. The feature model as presented in the paper has adopted very basic models for tasks and resources and lacked solver based solutions to the scheduling problems. Furthermore, the proposed idea has not been implemented. Furthermore, the paper does not present any method to verify the correctness of feature models and to analyze the schedule-ability of the problems.

There are many researchers focusing on scheduling problems, and such much research work has been published in this area. However, most of these publications do not aim at expressivity and reusability from the perspective of software engineering.

Many application frameworks have been defined in the literature and applied in various areas (Johnson and Foote 1988), (Fayad et al. 1999). However, we assume that there has not been any framework implementation on schedulers except the work presented in this paper.

There have been various publications on the formal verification of feature models (Batory 2005). Most of these publications do not integrate the verification process in a design environment. Moreover, verification of feature model instantiations in the area of scheduling has not been studied before.

The application of constraint solvers in different sub-domains of computer science is not new. Many different kinds of solvers have been studied and presented in the literature (de Moura and Bjørner 2008; Hutter et al. 2010). Furthermore, schedule-ability problems have been extensively studied (Fersman et al. 2006; Tindell and Clark 1994). The constraint-based implementations have been introduced in the areas of planning and scheduling (Fromherz 2001; Heinz et al. 2013; Hooker 2007) and various solvers have been utilized in these studies. However, the resource models in these articles are not expressive enough to support more complex hierarchical resource structures. There exists a study (Kim et al. 2016) which presents formal framework to analyze a family of products with respect to real-time properties. It uses UPPAAL model checker to detect whether any task is missing its deadline in the taskset and to perform schedulability analysis. However, its task and resource models are not generic and lack optimization of complex scheduling objectives.

THE OBJECTIVES

The objective of our work is to define methods, tools and techniques to create scheduling applications with the following characteristics:

Reduced complexity: To deal with the complexity, it is important to identify the necessary abstractions in the scheduling domain, represent these adequately and encapsulate the unnecessary details from the designer.

Enhanced reuse and expressivity: To manage the cost of development of complex scheduling applications, reuse becomes important. While providing reuse, special attention must be paid to explicitly deal with the variations in the scheduling domain. Enhanced reuse can be accomplished by providing dedicated reuse mechanisms (a) in the development processes and (b) in the software models, both tailored to the scheduling domain. Furthermore, due to the large variety, the software models and the corresponding libraries must be expressive enough to cover a large category of scheduling applications.

Verification: In the process of configuring applications from the corresponding software models and libraries, the created product (a) must not violate the constraints of the domain, and (b) the product must be able to schedule the activities as desired.

BACKGROUND

The background material of this paper is within the context of scheduling theory, Software Product Line Engineering (SPLE), and schedulability analysis, and verification of feature models, which are briefly summarized in the following three subsections.
The Scheduling Theory

Scheduling is the process of allocating resources to activities. Constraints and objectives are the two basic factors that influence this allocation process. The constraints restrict the allocation space; whereas the objectives are the criteria to be satisfied within the allocation space. From this perspective, a scheduling problem can be seen as a sort of an optimization problem. The definitions explained in this subsection are used to derive the domain model presented in the section where the feature models are derived.

In the literature, there exist some differences in the terminology used. For instance, activities are termed as tasks in (Buttazzo 2011); jobs and operations in (Graham et al. 1979; Pinedo 2008). Throughout the article, we adopt the terms tasks.

A task is defined by the following attributes: $r_j$ is the release time of the task $j$. It is also called as the earliest start time. $c_j$ is the execution time of the task $j$. It is the worst-case execution time (wecet) required by a task to complete its execution. $d_j$ is the deadline (due date) of the task $j$. It is also called as the latest finish time. $p_j$ is the period of the task $j$. It is the inter-arrival time between two consecutive instances of the task unless it has only one. $w_j$ is the priority of the task $j$. It defines the relative importance of the task among the others. It is also termed as the weight.

Graham (Graham et al. 1979) has introduced 3-field notation to define scheduling problems:

\[ \alpha \beta \gamma \]

where $\alpha$ is the machine environment (resources) on which the tasks (activities) run; $\beta$ represents the task characteristics (constraints) which are the restrictions based on tasks and resources; and $\gamma$ refers to the optimality criteria (objectives) which need to be optimized.

Machine Environment

Machine environment is defined as a pair $\alpha_1\alpha_2$, namely machine identifier and number of the machines, respectively. The knowledge presented in this section is used in the section Evaluation to assess the expressivity and reusability of our approach.

In (Graham et al. 1979), 6 fundamental machine identifiers ($\alpha_j$) are introduced:

$I$: It refers to a single machine environment on which the tasks can execute sequentially.

$P$: Machines are identical and in parallel, meaning that the execution speed of the tasks does not vary from one machine to the other and the tasks can be executed in parallel unless there is restriction.

$Q$: Unlike identical machines, the resources have different execution speed; therefore the execution time of the task depends on the speed of the machine. The following ratio showing the operational load of the task must be the same for each resource: $c_jv_j = c_jv_2$. Here, $v$ refers to the speed of the machine.

$O$: It is a machine environment called Open-Shop, where there is no restriction about the execution order of the operations, but they cannot be processed in parallel. Moreover, there exists one-to-one relation between the operations and the resources, meaning that a task has to execute on each machine.

$F$: The Flow-shop environment differs from the Open-Shop in one aspect. Each task has a predetermined path on each one of the machines.

$J$: In Job-shop machine environment, like the Flow-shop environment, a task has a route of execution, but it does not need to visit each machine.

We express the number of the machines ($N_\alpha$) using positive natural numbers $N^\alpha$. For instance, M-machine Flow-Shop environment is denoted as $\alpha = FM$ where $M \in Z^\alpha$.

Task Characteristics

In this subsection, we will further elaborate on the task characteristics ($\beta$) introduced in Definition 1. Constraints are defined as follows:

$\beta_1 = \{pmtn, \epsilon\}$: This expresses the preemption ability (preemptability) of a task. It gives a capability to the scheduler to stop and restart the task.

$\beta_2 = \{r_p, \epsilon\}$: If this symbol exists in $\beta$, then a task may have specific release time. Otherwise, any task may start at anytime.

$\beta_3 = \{prec, \epsilon\}$: The precedence relation of tasks blocks the commence of one task if it requires the completion of another.

$\beta_4 = \{M_j, \epsilon\}$: Machine Eligibility constraint obliges tasks to run on only specific subset of all resources. It is valid for only parallel machine environment.

$\beta_5 = \{p_j = p, \epsilon\}$: It represents the tasks have fixed execution time $p$.

$\beta_6 = \{d_j = d, \epsilon\}$: If this constraint is defined in the list of task characteristics, then each task is supposed to complete the executions before the fixed deadline $d$.

$\beta_7 = \{s_{jk}, \epsilon\}$: The sequence dependent setup time represents the required amount of time for a machine to start to execute
the task \( k \) after finishing the task \( j \). Unless it is specified, all the setup times are assumed to be zero initially.

\[ \beta_0 = \{ \text{batch}(b), \epsilon \} \]: A resource may execute \( b \) tasks simultaneously if this constraint is defined. The entire batch is finished when the last task of the batch has been completed and the execution times of the tasks may not be equal.

\[ \beta_0 = \{ \text{prmu}, \epsilon \} \]: It is only valid in flow-shop environment. The queues in front of a machine have First-In-First-Out policy. Therefore, the order of the tasks is maintained on each machine.

**Scheduling Objectives**

In the literature, we have found 7 criteria formulated based on the completion times of the tasks, which is denoted as \( C_j \) (Brucker 2007). These objectives are as follows:

The objective function that is chosen to be optimized is deployed to determine the quality of the calculated schedule. From this perspective, the aim of minimizing the objective function Lateness is to complete the tasks immediately after their release time, whereas the objective function Earliness is the complement of Lateness. Minimizing Tardiness aims to avoid only the deadline misses regardless to the placement of the task in time. The quality of the schedule remains the same if the task is scheduled and completed before its deadline. The remaining items in the list (Table 1) are derived from these functions and are considered self-explanatory. In addition, the common objective functions are formulated as maximum, summation and weighted summation over the tasks. For instance, the maximum \( L_{\text{max}} = \max_j L_j \), the summation \( L_{\text{sum}} = \sum_j L_j \) and the weighted summation \( L_{\text{sumw}} = \sum_j w_j L_j \) are the versions of the Lateness objective. The linear combinations of these formulas are also considered.

**SPLE Approaches**

SPLE as a software development method (Lee et al. 2002) has been introduced to increase efficiency and effectiveness in developing software product families with the help of dedicated software reuse techniques. It constitutes of two development processes, namely domain and application engineering. In the former, commonalities and variabilities are determined with respect to the requirements over family of products. To express commonalities and variabilities, the feature model diagrams are commonly adopted (Batory 2005; Kang et al. 1990; Lee et al. 2002). In the application-engineering phase, products are defined by binding the variabilities. The term binding (Pohl et al. 2005) refers to either fixing or removing the optional features in the feature model. To create executable software system, the software engineer may deploy a corresponding library and/or application framework or utilize a code generator tool, if any.

**Schedulability Analysis and Verification of a Feature Model**

The purpose of schedulability analysis is to determine whether resources can be allocated to a set of tasks by satisfying the constraints. In the literature, there are two
possible ways to perform schedulability analysis. Firstly, by using proof by induction, the processor utilization (Liu and Layland 1973) or processor load (Davis and Burns 2011) is calculated for boundary task set conditions. A taskset is said to be schedulable in any case with a given scheduling algorithm and on the corresponding resources unless the utilization factor does exceed the boundary value. In the second approach, schedulability tests are performed using a model-checking tool (Annell et al. 2004; David et al. 2009), which generates all possible states of the scheduling process. If there exists a path in the state space in which all the tasks complete by complying the constraints, the system is said to be schedulable.

A feature model encompasses cross-tree constraints as well as various kinds of features. There are two types of relations in terms of cross-tree constraints, namely requires and excludes. In requires relation like req( fa, fb ), binding the feature fa makes the existence of fb inevitable in the same product configuration; whereas for the relations like exc( fa, fb ), there exists no product configuration in which both the features fa and fb are together. Due to these kinds of relations, the feature model is prone to the inconsistencies. To detect the inconsistencies, the verification process is performed on the feature model. It includes the following steps as defined in (Zhang et al. 2004): Firstly, the relations between features and the constraints are identified. Secondly, the constraints are translated in a formal language. In the third step, the state-space exploration and constraint simplification is performed to decide the set of products by satisfying the logical expressions from the previous step.

**OUR SPLC APPROACH FOR SCHEDULERS**

We have developed a novel SPLC approach for configuring applications that incorporate schedulers. Fig. 1 depicts our application engineering process.

The specification is constructed as a dedicated configuration of the feature model. The description of the feature model is given in the next section. By the help of Configuration Tool for Feature Model, which is shown on the top-left corner of the figure, the software engineer binds the variabilities defined in the feature model to specify a dedicated application. The binding process may result in zero or more configured feature models. A zero configuration indicates that the binding process does not result in any application, possibly because the specification is too restrictive. In this case, the software engineer must reconsider the configuration. A single configuration refers to a desired specification, in which all the variabilities are bound. If there is more than one configuration, not all variabilities are bound and as such the configuration refers to multiple applications. In this case, the software engineer needs to complete the binding specifications.

In the figure, Verification Tool for Configuration takes the Configured Feature Model data (also called partial configuration in the literature) and checks the number of configurations. If there is a single configuration (also called product), Framework Configuration & Instantiation Wizard is used to guide the software engineer in providing the necessary detailed information for creating the implementation of scheduling application. In the current implementation, the wizard incorporates a traversal algorithm, which walks through the relevant branches of the feature tree and generates a report, which describes the configuration of the application and the parameters to be instantiated. The software engineer is requested to follow the guideline and supply the missing parameters so that the scheduling application can be instantiated. Schedulability Analysis Tool can check the schedulability of the application whether there exists a valid schedule for the given set of requirements or not.

In our approach variability of the product line is supported in two complementary ways. Product-level variability is realized by the configuration wizard and by the instantiations provided by the software engineer. As a result, a dedicated constraint specification is generated. Due to our solver-based approach, at the scheduling level variability is provided by the run-time environment of the solver. More information can be found in the section Evaluation.

**A DOMAIN-SPECIFIC MODEL FOR SCHEDULERS**

In this section, we present our domain analysis and the feature model for schedulers. Most studies in the scheduling domain offer dedicated algorithms that are claimed to be a beneficial in certain context (Brucker 2007; Burns 1994; Buttazzo 2011; Davis and Burns 2011; Dertouzos and Mok 1989; Hong and Leung 1988; Liu and Layland 1973; Pinedo 2008; Sha et al. 2004; Sprunt et al. 1989). Our aim in the domain-engineering phase, however, is to seek for a generic model, which can represent the common aspects of most relevant scheduling techniques. To realize this, we aim to identify commonalities and variabilities in the scheduling and related domains; these refer to mandatory and optional features in the feature model, respectively. The purpose of the feature model is to support our SPLC approach as such it

---

Fig. 1 The application of our SPLC method. Rectangles and ellipses represent the operations and the data, respectively.
defines the mandatory and optional features in the scheduling domain. In the application-engineering phase, it must be possible to conveniently configure and instantiate dedicated schedulers derived from the feature model.

After an extensive comparison of the models adopted in different approaches, we have found out that the following model is expressive and generic enough to represent a wide range of scheduling solutions:

\[ Sch = (T, R, C, S) \], (2)

where \( Sch \) is the scheduler; \( T \) is the task; \( R \) is the resource; \( C \) is the scheduling characteristics; and \( S \) is the scheduling strategy. The corresponding feature model is depicted in Fig. 2.

The term \( R \) in Definition 2 has a similar meaning as \( \alpha \) used in Definition 1. After analyzing the related literature in detail, we have observed that the definitions of the terms \( \beta \) and \( \gamma \) strongly relate to all the fields of Definition 1. For example, \( \beta \) may define both the characteristics of tasks and characteristics of scheduling processes. Similarly, \( \gamma \) relates to the objectives related to both tasks and resources. Although we believe that Definition 1 is useful for expressing the domain of scheduling problems in general, it is less optimal for reuse purposes. Moreover, variations in the implementation of schedulers tend to be classified as variations in the definitions of tasks, resources, characteristics of the scheduling processes and scheduling strategies. The fields of Definition 2, therefore, are considered to be more suitable to encapsulate these variations.

There are also various cross-tree constraints between inter- or intra-models. These constraints are necessary because naturally sub-models are not orthogonal to each other. Such constraints limit the number of scheduling applications that can be configured. Elaboration on the cross-tree constraints are considered outside the scope of this paper. A more interested reader can refer to the feature model definition in the repository.

In the following subsections, the sub-features of the model for schedulers (Fig. 2) and their structures are explained in detail.

**The Task Model**

Most publications in the scheduling domain come along with their own task models. Nevertheless, they largely share common terminology. Based on our literature analysis, we generalize the task model with the following formula where its feature model is shown in Fig. 3:

\[ T = (T_{\text{time}}, T_{\text{req}}, T_{\text{dep}}, T_{\text{pro}}, T_{\text{pan}}, T_{\text{gran}}, T_{\text{obj}}) \], (3)

In the definition above, the symbols are explained as follows:

\( T_{\text{time}} \): It corresponds to the fundamental time-related attributes of a task, which identify the task. These are:

\[ T_{\text{time}} = (r, c, d, p) \], (4)

where the parameters above correspond to exactly the same task attributes explained in the section where the scheduling theory was presented. Moreover, we defined the time interval between release time and deadline of a task as *life-scope*. More information can be found in (Buttazzo 2011).

\( T_{\text{req}} \): It represents the *resource requirement*, mutex requirement and *deadline requirement* of a task. On the contrary to the traditional notation in Definition 1, a task may have more complex resource requirements. For instance, expressing a system with *shared* (Buttazzo 2011) and *multi-unit* resources (Holenderski et al. 2012) is impossible using the traditional notation. Therefore, the *resource requirement* is supposed to be specified accordingly for each task in our resource model. The *mutex requirement* is the principal feature related to the shared resources. If there is a task, which modifies the internal structure of a resource, such as writing to the memory space, the other tasks have to wait until it is done. For instance, assume that there exist three tasks using the same-shared resource and \( r \) has mutex requirement. This means that the scheduler produces a schedule in which the life-scope of the other two tasks cannot overlap with the life-scope of the task \( r \). The *deadline requirement* of a task represents the consequences of missing the deadline on a system. These are hard, firm and soft which have been introduced and explained in (Burns 1991; Buttazzo 2011; Sprunt et al. 1989).
Tdep: It refers to the term data dependency between tasks. We adopted the token-based dependency approach explained in (Ahmad et al. 2014; Lee and Messerschmitt 1987). It blocks the execution of a task until the dependent data (token) is fired even if its release time has already passed. In addition, we have defined sequence dependent setup time (Pinedo 2008) in the context of data dependency such that after the dependent token is fired, a task may either immediately start to execute or needs to wait the amount of setup time.

Tpr is: It represents the priority of a task, which is also termed as weight in the section Scheduling Theory.

Tpmrn: It refers to the preemption ability (preemptability) of a task. The preemptable tasks can be interrupted in time and continue afterwards.

Tgran: It represents the granularity of a task. It can be either terminal or composite. The tasks in the literature have terminal granularity according to our definition; whereas the composite task can be decomposed into terminal and other composite tasks, which are contextually related to each other.

Tobj: In the literature, the objective is the performance measure on the overall system to be followed in the schedule (Buttazzo 2011). Nevertheless, a task might also have explicit individual objectives. For this reason, we defined the task-related objectives.

The Resource Model

Most publications in the scheduling domain have introduced their own specific resource model. Nevertheless, although there are differences in resource models, most publications adopt similar terminology. Based on our analysis of the literature, for the sake of expressivity and generality, we propose the following model shown in Fig. 4:

\[ R = (R_{\text{cap}}, R_{\text{type}}, R_{\text{mode}}, R_{\text{pow}}, R_{\text{obj}}) \]  

(5)

Rcap: It represents the capacity of a resource. As classified in (Holenderski et al. 2012), there exist two kinds of resources, namely single- and multi-unit. In our model, all resources are defined as multi-unit resources. Nonetheless, a single-unit resource can be expressed by setting the capacity value to one.

Rtype: It represents the type of a resource. In (Zhao et al. 1987), the resources have been categorized in two groups, namely active and passive. During the duration when a task performs its actual operation on an active resource, it may also require other passive resources to perform supplementary operations such as sensing the environment, storing data, etc. The central processing unit in computers can be given as an example to the active resources; whereas bus, memory, peripheral devices are the examples of the passive resources. Furthermore, another factor affecting the allocation of the resources is the accessibility of the passive resources with respect to the active resource where a task runs. For instance, in computers, the cache memory embedded into a processor can only be accessed by a task running on the same processor. To realize accessibility, we defined a new abstraction named composite resource and we represent the resource system as tree where the terminal nodes are active or passive resources and all other internal nodes correspond to the composite resources. A task running on an active resource \( R_a \) has access to the passive resources, which are the siblings of \( R_a \) or the terminal resources of its ancestors. We defined the abstraction attribute of the type of the resource as follows:

\[ \{\text{ACTIVE, PASSIVE, COMPOSITE}\} \]  

(6)

In addition, we defined the identifier \( \psi \) that helps user to specify human-understandable name for any type of the resource. For instance, the abstraction and identifier for processing unit might be \( \text{ACTIVE} \) and “CPU”, respectively. Consequently, the type of the resource composed of two attributes, which can be formalized as follows:

\[ R_{\text{type}} = (\psi, \gamma) \]  

(7)

It is crucial to note that there may be more than one resource with the same type, but each resource has only one type. As stated in the previous section, the resource requirement of a task is determined by using the type of the resource and

![Fig. 4 A feature model representation for resources.](image-url)
required capacity on that kind of resources.

Rnode: It corresponds to the mode of the resource. In (Buttazzo 2011; Zhao et al. 1987), the resources are grouped in two groups based on share-ability among tasks. The shared resource can be occupied by many tasks each of which requires at most the same amount of capacity that the resource provides. On the other hand, the exclusive mode can be divided into two sub-categories: capacity-based and semantic-based. For a resource running in capacity-based exclusive mode, the total capacity utilized by the tasks cannot exceed the capacity of the resource; whereas the semantic-based exclusive mode allows only one of the mutually exclusive resources to run at a time.

Rpow: It represents the power consumption of the resource. Due to the increasing trend in mobile devices, one of the main concerns is to reduce power consumption of resources for sake of extending battery life. For this reason, the Dynamic Voltage Scaling (DVS) has been introduced (Chen and Kuo 2005; Matic et al. 2006; Pillai and Shin 2001; Quan and Hu 2001; Seo et al. 2008; Shin et al. 2000). For the resources with discrete-state power consumption, the power scale is divided into finite number of set. On the other hand, continuous-state power consumption provides capability for a resource to adjust the speed scale of the resource continuously within a range.

Robj: Because of the individual performance requirements of the resource, we defined the resource-related objectives such as minimizing power consumption, maximizing utilization, etc.

A Model for the Scheduling Characteristic

In this section, we present a model for the parameter C (scheduling characteristics) in Definition 2. Our model, which is shown in Definition 8, extends the models presented in (Buttazzo 2011; Davis and Burns 2011; Sha et al. 2004) for the sake of expressivity. The corresponding sub-feature model can be found in the link1:

\[ C = (C_{\text{type}}, C_{\text{pmin}}, C_{\text{mig}}, C_{\text{pol}}, C_{\text{reso}}, C_{\text{prio}}, C_{\text{window}}, C_{\text{obj}}) \]  

(Ctype): It refers to the type of the scheduling process. It can be either offline or online. In offline scheduling, the resources are allocated to the tasks within predefined time interval; while in online scheduling, the most convenient tasks are chosen for each resource at instant of time. For both, complying with the constraints and satisfying the objectives explained in the subsection where the scheduling objectives were defined are the main concerns.

(Cpmin): It shows whether the scheduling process is preemptive or not. Although tasks are specified as preemptable, the scheduler produces non-preemptive solutions unless it is preemptive.

(Cmig): It specifies whether the scheduling process supports migration or not. The migration is the capability of the scheduler to suspend a running task on a resource and moving its execution context to another. There are two migration strategies, namely task-level and job-level. While task-level migration supports migration of only the task instances, but each task instance must be executed only on the same resource, job-level migration does not have such a restriction if it is also preemptable.

(Cpol): It represents the scheduling policy, which is a criterion to decide on which task takes the permission to utilize a resource at an instant of time.

(Cres): It refers to the time resolution of the scheduling process. It is the minimum time interval within which resources are allocated to tasks. In online scheduling, it refers to time quantum (Silberschatz et al. 1998). In offline scheduling, it represents the possible context switch (preemption) points in time.

(Cprior): It defines the strategy about priority assignment. It may be either fixed or dynamic. In fixed-priority assignment, the tasks have static priorities that do not change in time; whereas if the policy depends on the attribute varying from one task to another, such as deadline, the priorities are recalculated based on the scheduling policy.

(Cwindow): It corresponds to the scheduling window whose boundaries determine the duration of the time interval when the tasks are scheduled. Specifically, it equals to the time resolution in online scheduling.

(Cobj): It corresponds to the objectives of the scheduling process. It directly relates to the objectives defined previously in the subsection Scheduling Theory.

A Model for the Scheduling Strategy

In our domain model, a scheduling strategy is termed as Solver. A particular solver accepts the scheduling specification expressed in a given format as input and produces a schedule in a given format as output. To foster reuse, we aim at adopting existing solvers within our design environment.

A feature model representation for scheduling strategies is shown in the link2:

We formulate a model for scheduling strategies as:

\[ S = (S_{\text{solver}}, S_{\text{in}}, S_{\text{out}}) \]  

References:


where \( S_{\text{solver}} \) is a set of solvers; \( S_m \) is the specification of the scheduling problem; and \( S_{\text{sol}} \) represents the schedule.

The sub-features of \( S_{\text{solver}} \) from left to right \( \text{SCIP} \) to \( \text{Walksat} \) represent the supported solvers in our framework. Other corresponds to a future solver implementation that can be plugged into the system. The dedicated sub-features of \( S_m \) and \( S_{\text{sol}} \) which correspond to the adopted solvers are not shown in the figure. This is, because, in our framework a generic input/output language is used for all solvers. Translation to a specific language is encapsulated.

EXPERIMENTS

This section demonstrates how the feature model can be configured to express 7 different scheduling examples and one optimization example. Each example is presented in the following order: (i) Explaining the problem; (ii) Expressing the problem using the traditional notation formalized in Definition 1; and (iii) Presenting the example as a configured feature model. The motivation for using this notation is due to its common usage in the field of scheduling theory. Since the resulted configurations are too large to display in this paper, associated with each example, we give links to our repository where the corresponding feature models can be found.

Rate Monotonic Scheduling (RMS) Problem

According to the articles (Lehoczky et al. 1989; Liu and Layland 1973), in the problem RMS, the tasks are assumed to be preemptable and have various release time requirements. Furthermore, there exists only one resource with single-unit capacity and the optimality criteria is to minimize the lateness penalty for each task. Based on this definition, the 3-field notation of this problem is as follows:

\[
1 \mid r_j, pmtn \mid L_j \quad (10)
\]

This example can be expressed directly by our feature model. The configuration is given in the repository\(^3\).

Multiple Resource Scheduling Problem (MRSP)

In this problem, unlike previous examples, we have two tasks and two resources. The tasks are either periodic or aperiodic. While the periodic tasks are not preemptable, the aperiodic tasks are preemptable. In addition, the aperiodic tasks depend on the periodic tasks with setup time. We assume that there are two active and two passive resources. The active resources have fixed-state scalable power consumption; whereas, the power consumption of the passive resources is assumed to be constant per capacity. Furthermore, the passive resources can execute more than one task at a time with respect to a given capacity (batch).

From the perspective of scheduling process, the scheduler is offline and preemptive. Moreover, it has the capability of job-level migration. The scheduling policy is First-In-First-Out and the priority assignment of the tasks is dynamic. The objectives are to minimize the power consumption and lateness, separately. The definition of this problem is:

\[
Q2 \mid r_j, d_j, prec, pmtn, M_j, sjk, batch \mid L_j \quad (11)
\]

The example can be directly expressed by our feature model\(^3\).

Elevator Scheduling Problem (ESP)

In this example, we assume that we have two cars (elevator cabins). Therefore, the scheduling process consists of two phases: (i) dispatching the passengers to the cars (passenger-to-car assignment) and (ii) stopping cars on the corresponding floors according to the requests of passengers (Koehler and Ottiger 2002; Seckinger 1999). Tasks are categorized into two groups, namely Car- and Hall-Call. The former is the request of a passenger inside a car to go to the desired destination floor index; whereas the latter is the request to indicate the desired travel direction of a passenger on the floors using \textit{UP} and \textit{DOWN} button calls. To ease the computation, we assume that moving between successive floors and the time resolution are unit time. Since the elevator should visit the closest destination floor first, the scheduling policy is set to Shortest Job First and the priority assignment is dynamic.

In this example, we have defined our objective is to minimize the waiting time of the passengers. Therefore, minimizing the total lateness objective is chosen. According to the specifications above, the problem domain can be formalized as follows:

\[
P2 \mid r_j, d_j \mid L_j \quad (12)
\]

The configured feature model can be found in the repository\(^4\).

Flow-shop Scheduling Problem (FSP) with Permutation

In the subsection \textit{Machine Environment}, the \textit{flowshop} machine has been defined. Unlike ordinary flowshop problem, the permutation constraint inhibits the sequence changes of tasks between machines (Pinedo 2008). The execution orders of the tasks are the same on different machines. The corresponding definition of the problem is as follows:

\[
F4 \mid prec, prmu \mid C_{\max} \quad (13)
\]

The configured feature model can be found in the repository\(^5\).

---

3. \( https://github.com/gorhan/LFOS/tree/master/FM_LFOS/RMS \)

4. \( https://github.com/gorhan/LFOS/tree/master/FM_LFOS/MRSP \)

5. \( https://github.com/gorhan/LFOS/tree/master/FM_LFOS/ESP \)
Job-shop Scheduling Problem (JSP)

JSP is one of the well-known combinatorial optimization problems (Cheng et al. 1996; Garey et al. 1976). We have defined four machines and three tasks with 3, 4, 3 instances each of which requires one of the machines. Since the execution order of the task instances on each machine is predefined, there exists natural precedence relation between tasks. The aim of the scheduling process is to minimize the makespan. According to this scenario, the problem definition is as follows:

\[
J4 | M_j, prec | C_{max} \quad (14)
\]

The configured feature model can be seen in the repository\(^6\).

Open-shop Scheduling Problem with Preemption (OSP/PMTN)

Since the Openshop machine environment has been explained in the subsection Machine Environment, there is no need to mention its explanation. In this example, we assumed that the tasks are preemptable, and so the scheduling strategy, and we have three resources and 5 tasks each of which has 3 instances assigned to the machines one-by-one. Each task instance has its own release time and deadline. The objective is to minimize the Makespan.

The definition of this problem is:

\[
O3 | r_j, d_j, M_j, pmtn | C_{max} \quad (15)
\]

The configured feature model can be seen in the repository\(^6\).

Open-shop Scheduling Problem without Preemption (OSP)

Only the difference of this example from the previous one is the preemption capability of the tasks. Therefore, the definition of the problem using the traditional notation is as follows:

\[
O3 | r_j, d_j, M_j | C_{max} \quad (16)
\]

The configured feature model can be seen in the repository\(^6\).

Travelling Salesman Problem (TSP) as an optimization problem

TSP is an optimization problem (Flood 1956) (also known as a path planning problem). The aim is to find the minimum traveling path for a salesman who is visiting each city on a route and finally arriving at the departure city. We need to map the concepts of this problem to the concepts in the scheduling domain. We model the traveling cost between two cities as sequence dependent setup time. Furthermore, salesman’s worst case staying time in a city is represented as execution time of the corresponding task. Finally, for simplicity, we assume that all tasks have the same execution time.

The objective is to minimize the criteria Makespan (Butazzo 2011). The corresponding problem definition is as follows:

\[
1 \mid s_{jk}, p_j = 1 \mid C_{max} \quad (17)
\]

The configured feature model can be seen in the repository\(^10\).

EVALUATION

In this section, we assess our approach explained in three previous sections with respect to the objectives described in the section Objectives.

Complexity of Developing Scheduling Software

For the software engineer, implementing software systems that incorporate scheduling systems can be experienced as a complex process due to the following reasons. First of all, the software engineer has to define and implement the related tasks, resources, associated parameters, objectives, strategies, and the constraints related to these. Due to the inter-dependencies, this can be a complex process. For example, the software engineer may need to define and implement the following constraints in a very precise and robust manner: The tasks have to be scheduled within their life-scope; the periodic tasks have to be spawned at each inter-arrival time; the resource requirements of the allocation have to be realized for each task; the precedence relations have to be satisfied for each allocation; the capacity constraints of resources have to be satisfied; the preemption capability is supposed to be realized; the migration capability has to be satisfied; the mutual exclusion constraint among resources have to be satisfied. Dealing with all these concerns can be a very time consuming and error-prone task.

Our Assessment Method

In general, feature models form the backbone of any PLE process as they offer the necessary abstractions that derive the various implementations in the domain of the PLE

---

https://github.com/gorhan/LFOS/tree/master/FM_LFOS/FSP

https://github.com/gorhan/LFOS/tree/master/FM_LFOS/JSP

https://github.com/gorhan/LFOS/tree/master/FM_LFOS/OSP_PMTN

https://github.com/gorhan/LFOS/tree/master/FM_LFOS/OSP

https://github.com/gorhan/LFOS/tree/master/FM_LFOS/TSP
approach. From this perspective, the “quality of the feature model” is one of the important factors that determine the “quality of the PLE approach”. Secondly, the “quality of a PLE approach” is also determined by the configurability of the feature model and ease of implementability of the configured model.

The method of evaluating our feature model is, first, based on adopting a thorough domain analysis, and matching the model and the available background knowledge in the scheduling domain. Furthermore, as presented previously, the feature model is configured with respect to the canonical applications in the scheduling domain. By this way, it is aimed to demonstrate that the model is capable of adequately expressing the known examples. In parallel, with the help of our wizard, scheduling implementations are generated.

**Dealing with Complexity**

In our approach, the tool *Framework Configuration and Instantiation Wizard* supports configuration of a product. As an implementation of this tool, we have developed an object-oriented scheduler framework based on various solvers. The constraints that are associated with the scheduling problem are handled internally so that the software engineer does not need to concern about defining the constraints of the problem in specific format; only the required parameters must be defined as a configuration of the feature model. Obviously, all these possibilities reduce the complexity of developing scheduling applications compared to implementing these from scratch.

**Expressivity**

Naturally, the adopted feature model notation and the associated tools support variability in the domain model. In the previous sections, we derive a generic model that is capable of expressing a large category of different schedulers and present a set of illustrative examples from the scheduling domain to demonstrate the expressivity of our model. To justify the coverage of these examples, we first refer to the 3-field notation introduced in Definition 1, which is used in the literature as a common notation to express scheduling problems. Each example given in this paper is explained using the 3-field notation.

Table 2 depicts the scheduling domain and examples are shown in the table as numbers in cells based on the corresponding subsection number. The parameters in the rows of the table corresponds to the *task characteristics* and the set of all these parameters are denoted by \( \beta \); whereas the columns of the table are grouped into two categories, namely the *machine identifier* and the number of machines which are denoted by \( \alpha_1 \) and \( \alpha_2 \), respectively. Within these categories, the parameters \( P, Q, F, O, J, 1, M \) are explained in detail in the subsection *Machine Environment*. As also argued in the literature (Brucker and Knust 2009), the scheduling domain can be represented using these parameters as shown in Table 2. As it can be seen, in each column and row from top to bottom and left to right, respectively, at least one example resides. This illustrates that the examples cover at least one case of the parameters in the scheduling domain. Since the examples are derived from the feature model, we claim that our feature model is expressive enough to cover a large category of scheduling applications.

**Reusability**

Like all SPLE methods, our approach as presented in
the section Our SPL Approach is based on a method, which aims at enhancing reuse. In addition, the feature model and the associated tools are eligible in expressing commonalities and variabilities in the domain. As illustrated by the examples, our feature model is shown to be expressive. Naturally all these characteristics are considered necessary to foster software reuse. Furthermore, configuration of a software product from the feature model is supported by the tool Framework Configuration and Instantiation Wizard, which is defined on the top of an object-oriented scheduler framework based on various solvers. The exact percentage of reuse in a particular software system that incorporates schedulers naturally depends on the requirements of that system. The feature model as presented in this paper only supports the scheduling components of applications. The application specific remaining parts of software, naturally cannot be supported by our feature model. For example, in case of facility scheduling at airports, there may be many software components representing planes, runways, gates, staff members, etc. are not directly supported by our tool. New feature models dedicated to the application of interest, of course, can be developed and integrated with our tool if necessary. In the examples shown in the previous section, since our main focus is to demonstrate the expressivity of our feature model, the application-specific code is kept minimum. Application-specific requirements are mapped to scheduling concepts such as tasks and resources.

We will now elaborate further on the reusability of code by first referring to the number of lines of code (LOC) metrics. There is much debate on the preciseness of this metrics. Nevertheless it gives a reasonable indication for our purpose, because as it will be shown the reuse percentage in our approach is much higher than the expected error margin of this metrics.

From the perspective of the code size, our SPL repository can be divided into three major components: EB, code that represents the essential abstractions in the scheduling domain, such as tasks and resources; AD, the adapter, the software module that handles the constraints which are generated by our framework and couples these to the solver; and finally SOL, the solver, which analyzes the constraints provided by AD, and computes a schedule if possible.

The LOC metrics of our SPL includes Python and C/C++ code. In AD, there exist 14856 and 335963 LOC implementation written in Python and C/C++, respectively. SOL has been developed using C/C++ in 591941 LOC. Finally, our EB implementation consists of 3215 lines of Python code.

In Table 3, the columns and rows show the examples given in the previous section and their characteristics, respectively. The first column indicates the LOC metrics of the example programs.

By comparing the LOC metrics for AD, SOL and EB; and the row in Table 3, we can see that the additional code that is necessary for expressing the examples are much less than the reused code. The relatively high number of LOC in the elevator example (ESP), is due to addition of several application-specific classes to the implementation such as elevator, hall and car calls, etc. In our approach, high-degree of reuse is naturally provided for the scheduling part of the software system. Detailed information can be found at our repository whose link is given as a footnote in the subsection ESP.

**Verification**

In the section Objectives the verification requirements were defined in twofold: models must not violate the constraints of the domain, and the product must be able to schedule the activities as desired.

To detect whether the configured application violates the domain constraints or not, the model of the application is checked using Clafer (Antkiewicz et al. 2013). The complexity of the model-checking process mainly depends on the characteristics of the feature model and as such it is more or less independent of the examples given in the section Experiments. We have verified the configurations of all the examples and in case all variabilities are bound. Approximately 0.25 seconds are needed for the verification of the configured applications using the MacBook Pro computer on 2.6 GHz Intel Core i5 processor and 8 GB 1600 MHz DDR3 memory with MacOS X 10.9.5.

To determine the schedulability of the activities, the scheduling problems are translated into constraint specifications and verified using appropriate solvers (David et al. 2009).
In Table 4 the first row indicates the complexity of the scheduling problem by referring to the number of constraints generated by the framework. The second row indicates the execution time metrics of the solver necessary to compute the schedule.

The number of constraints in the first row of Table 4 varies considerably between 720 and 13205, based on the complexity of the scheduling example, which is proportionally related to the dependencies among the tasks. For example, as it is stated in the literature (Brucker and Knust 2009), within the context of three well-known examples, the number of task dependencies and consequently the number of constraints to be solved are known to be lower in open-shop (OSP), higher in job-shop (JSP) and the highest in flow-shop (FSP) problems. We observe this fact also in the first row of Table 4 where the number of constraints are subsequently 770, 2199, 13205. We consider the time necessary to compute the desired schedule in FSP is acceptable because in general such kind of a scheduling problem is applied in off-line production processing planning which does not have to change frequently. The number of constraints for the example ESP (Elevator Scheduling Problem) is stated to be "not fixed" (NF) because the number of tasks and consequently the number of constraints to be solved change dynamically with respect to the elevator calls and the position of the car.

The second row of Table 4 indicates the time necessary for the solver to compute the desired schedule. If the solver does not give a result, this means that the problem is not schedulable. As it can be expected, the number of constraints and the time necessary for the solver to compute the problem is related. However, higher number of constraints does not always result in higher computation time for the solver. If the certain constraints are interrelated, such as in the case of MRSP, the time needed for the solver may increase accordingly.

Implementation Aspects

Our design environment is a comprehensive software system which incorporates three kinds of software subsystems: (i) The sub-system that implements the essential building blocks (EB) in the scheduling domain, such as tasks, resources, scheduling characteristics and strategy, as explained in the section where the domain models were presented; (ii) Third-party software that is used in the scheduling process (AD + SOL) and as such together with EB, they form part of the SPLE repository; (iii) Third-party software which is adopted to implement certain supporting functions such as input-output handling, visualization, etc. These are not considered part of the SPLE repository and as such they are not taken into account while reasoning about the degree of reuse in the subsection Reusability.

We have implemented the subsystem EB using the Python language (van Rossum and Drake 2011). The third-party software which constitutes AD + SOL are Numberjack (Hebrard et al. 2010), SCIP (Gamrath et al. 2016), MiniSat (Sörensson and Een 2002), MipWrapper (Hebrard et al. 2010), Mistral (Dillig et al. 2009), Mistral2 (Hebrard et al. 2010), SatWrapper (Hebrard et al. 2010), Toulbar2 (Cooper et al. 2010) and Walksat (Selman et al. 1995). To implement the supporting functions, we have integrated the following third-party tools: Clafier (Antkiewicz et al. 2013), Alloy (Jackson 2012), Chocosolver (Prud’homme et al. 2016) and PyPlot (Hunter 2007).

A guide for the implementation of schedulers is generated by our wizard tool. Interested reader can refer to our repository. We have used the same abbreviations for the scheduling examples while naming files except ESP located under another directory. The framework implementation (AD + SOL + EB) can also be found in the link.

CONCLUSIONS AND FUTURE WORK

We have introduced a domain specific model to manage the complexity in designing scheduling systems. Furthermore, our SPLE method and the associated tools hide the coding details from the software engineer. To assess the expressivity of our model, in the section Experiments we have provided a number of canonical examples. We have presented LOC metrics of reuse per example given. Also, the adopted verification tools and their scalability are presented in that section.

Currently, the software is instantiated manually based on a guideline that is generated by the Framework Configuration and Instantiation Wizard. As the future work, it is the intention to implement a more interactive application generator incorporating a model-directed editor. In addition, we plan to incorporate our SPLE method, the associated software repository and the tools in a category of radar systems produced by the company Aselsan A.S. Furthermore, we are in close cooperation with a company in applying our method to designing elevator systems.

ACKNOWLEDGMENT

This research has been cooperatively carried out by the company Aselsan A.S. and University of Twente, and partially-funded by Aselsan A.S.

REFERENCES


11 https://github.com/gorhan/LFOS/tree/master/Tests
12 https://github.com/gorhan/LFOS/tree/master/ElevatorScheduler
13 https://github.com/gorhan/LFOS/tree/master/LFOS
14 http://fmt.cs.utwente.nl/research/projects/aselsan/
scheduling of aircraft arrivals at congested airports, in ‘2006 IEEE Aerospace Conference’, pp. 11 pp.–.


COMPONENT INTEGRATION THROUGH CONNECTOR SUPPORTED PROCESS MODELS

Anil Cetinkaya, M. Cagri Kaya, Ali H. Dogru
Department of Computer Engineering, Middle East Technical University
Ankara, Turkey
{ cetinkaya@ceng.metu.edu.tr kaya@ceng.metu.edu.tr dogru@ceng.metu.edu.tr }

ABSTRACT

Invocation times for messages among components are decided by a central process that assumes the composition responsibility and arbitrates the message traffic via connectors. This research proposes connectors in component-oriented development to be enhanced with the capabilities for the required synchronization. Existing component technologies are used without modification and separation of concerns is provided for the central process, components, and connectors. Recently a study laid down the connector types and categories that imply further expected capabilities that are preferred to be offered in this research through partially completed connector instances. Also loaded with late-binding variability management properties, connectors come close to components where reusable and readily available connectors can be part of software development. Suggested connector structure includes slots for registering the identification of the two components on either side of a message, along with their required and published methods. The main mechanism operates upon the reception of a Triggering signal from a central process model that synchronizes with the requesting component and relays the request to the responder component. Besides implementing the required infrastructure in a language, an example application is demonstrated.

INTRODUCTION

Coordination is an important issue in synthesizing software intensive systems leveraging on components. Given the same set of components, different flow specifications can yield various systems through coordination of existing components. A central flow model used in the ordering of message invocations is a powerful mechanism that should be further supported in an effort to provide the developers an effective integration tool.

Coordination has been addressed at different levels and recently connectors started assuming more responsibilities in this direction. A central control model is promoted in this research, that offers a top-down view for coordination while not limiting other alternatives. That also accommodates different modes of elementary coordination tasks for example bottom-up. Here Top-down mode is defined where a message is triggered by the central flow model and the component collaborates. Bottom-up is the reverse mode where a component starts a request. These modes can also be viewed through an analogy to Poll based (top-down) or interrupt based (bottom-up) communication.

A central control implies the desired separation of concerns such as the components not being aware of the algorithm they are used for. Actually, one component can be utilized in the solution of more than one algorithm in a system. Also, they should be used without code-level modification – some configuration is allowed anyway, in general component utilization.

A distributed control, on the other hand makes it possible to employ distributed algorithms that are more powerful. However, the components when taking part in the implementation of distributed algorithms need to be solution aware: every component knows what problem the collaborative system is solving and their local duty that requires knowledge about the other components. The cohesive nature of components is therefore sacrificed and components become system dependent modules rather than off the shelf and more reusable units.

In recent study (Cetinkaya, 2017) variability in connectors was offered as a capability that is leveraged in this research. Besides the implied duties that surface in the classification (Mehta et al., 2000) a configuration mechanism to comply with variability management, connectors become versatile building material for integration.

Serving different Coordination Needs

Considering a banking operation using an ATM, some requirements can be demonstrated. A central coordination, trying to arbitrate the message traffic for ordering the intermediate actions for money withdraw operation would prefer a prescribed order of activities. Those activities will be controlled through message invocations. A typical scenario could be ordering the activities such as user authentication, function selection, checking the balance, and checking the amount of cash in the ATM. However, the authentication cannot be started by the central flow control. On the contrary, user starts this activity by inserting a bank-card to the ATM. This is an example where a top-down control needs to work with a bottom-up invocation.

Our basic solution for such incompatible interaction needs is the basic synchronized communication definition that is “sender blocked”. However, we define two different initiators for an interaction where both could be assigned the blocked role. In general, any message is regarded as an
activity requested by one party and responded by the other. However, in our case there are two requestors: The coordinating authority and the message sender. Once both are on, the activity can start: the responder can be invoked.

The sequence of events for synchronizing the method request and invocation can be modeled as in the following rules according to an event algebra. According to this representation (Kaya et al. 2010), the condition part (left) of the rule suggests that the event to the left of the “;” symbol should have happened before the one on its right. If this is true, the action on the right of the arrow will be triggered. The two cases correspond to the request or the invocation happening first where r stands for the request, i stands for invocation, and a is for activation of the method:

\[ r ; i \rightarrow a \]

\[ i ; r \rightarrow a \]

This means the central control cannot decide about the activation time, if the requestor is not ready. However, the later one between the request and invocation starts the activation. It is possible therefore to allow the connector to represent this compromised synchronization:

\[ (r ; i) \mid (i ; r) \rightarrow a \] (1)

Following the ATM example, the central control would be waiting for an external event by enabling it: An invocation for user authentication will result in the system waiting for the users inserting the banking card. Once the card is inserted, the authentication action will immediately be started.

**Bottom-up coordination of an interaction**

Proposed connectors are equipped with methods that can be called from either party (the central process or a requestor component). By representing any kind of communication needs in the methods, different modes can be served. However, the central flow represented in a process model should be capable of working in different modes. A component trying to start an activity by informing the center can be responded in an interrupt manner. The central process should now have published methods that can be called from outside (bottom).

Actually, process models have the required capabilities and can be adapted to any mode. However, the system integration should be made simple and consistent. If there is a real requirement for different modes of communication, then it is inevitable that the process model will be a complex one allowing different patterns of flows (Ambler, 1998).

As a summary, the composition of components is conducted through a process model specification as well as the incorporation of new connectors with a heavy protocol. Implementation details will be included in the later sections after the introduction of related concepts.

**Components**

A software component is a piece of software which is implemented according to a component model. Components are usually pre-implemented. They can be used independently as well as integrated with other components to yield larger and more complex components, and eventually systems. Thanks to their technology that is offering reusability, they allow creating software systems with reuse instead of creating from scratch.

Components can be considered as structural pieces of software systems. Each component includes interfaces to show its provided services and the services required by the component to work properly. Thus, the functions and behaviors can be pre-defined at some abstraction level. Components are the fundamental enabling technologies for “build by integration” (Dogru and Tanik, 2003).

**Connectors**

Most of the modern software systems consist of many complex components. Management of interactions among these components is a crucial task to ensure stability and efficiency of the system. In traditional systems, handling of these interactions is done through components themselves. Adding communication responsibilities to components in addition to their expected functionalities is against the separation of concerns principle and makes it difficult to manage the system development.

Connectors are defined as “architectural elements tasked with effecting and regulating interactions among components” in (Mehta, 2000). Therefore, connectors can be considered as architectural abstractions that are responsible for the interaction among components. In simple terms, they are used for transferring control and data among components.

Mehta et al. define coordination as “transfer of control among components” and assumes it as a responsibility for connectors. Coordination hence is categorized as a service type for connectors. In their classification, connector types that provide coordination are procedure call, event, and arbitrator.

**XCOSEML**

XCOSEML was first introduced in (Kaya et al., 2014) as a textual version of COSEML with variability capability. In the same study, dynamic constructs were added to COSEML. Dynamic constructs were a composition specification, and atomic and composite interactions in that composition. It was an adaptation of service choreography inspired from (Suloglu, 2013).

In the following work, XCOSEML was extended by adding detailed connector definition and variability (Cetinkaya et al., 2016). The earlier version of connectors as defined for COSEML was abstract and contained no processing (code) in them. According to this recent work, connectors are conforming to the classification of (Mehta et al., 2000) and have a similar text-based style defined in (Oussalah et al., 2004).

The dynamic specifications introduced by XCOSEML to its predecessor, is actually modeling a process. The overall process model of a system can be compiled from the interaction declarations contained in the “packages” that can be regarded as a super component. A package integrates smaller components by means of interactions. Despite the
capability to represent a process model inclusively in the XOSEML specifications, it is also possible for the components to work with an exclusive process model. Tools supporting executable process models make it easy for such representations to interact with any kind of software at run time.

SUGGESTED APPROACH

This research is supporting the goal of composing big software intensive systems without code writing. In general, efforts in line with this concern can be classified into two approaches: compositional and generative. While mostly Model Driven Approaches define the modern efforts in the generative category, more reuse based approaches define the compositional category. Recent developments along software architecture all support the compositional approach more or less. Some examples are Software Product Lines (SPL), Service Oriented Architecture (SoA), and Component Oriented Software Engineering (Dogru & Tanik, 2003).

None of the new approaches offer a methodology for exclusively “build by integration” as an alternative to code writing. However, the set of technologies developed in support for such approaches are very suitable for their utilization towards our goal. The variability management concept is inspired by the SPL approaches. Utilizations of a central process model and invocation of services in a top-down manner are inspired by the SOA.

Variability and Domains

As investigated in SPL, variability is one of the two main concepts in the modeling of a domain. A domain accommodates a set of products all of which are defined with their common and variable parts. Whereas the variability concept constitutes the most important specification component for the “next product”, this concept also provides a powerful mechanism for our connectors. Connectors include in their configuration capability a foundation for reuse, and late binding capability offering more flexibility. Consequently, connectors become like components in the composition of systems – they can be selected from an existing set and modified for integration. Some degree of modifiability and application domain scoping are the necessary conditions for “off the shelf” selection of a systems constituents. Those constituents used to be only components whereas now connectors are treated likewise.

As in the case of SPL, as a domain oriented approach where the software assets are enumerated, connectors need to be classified within a finite list before they are offered for reuse. In a specific SPL, a domain of application needs to be determined and the environment should be populated with software assets (such as components). This pre-determined set of assets corresponds to the scope limitation of a domain: As in the generative case which actually implements automated code generation, success depends on the narrowness of the scope. Automated code generation is not possible as a silver bullet today for any kind of requirements. Rather, an application area needs to be determined where the tools can decide with intelligent capabilities, based on previous solutions and existing concepts, to substitute a notion in the problem domain to a structure in the solution domain. A finite set of elements together with their adaptability has been the requirement for usable component libraries. The same should be valid for a usable connector library.

The domain concept is also an important foundation for either the compositional or the generative approaches. As a domain is more matured, with its tools improving and a critical number of engineers having command over the concepts, tools, and assets defined for the domain, development becomes easier, faster, and more predictable, suggestively improving quality in many directions. For effective reuse, business domains have been attractive for the developer: bringing our models closer to the problem domain rather than the solution domain.

Business domain empowerment has been more successful in some levels of development and not as much in some other levels. Components, by definition have been more application-field oriented, such as the web services utilized in the SOA approach. A component is different from a web service in how it is implemented and integrated in the application. However, it will not be wrong to generalize a web service as a kind of a component, ignoring the technology and concentrating on the web services offering a partial solution.

On the other hand, Domain Specific Languages (DSL) started having a wide application area, as another domain related tool. Unfortunately, the early use of DSLs has been more effective in “technological domains” rather than business domains. There could be numerous DSL constructions in different organizations, which are only used in that locality with the additional problems such as the maintenance and training for a multitude of different languages. However, technological domains have been luckier for being established and offering wide use. For example, SQL is a DSL that serves the relational database domain. However, this is a technology domain that can be used in many different business domains. From one perspective, SQL can be evaluated as offering high levels of separation of concerns, from another one, it becomes like a general purpose language rather than being domain specific.

Form this point of view, connectors will be on the technological domain, for connecting components that are more business domain centric.

It is quite possible though, for sets of domain specific connectors to surface. Coordination and even adaptation may sound generic problems in the beginning. However, those functions even can have domain specific enhancements and common usage. It is quite possible therefore to confront in the near future, “toolbars” of connectors as well as components in domain specific development environments.

The technological domains for the connectors are two in this research, with the message relaying and arbitration being one and adaptation being the other. The adaptation requirements are implemented as suggested in the definitive
study (Cetinkaya, 2017) and the messaging infrastructure is implemented to serve the three-port architecture with the invoker, requestor, and responder ports. This architecture serves for any kind of communication needs especially due to the fact that those ports are supported with methods that can include any kind of code. However, in this research we are promoting a central control for the cases where there is no complex coordination but there is a big demand for business software.

**Methodology**

A compositional approach is assumed in this study. The software units coded as components will be integrated for the application to emerge. Connectors will be used in connecting two components, or a component to the central process. The central process is like the main function, if an analogy is to be made to a program written in the C language.

It has to be recalled that such an approach has to be domain-centric. Having a well-established set of components at the disposal of the developers is a beginning point. We also suggest the existence of a well-defined set of connectors. Both components and the connectors are subject to modification, as reuse requirements should dictate. Modification here is implemented as the realization of the variability specifications.

Having a good understanding of the components at hand, the developers should design the central control in terms of a process model. The activities in this model should be implemented by the methods of the components to be integrated. Depending on the kinds of components, their connections should be served by a connector, selected for any two-component connection.

Basic activities for composing a system are listed below. They are further explained in the following sections.

1. Process modeling for the central flow specification
2. Process decomposition
3. Component matching
4. Connector matching
5. Verification

**Process Modeling**

Usually applications contain different system functions that do not have much of a common process: once a system function is selected, a process specific to that function is enacted. For such cases, the first activity group, or actually the first process pattern corresponds to a conditional multi-way decision point (branching). An analogy to the more familiar UML modeling can be made: the top-level entities as followed from requirements are the use-cases declared in a use case diagram. The use cases correspond to system functions where no ordering information is usually defined among the activations of those system functions. Actually, it was not even possible in the earlier versions of UML to combine those system functions. With version 2.0 the option to combine the system functions in a scheme that allows ordering was offered: Use cases can be detailed in sequence diagrams that can be modeled as activity nodes in a process model – that is the activity diagram in UML. This new diagram is called the interaction overview diagram.

Overall, the process model to specify the central flow corresponds to Use Cases and higher-abstraction information in the collaboration models (drawn using collaboration/communications or sequence diagrams) used in UML. Later steps in our methodology concerning the connection of components relate more to the messages in the collaboration models.

If the development takes place in an SPL conscious organization, it is possible that early models for the specification of the product are some kind of instantiations of a more general domain model. The classic domain model would involve “features” most probably paving the road for a static model as they are a good starting point in the determination of the set of components to be mapped. The variability information in the feature model (or provided in a complementary variability model) will be used in the selection among alternative sets for components. For variabilities of less magnitude, configuration of components will do the job. Later work that introduced variability in the process models also made it possible to apply the SPL philosophy to dynamic models. Therefore, it is possible to start with process modeling that already was partially offered by means of its pre-existing constituents in the domain model. As a result, as in most of the steps of the methodology, the process modeling step also includes variability resolution. In its simplest form, variability resolution in the process model can be explained as selecting among the sub-flows of a decision point for the inclusion or exclusion of those flows.

**Process Decomposition**

As in the analogy made above, after starting with use-case-level information, more details of ordered activities is required. This all by itself defines decomposition. A top-down specification is naturally fit for the goal. While defining the process is corresponding to “what to build” in its earlier stages, our domain oriented advantage makes it possible to decide about “how to build” not after too long (Dogru, 2006): this latter concern corresponds to the “leaf-level” activities in the process model. Domain-experienced developers should be concerned about specifying tasks in the process model that can be matched with existing methods in the available set of components. In a mature domain most of the domain process model should be expected to correspond to existing solutions anyway.

Actually, some process composition may be foreseen during the overall process model defining activities: a partial solution offered in the domain could be copied and inserted in a different location due to an expectation in the design that allocates a different execution time than a default.

Anyways, before component connections, a process model is finalized where tasks can be connected to existing code. Now it is time to conduct detailed selection, configuration, and connection over the components.
Component Matching

Most of the tasks in the process model would require executable code for enactment. Those tasks should be taken one at a time and matched to the desired code by assignment of a suitable component. A published method in the component would provide the required functionality (code). There may be cases where a component cannot directly offer the requirement, where the variability management will be recalled to aid in the configuration of the component. Components in their fundamental descriptions carry such a modification capability (Szypersky, 1998).

For configuration of the components, the configuration interface is used. In a well-supported environment, higher-level variability information should propagate to lower-level ones and eventually affect those interfaces. A completely supported variability propagation capability will end in the automatic configuration of the whole system. Of course, this is a complex task because variability management also requires the constraints among different variation points to be resolved. Every selection made by a developer could change the landscape of the allowable options drastically, system wide.

Our suggestion proposes a more process model based selection of the components. However, it is possible due to the popularity of the SPL techniques that developers may directly decide on an alternative component based on the feature model, only considering the static view before the dynamic view is considered. There is room for such freedom, however bringing further constraints in to play.

Connector Deployment

Even if not in a methodology proposed here, parts of the activities defined above were employed more or less in existing approaches. Inclusion of connectors becomes the important contribution in this research.

After a component is matched, it needs to be connected and the connection is always through a connector. Any pair of components that require connection, will require at least one connector. While connectors are conducting heavier duties than only conducting one message, every major service request between the same pair of components may require a separate connector. Connectors are basically two-port structures for the execution of the functionality in terms of the requestor and the responder for the function, a third port is also used for the triggering party that would for majority of the cases be the central process model. On the other hand, many connections would require one of the pair of ports to be the central process, ending with the requestor and the trigger being the same party (the central process). What is left is only the responder role assuming the duty of a requested service to be conducted by another component.

There will be cases where a component could invoke another component without the central process being involved. We do not want to emphasize these connections in the main model however such segments of the model should be available in related views. The suggested connector structure would serve such cases also.

A typical deployment of a connector involves the following steps:

1. Select the category and type of the connector
2. Determine the synchronization requirements
3. Declare the connection points for published and required methods
4. Apply variability resolution for fine tuning

Finally, any special handling could be served through code development for the internal methods of a connector. We try to avoid code writing completely if possible. However, a domain matures with more exceptional developments in the beginning and less, as maturity improves. Connector structures to complete the spectrum of structures in the constructive approach, are new constituents. There may be unpredicted expectations that may yet be very common in their infancy. Such enhancements will take place later as part of the pre-defined variants, if proven useful.

CASE STUDY: E-COMMERCE SYSTEM

An e-commerce system is developed to demonstrate the capabilities of the proposed approach. The system consists of a customer, an e-shop, a bank and a supplier. The system environment is shown in Fig. 1. Coordination of the components is managed through the connectors via the central process.

![Fig. 1 The e-commerce system environment](image)

The elements used in the system are shown in Fig. 2. It is shown that the “customer” component starts the interaction by requesting an item from the “e-shop” component through the related connector. When the “e-shop” component receives the request, it invokes the connector that is responsible for the communication between the “e-shop” and the “bank” components to check if the customer’s payment information matches his/her credentials in the bank database for security reasons. In this phase, requested items price is held in the connector by using a buffer (such as a stack or a queue) until the connector is invoked by the process to finish the transaction. Transaction fee will be held by the bank until a confirmation or a deadline. In the meantime, the “e-shop” component interacts with the “supplier” component through the corresponding connector to check the availability of the requested item. If the item is available, the “e-shop” component will be provided with the requested item by the “supplier”. Then, the connector between the “e-shop” and the “bank” will be invoked by the process to finalize the transaction. When the connector is invoked by the process,
the customer will be charged with the amount held in the buffer. Since the connectors are enhanced with some capabilities, if the currencies don’t match with each other, necessary conversion is handled by the connector in real time. On the other hand, if the item cannot be supplied until the deadline, the bank will release the deposit and transaction will be cancelled. A sequence diagram is given in Fig. 3 in order to show how the system functions.

RELATED WORK

Considering the variability enhanced connectors usage in coordination, especially aligned towards practical composition goal there is not much to cite. Recent activity has evolved from more restrictive to more flexible approaches, basically revolving around the same infrastructure. Çetinkaya et al. (2017) first offered a solution where components needed to be tampered. Soon the work of Kaya et al. (2007) followed that relieved the components from the integration code and carried such responsibility to connectors. Here we further relieve the connector from the pressures to abide with the central arbitration for every interaction: it is easier to accommodate locally initiated activities. Ordering among the actions are more exclusively stated.

CONCLUSION

In this study, composition responsibility of a system that is composed of components is managed through a central process through the help of connectors. In order to achieve that goal, connectors are enhanced with some capabilities such as adaptation. Also, the connectors are modified to be suitable for being invoked to initiate messages. A connector can be invoked through a component or the central process. Thus, synchronization of a system is made more efficient and manageable through a central process. With this approach, a means to organize different kinds of invocations such as those sourcing from the peripheral devices or from the central process, are all managed in a top-down control paradigm through a central process.

Future work will utilize this infrastructure for providing connectors the reuse capabilities of the components. Domain specific composition environments should list besides a set of components, a set of connectors. Experimenting with existing component models will definitely prove beneficial.

REFERENCES


Fig 3. Sequence diagram for a successful purchase operation
A VARIABILITY PERSPECTIVE FOR IOT HETEROGENEITY IN SMART CITIES

Selin Unal\textsuperscript{1,2} and Ali Hikmet Dogru\textsuperscript{2}

\textsuperscript{1} TÜBİTAK İLTAREN
\textsuperscript{2} Middle East Technical University
\{selin.unal@tubitak.gov.tr, dogru@ceng.metu.edu.tr\}

ABSTRACT

This research is one step towards managing heterogeneity in Internet of things (IoT) field. A feature-model based variability is represented for a pilot domain model and an example development is presented where the variability model guides software component composition. Heterogeneity is tried to be solved through variability compliant component technologies. Especially with the late developments, connectors have gained capabilities to assume the important task of integration that further magnifies the heterogeneity problem. It is aimed to demonstrate the feasibility of such an approach by offering the feel of intellectual control over an example problem.

INTRODUCTION

Today many devices are connected to the Internet and the number is growing day by day. According to CISCO there will be 50 billion devices connected to the Internet in 2020 (Evans, 2011). IoT (also known as Internet of Objects) corresponds to networked devices that are communicating with each other or with humans through the Internet. There are different kinds of domains, protocols, and network connectivity methods for these devices that are connected to the internet (Abeele, Hoebek, Moerman, & Demeester, 2015). All these different types of devices use their own language for communication. Since these devices do not use the same language to communicate at the data level, IoT world faces the heterogeneity problem.

IOT Heterogeneity Examples from selected Domains

There are some provided solutions for the heterogeneity problem in IoT. For example, for forecasting drought a study offers a semantic middleware application which uses ontology through the RDF (RDF xml syntax) and Web Ontology Language (OWL) (Noy, 2014) technologies to solve this heterogeneity problem. In this study to forecast drought, the application needs to collect and sense data from heterogeneous sensors which are located in different locations (Akanbi, & Masinde, 2016). Provided semantic middleware solution aims to hide the complexities and eliminate the data heterogeneity from multiple data sources. This solution also provides abstraction for complex network communications and presents the data in machine readable forms for easy integration. Shortly, this study aims to provide ontology-based semantic middleware solution that will ease the representation and integration of heterogeneous sensor data.

In another survey it is stated that there are four basic challenges in IOT world which are:

1. scalability of the things in the internet and managing them,
2. deep heterogeneity (Teixera, Hachem, Issarny, & Georgantas, 2011),
3. dynamic and unknown network topology, and
4. incomplete and inaccurate metadata

In consideration of these problems, this survey provides an ontology-based middleware solution. This middleware solution is based on the Service Oriented Architecture (SOA) that abstracts things as services (Hachem, Teixeira, & Issarny, 2016). And since SOA is used this middleware provides loosely coupled services and increases reusability. And by using ontologies in this middleware they aim to solve the challenges in IoT they have stated.

In another project that is called virtualized programmable Interfaces for innovative cost-effective IoT deployments in smart cities (VITAL) (Kazmi, Jan, Zappa, & Serrano, 2016), the researchers try to solve heterogeneity problem for smart cities by using Linked Data Technologies. They have introduced an integrated virtualized paradigm for the development and deployment and operation of smart city applications, which emphasizes the collection and processing of data streams from heterogeneous sensors and IoT platforms across the urban environment. VITAL is a system of systems; a system that can support any underlying IoT system. VITAL uses linked data (Heath, Bizer, 2011) standards for modeling and accessing data including RDF (RDF xml syntax) as a basic data model, JSON-LD (Jason ld) as the data format, and ontologies to specify the data in a formal way.

Connected devices are used in various domains like smart city, building control, logistic, energy management, waste management, mobility, healthcare and transportation. In this work we concentrate on the smart city domain for demonstrating our work on the resolution of heterogeneity.

MODELING THE SMART CITIES DOMAIN

In this section, the selected domain is studied and modeled in terms of IoT variability. Feature Modeling is selected as the fundamental modeling environment for presentation purposes. Our objective has not been to...
completely model this domain. Only a pilot study has been conducted that considers a realistic domain in part.

The Smart City projects aims to make cities more livable and more comfortable for people. In the cities people are facing many problems like parking, heavy traffic, polluted air, finding available charging stations, knowing whether trash bins are full or not, and so on. The example smart city project presented in this article aims to solve the heterogeneity problems created by using sensed data that the application gathers from different devices in real time. There are many sources to collect data from in the city. For example, each street light can gather and send information by attaching sensors and communicating devices. By using these data, an application can make a decision for example if there is an available parking slot or not in real time. Another example is gathering weather information from an enhanced street light to be used in assessing weather conditions and to prevent flood. Another example is assessing the occurrence of an accident and necessary notifications are conducted so that actions can be taken accordingly. As we can see from the given examples smart city applications are very important in today’s life. The smart city domain is closely related to IOT world since there are many different kinds of devices that need to talk continuously. Some devices need to gather information and such information needs to be interpreted and finally produced results need to be transmitted to the necessary places. As we can see here, smart city applications face the heterogeneity problem since there is no common way to communicate for devices and this heterogeneity problem makes harder the applications to work appropriately.

Since a smart city project is selected as an example, in this section we will show the core infrastructure elements of a smart city (Alshafie and Saife, 2017). Some of the constituents of this domain can be listed as:

- Smart homes and smart buildings,
- Smart parking,
- Smart healthcare,
- Smart transportation and smart traffic,
- Smart security, and
- Smart environment.

All of the components stated above aim to improve the quality of life for the citizens. Smart homes can be supported for detecting fire, monitoring temperature, security systems, and social networking. By utilizing the various data collected from different sensors these solutions can be achieved. Smart parking areas enable managing the arrival or leaving times of different cars by sensing the data from the cars. In this way other citizens can be informed about the available parking lots. In terms of smart healthcare, the status of the patients in the hospital can be observed and in this way their treatment can be manageable more secure and faster. Location of ambulances and blood products and organ transferring can be monitored for availability in real time. For smart transportation, data can be used for analyzing and taking action accordingly. For example, informed about the presence of a passenger at a specific bus station, the bus driver will drive there and will not otherwise. Drivers can be informed online about traffic density at specific roads. Smart environment can be used for renewal of energy, sensing weather conditions, and taking action accordingly by using sensed data from the environment. Smart security can be used for improving the security of a city by using data collected from streets, parks and any place that is concerned.

The key components of the smart city are shown in Fig. 1.

![Feature Model for the Smart City Domain](image)

**Feature Model for the Smart City Domain**

As an example domain we select smart city and we will focus on the smart parking element of it. A partial feature model of the smart city is provided in Figures 2 through 7. The model is too big to fit into one page therefore it is divided into parts. For the purpose of this work only smart parking is divided into subparts. If needed other parts of this model can be constructed in detail.

Assuming we need to build smart parking for a selected street of a selected city, we will use the provided feature model. We are examining each figure and while we are choosing the components of the Smart Parking we will explain the heterogeneity problem and provide a solution as we go step by step.

In Figure 2, we have alternative options that are Smart Parking, Smart Home and Buildings, Smart Healthcare, Smart Traffic, Smart Environment and Smart Security to build smart city model. According to the type of the connections under the root, we need to choose at least one of those options. On the assumption that we want to build only smart parking for this work, we choose only Smart Parking option from Fig. 2.
In Figure 3, we have mandatory options that are Data Sensors, Data Center, Data Requesters and Data Sources. In this part these options will be explained to indicate what is intended. We have Data Sensors that can sense data with specific communication protocols and send this data to data centers with specific communication protocols. Data Centers can get the data from sensors and interpret them. Data requesters are applications that want to know the status of parking areas at specific times. Data sources are the sources of the data for the Data Sensors. We need to choose all of these options according to Fig. 3. We have Data Sensors, Data Center, Data Requesters, and Data Sources defined up to this point.

In Figure 4, we have some mandatory and optional options for Data Sensors. As mandatory options we need to choose Protocol Type, Brand, and Sensor Type. As optional options we have Operating Distance and Accuracy. For Protocol Type we have alternative options that are ZigBee, Bluetooth, LORA, and WIFI. For the purpose of this work we choose the WIFI protocol among the provided protocols. For Brand option we choose BrandA (Assuming BrandA or BrandB has different kinds of devices communicating with provided protocols in this model). For sensor type we choose camera. As Operating Distance and Accuracy are optional features. We choose 15 meters for Operating Distance and 90% as Accuracy. After all we have a BrandA camera talking with WIFI operating in 15 meter and has 90% accuracy for the Data Sensors part of the Smart Parking.
In Figure 5, we have the “Data Center” feature which will collect the sensed data from sensors and interpret them. At this point we face with the heterogeneity problem that since we chose the WIFI protocol for the data sensors we need to choose the same for the Data Center. In this way we are providing a solution for the heterogeneity problem at the variability level. The compatibility issue is addressed at an early phase, at a modeling level. Such selections in the variability representation are expected to be propagated toward configuration of the executable outcome. The desired automation of such a configuration will depend on the capabilities of the development environment.

In Figure 6, we demonstrate the Data Requesters feature that will query the situation of the parking areas in real time. For the Data Requesters Smart Phone application and Web application alternatives are provided. Since these applications will query data through the internet we can choose one of them among provided options. For the purpose of this work we choose the Smart Phone IOS application feature.
In Figure 7, we have Data Sources and we have some mandatory and optional options. As mandatory options we have Data Source Type and Connectivity Type. As an optional part we have Distance. For Data Source Type we need to choose out of the provided options which are mobile phone and car. We choose car for the data source type for the purpose of this work. And for the connectivity type we are facing the heterogeneity problem. We have chosen the WIFI protocol for the sensor before. Since we chose WIFI, again we need to choose this protocol to overcome the heterogeneity. In this way we can solve the heterogeneity problem in variability level. Therefore, we choose WIFI communication type. As Distance feature we choose 15meter option. After all we have car talking with WIFI operating in 15meter.

So far we have Smart City, Smart Parking, Data Sensors Protocol Type: WIFI, Operating Distance: 15m, Accuracy: 90%, Brand: BrandA, Sensor Type: Camera, Data Center—> WIFI/PC, Data Requesters: Smart Phone/IOS application, Data Sources—> Data Source Type: Car, Connectivity: WIFI, Distance: 15meter is selected. Again, at the variability level we choose appropriate communication protocols for the selected options. By choosing compatible protocols we provide a solution at the variability level for addressing the heterogeneity problem in IOT at variability modeling level. In Figure 8, a feature model is presented for the decisions about the planned smart city application. In Software Product Line Engineering (SPLE) this is referred to as the “product feature model.”

---

Fig. 7 Smart Parking Feature Model Sources

---

Fig. 8 Product Feature Model
COSEML DIAGRAM FOR THE SMART CITY DOMAIN

A partial COSEML diagram for the smart city is provided in Figures 9 and 10. The model is too big to fit into one page therefore it is divided into parts. For the purpose of this work only smart parking is divided into subparts and only components corresponding to the selected features from the feature diagram are provided. This COSEML model therefore corresponds to the product rather than the domain.

In Figure 9 the Smart City applications COSEML diagram is provided including the packages that are Smart Parking, Smart Home and Building, Smart Healthcare, Smart Traffic, Smart Environment and Smart Security. Since we develop the Smart Parking capability for this example, other packages are not decomposed in detail. In Figure 10, the Smart Parking package is decomposed. It includes the Data Sources, Data Sensors and Data Centers packages. The Data Sources package includes car and mobile phone packages. The Data Sensors package includes camera, street lights, and traffic lights packages. Each package is implemented through specific components. Since we only aim to show connectivity between the car and camera components only the related decomposition is provided in the diagram.

As we insert the components for the smart parking application according to the communication protocols, the heterogeneity is solved by adhering to our previous decisions at the feature modeling level. The conforming components take place in the COSEML diagram. In this way we can
manage IOT heterogeneity problem in the components, hence the implementation level.

**CONCLUSION AND FUTURE WORK**

By using provided feature model, we have chosen compatible features while building selected domain’s components. After this, selected features are matched with the components from the component diagram. In this way we provide a solution for heterogeneity problem at variability level.

**REFERENCES**


Case Study on Improving Data Accessibility and Reducing Interface Complexity in the i2b2 Web Client

Sai Saran Chintha, Dr. Leon Jololian
University of Alabama at Birmingham, Department of Electrical and Computer Engineering
{saisaran@uab.edu, leon@uab.edu}

ABSTRACT

Clinical informatics has largely focused on providing useful tools that can help medical researchers curate effective clinical studies and therefore, improve the overall quality of healthcare. However, the tools themselves are often not designed with the clinical researcher in mind. The usability of these systems and their ability to integrate with the clinicians’ workflow have not been given adequate attention. There is a growing need to improve the user interface of these tools to support the researchers and help them to use it effectively without requiring significant preliminary training. At the UAB Informatics Institute, we worked on a case study (in progress) that focuses on completely redesigning the i2b2 Web Client from the ground up and improving data accessibility with the clinician in mind. We use lightweight, event-based web technologies both on the client and the server such as React and Node.js respectively. We modify the data interchange format to JSON and demonstrate how it can be a lightweight alternative to XML. We propose a change in design language, illustrate the new proposed design and compare it to the existing i2b2 Web Client. The working code repository is available at www.github.com/saichintha/i2b2-client-react and www.github.com/saichintha/i2b2-backend.

INTRODUCTION

Integrating Biology and the Bedside (i2b2) Web Client [14] is an open-source clinical informatics tool that provides useful ways for clinical researchers to access patient health records and find patient cohorts given a set of parameters. The parameters can vary from broad diagnoses such as diabetes to very fine-grained details such as the iron level from a complete blood count (CBC) test. It allows researchers to gauge the feasibility of a study, identify potential patients for study participation and analyze clinical data. Numerous clinical trials and research studies in the recent years have increasingly used i2b2 to investigate and find patient cohorts for discovery research [2]. i2b2 is now widely used at 34 academic health centers in the US with several more around the world and at various health maintenance organizations (HMO) [1]. Ongoing work on i2b2 is being curious but various academic institutions such as The University of Kansas [7], University of California at Davis [8], The University of Texas, to name a few.

Ontology Browser

As defined by OpenClinical [12], an ontology in the field of medicine is “mainly focussed on the representation and (re-)organization of medical terminologies. Physicians developed their own specialized languages and lexicons to help them store and communicate general medical knowledge and patient-related information efficiently”. The ontology in i2b2 is similar to the definition above in the sense that it collects patient data and organizes the various demographics, diagnoses etc. into a hierarchal path within the database [13]. The i2b2 database is structured primarily after a well-defined ontology that spans across a multitude of data on the various diagnoses, genome demographics, laboratory tests, medication, procedures, visit details, and reports. Data within each level of the ontology is further organized in more specific groups. For example, the “Diagnoses” path in the ontology organizes further such as circulatory system, endocrine disorders, congenital anomalies, neoplasms etc. The specificity of the terms, also known as concepts within i2b2, increase as the path goes deeper and deeper in the ontology.

Querying Mechanism

A typical i2b2 query returns the number of patients who fit the criteria defined by the various concepts and the concept relationships. The i2b2 query tool allows the user to select any concept from the ontology browser and add it to the query builder that is essentially made of several panes. The primary advantage of the i2b2 Web Client is the visual query
builder that lets clinicians gather concepts (diagnoses, medications, lab tests etc.) and organize them in a way to build an SQL query on the backend of the system.
The query tool consists of groups that can contain multiple concepts and are connected together by logical AND and OR operations. For instance, the concepts in Group 1 can be AND(ed) with the concepts in Group 2 which would result in a patient set that have concepts in Group 1 and concepts in Group 2 as well.

To be specific, consider the concept “Abdominal pain, generalized” was added to Group 1 and “Intrinsic asthma without mention of status asthmaticus” to Group 2. The default connection between the two groups is set to the logical AND operation. Therefore, the query tool returns the number of patients who have both the concepts “Abdominal pain, generalized” and “Intrinsic asthma without mention of status asthmaticus”. Patients who have one concept and not the other would not be part of the return patient count. The patient has to satisfy both the concepts to satisfy the criteria assigned. The array of these groups are unlimited and the connecting operator can be changed. This visual representation helps the researcher understand how the query is being built and illustrates the relationships among concepts.

In this regard, a case study has been conducted to illustrate how the i2b2 Web Client can be designed for a more user-centered, data-driven application that can help clinical researchers to be more effective. Design changes related to the ontology browser, group querying and keyword search in the i2b2 Web Client were made to reflect the user-centered design language.

**Approach**

**Application Architecture**

The application architecture is a three-tier architecture that envelopes the server (backend), the client (frontend), and the i2b2 database which adhere to specifics of the i2b2 start schema. The backend of the application is run in a Node.js [10] environment and is essentially a web server that responds to various HTTP calls. The frontend of the application is the implementation of the new design and are static HTML, CSS, JavaScript files that are needed to render the frontend. Various actions within the frontend prompts calls to be made to the backend, the backend server interacts with the i2b2 database and responds with the result formatted in JSON.

![Application Architecture Diagram](image)

Figure 1: The application architecture of the new proposed design and data interchange routes within the architecture tiers.

The revamped user interface has been built from the ground up using modern web technologies, revolving much around Facebook’s React ecosystem [9]. Various open-source libraries that work closely with React such as Redux [11] were also used to varying degrees. The backend web server that responds
to various API calls by the frontend was also built using server-side JavaScript using libraries such as Express. The overall architecture is best explained in the Figure 1 below.

Design Considerations
The design languages was implemented closely after the Material Design [4] specification as proposed by Google. Much of the design implementation throughout the user interface elements were used from a predefined set of elements in the library, material-ui. The material-ui library follows Material Design to every specification and pixel. Material Design was chosen among all other design languages for its familiarity among a wide range of people, ease of use and most importantly, the excellent set of design choices and selection of elements carefully crafted by Google that help simplify the most complicated of applications. The i2b2 Web Client consists of multiple panes spanning across the page that could overwhelm any new user and would certainly benefit from a redesign. Furthermore, the interplay of vibrant color schemes paired with various varying levels of motion, material choices provide a more open and free outlook to potential users.

Home Screen Changes
The most challenging task in executing an i2b2 query is to find the right concepts from the ontology browser. Although some of the challenges lie in the inconsistencies in the ontology design itself, many stem from the tree display architecture of the browser. Concepts can be hard to find if one is not aware of the potential parent concepts and could be blindsided by duplicate concepts with different parents. As such, the ontology browser was completely replaced with a search bar. The group query panes were also repositioned to be at the center of attention to the user as shown in Figure 2.

Figure 2: The home screen of the proposed design

Revamped Search
The search bar was built to provide the most relevant and most impactful concepts upfront to the user depending on the keyword entered. Search was designed to return concepts that have at least one patient is observed to have and the result concepts themselves are ordered by the number of patients that are observed to have that particular concept. The other panes on the screen were also designed to be hidden during search to focus the users’ attention on the task at hand and also to have more real estate to populate the results. The resulting concepts for a search on the keyword “heart” is shown in Figure 3. Other resulting concepts that are positioned lower are not included in the Figure for brevity.

Figure 3: Screenshot illustrating the user interface
during search. Each search result (concept) is shown ordered by patient count (the far right number) and the concepts relations tree view, patient demographic view button depicted from right to left of the patient number. The green button the far left of each result is used to add the particular concept to a group.

Search populates the screen with the resulting concepts order by patient count and provides information as to where the concepts originate from in the form of file path tags. Every concept that is returned in a search query can be further investigated for demographic information, relation to other concepts in the search results and be added to a group pane to form a query.

Patient Demographics and Concept Relations
The complete demographic information of patients that are observed to have a certain concept can retrieved right from the search result before ever adding it to a group pane. The user can also visualize how the concepts in the search results relate to each other in the whole i2b2 ontology from the tree view. Concepts are positioned in a tree architecture based solely on their concept path and can be searched within
the tree view for more efficient discovery. The demographic information pane and the tree view pane on the concept titled “Palpitations” from Figure 3 are shown below in Appendix 1.1 and Figure 4, respectively.

Group Querying

Any concept from the search results can added directly to a certain group in the group query panes to build a query. The green plus button lets a user add the current concept to a group by displaying the current state of the group query panes (state of the group panes consists of the concepts already in the groups and will change over time). The patient count of the particular concept also carries over to the group panes when added. For instance, consider the user also added the concept title “Angiocardiology of left heart structures” which a patient count of 4 to the Group 2 pane. When the user attempts to add another concept to the group, the group state will now contain the previous concepts and the display of the group panes will reflect the consistency. This particular example is shown in Appendix 1.2.

The home screen will now reflect the changes and will automatically generate a name for the particular query by using the first five characters of each concepts that has been added and separated by a hyphen. The changes are shown in Appendix 1.2. Concepts added to the a group can also be easily removed from the query by hovering the cursor over the concept in the group and clicking on the red button that appears as shown in Appendix 1.4. The “SEARCH” button when pressed would execute the query and display the results along with the demographic information for the result patient set and is shown in Figure 5.

The query building process was completely remodeled using JSON (JavaScript Object Notation) as the primary data structure to communicate information to and from the server backend. The concepts within the groups and the position of these concepts are modeled using JSON Arrays and are sent to the server through an HTTP POST request. Concept codes are placed within pre-defined SQL queries and the result data is sent back to the client using JSON. The request payload for the query containing “Heart Rate (LOINC: 8867-4)” in Group 1 and “WHITE BLOOD COUNT (LOINC:6690-2)” in Group 2 is shown in Appendix 1.5. The generated SQL query on the server backend is represented below.

```sql
SET search_path TO i2b2demodata; SELECT CONCEPT_CD, COUNT (*) FROM(SELECT CONCEPT_CD FROM OBSERVATION_FACT WHERE (CONCEPT_CD LIKE 'DEM|RACE%' OR CONCEPT_CD LIKE 'DEM|AGE%' OR CONCEPT_CD LIKE 'DEM|RELIGION%' OR CONCEPT_CD LIKE 'DEM|LANGUAGE%' OR CONCEPT_CD LIKE 'DEM|SEX') AND PATIENT_NUM IN (SELECT unnest(array(SELECT DISTINCT PATIENT_NUM FROM OBSERVATION_FACT WHERE CONCEPT_CD LIKE '%LOINC:8867-4%')) INTERSECT SELECT unnest(array(SELECT DISTINCT PATIENT_NUM FROM OBSERVATION_FACT WHERE CONCEPT_CD LIKE '%LOINC:6690-2%')))) A GROUP BY concept_cd ORDER BY 1;
```

The results returned from the i2b2 SQL database is then parsed into JSON object using an ORM (Object Relational Mapper) library called Sequelize. This JSON object is then returned back to the client. The response of the original HTTP POST

Figure 4: Screenshot illustrating the tree view visualizing of how the particular concept is in relation to all the other concepts in the search results. The particular concept is highlighted in green.

Query Builder
request is shown below in Figure 6. Demographic information such as age, language, sex, religion and race are returned along with the patient count for the particular concept. The age demographic information above the list shown is cropped out of the figure below for brevity.

**Figure 6:** Part of JSON response (as seen in the Chrome Developer Tools) received for the SQL Query shown above. “concept_cd” is the name of the concept and “count” is number of patients who have the particular concept.

---

**EVALUATION**

**First Impressions**

The design language and new concepts have been well received by the i2b2 steering committee at the UAB Informatics Institute [6]. The study was approved for further development and testing. The application is expected to be tested at various clinics at UAB including the UAB Comprehensive Cancer Center once the testing environment is set up.

**Design Comparisons**

The new proposed design adopts for a cleaner look overall with less clutter. In comparison to the i2b2 Web Client, the new design promotes a more welcoming aesthetic for potential new users and provides relevant information to the task hand upfront without requiring user input. However, this is subjective to the person and would require a substantial study among current i2b2 users to validate. We intend to continue our study further to include metric-based testing. First impressions of the application, as stated earlier, have been very positive.

**Comparing Search**

The search feature improves upon the existing keyword search in the i2b2 Web Client by retrieving search results in the order of the patient count and incorporating a multitude of user actions that can be performed right within the search results page. For example, a search for the keyword “heart” in the i2b2 Web Client returns more than 200 results as shown below in Figure 7. Many of the concepts return have a patient count of zero, which is not displayed to the user upfront, and would essentially render it unnecessary in the query building process.

The i2b2 Web Client uses folder/file icons to indicate that the concept shown has a deeper level in the ontology or is the last level in the ontology for that particular path. However, due to the specific naming of the concepts and close relationships, clinicians have found it challenging to pick the right concept out of multiple very similar concepts. This would usually entail a try-and-error workflow for the researcher.

**Figure 7:** Depicts the partial listing of the results for the keyword “heart” in the existing i2b2 Web Client. The yellow icons represent the particular concepts has more concepts embedded within it and the gray file icon represents an endpoint in the ontology path.

The search proposed improves upon the data transparency by showing the patient count for each concept upfront and ordering search results by patient count. The search for the keyword “heart” using the new search mechanism is shown in Appendix 1.3. The search results can also be viewed as tree to visually illustrate how each search results is related to every other search result. Demographic information is also a click away, the greatly enhances the steps that the researchers has to perform and takes away the guessing process when multiple very similar concepts are involved.

**XML vs JSON**

The underlying data models were also changed from XML in the existing i2b2 Web Client to JSON in the proposed design. As stated by a case study on the performance between the two data interchange formats, “JSON is significantly faster than XML” [5]. XML requires more system resources and computing power to be parsed when compared to JSON. The major advantage to using JSON is the native support in JavaScript environments such as Node.js. Native support eliminates the need for parsing of the data received or special packaging of the data to be sent. Additionally, the data transfer sizes that occur with XML is almost always higher than JSON. For instance, consider the query that contains “Blood Urea Nitrogen (GROUP: BUN)” in Group 1 and “PLASMA
CALCIUM (LOINC:2000-8)" in Group 2. When the concepts are added to the respective group in the existing i2b2 Web Client, the HTTP request headers that contain the query information sent in XML format is as shown below in Figure 8. XML containing information only for the “PLASMA CALCIUM“ concept is shown below for brevity.

Comparatively, consider the request headers when the same query request is sent from the proposed design using JSON. Although the data contained in the two figures differ, JSON is clearly more efficient to parse and extract information from.

**Figure 8:** Partial view of the XML sent through an HTTP request from the the existing i2b2 Web Client after the “Run Query” button is clicked. Depicts information on one concept, “LOINC:2000-8).

![XML Request](image)

**Figure 9:** Complete JSON that is sent as payload for a query request that contains the two concepts shown. Only the concept codes are transferred through the HTTP request.

**CONCLUSION**

Our research on the i2b2 Web Client at the UAB Informatics Institute and the UAB Center for Clinical and Translational Science, we report on a work in progress that focuses on rethinking the user interface and improving accessibility of information. We use the Material Design language specification by Google and weave the design elements to remodel the i2b2 Web Client. The web application was built on the React ecosystem and user interface libraries like material-ui were extensively used to implement the Material Design specification. Proposed design changes include the replacement of the current ontology browser int eh i2b2 Web Client with the new search at the center of attention to the user. The search was designed to return concepts matching the keyword ordered by the highest patient count. Every search result (concept) was implemented to quickly get demographic information of the patients who have the concept, visualize how the particular concept is related to other concepts within the search results. The tree view is implemented using the library, react-sortable-tree, which lets the user search for a element within the tree by pattern matching. Changes to the data interchange format between the server and client was also proposed to be JSON and the various advantages of using JSON compared to XML in the existing i2b2 Web Client was demonstrated. Design comparisons to the the existing i2b2 Web Client were also made by picking out certain elements in the UI and were evaluated subjectively. We intend to further study the advantages of the proposed changes in a clinical setting and evaluate against a set criteria. For now, the proposed design changes have been well-received within the UAB Informatics Institute.

**ACKNOWLEDGEMENT**

The authors would like to thank Dr. James Cimino, Director of UAB Informatics Institute, for the opportunity, support, and contribution throughout this research.

**REFERENCES**


[3] XUAN LU, ZHENPENG CHEN, XUANZHE LIU, and HUORAN LI, and HUORAN LI, Peking University, China TAO XIE, University of Illinois Urbana-Champion, USA QIAOZHU MEI, University of Michigan, USA. PRADO: Predicting App Adoption by Learning the Correlation between Developer-Controllable Properties and User Behaviors


APPENDIX

[1.1]

[1.2]
RADIAL FIBER ATROPHY: A NEW METRIC FOR TENSOR-BASED MORPHOMETRY

Jon R. Marstrander 1, Thomas Anthony 1, Veronica L. Powell 1, R. Glenn Brook 2, Mitchel D. Horton 2, Frank M. Skidmore 1,3
1: Department of Electrical and Computer Engineering, University of Alabama at Birmingham,
2: Joint Institute for Computational Science, University of Tennessee
3: Department of Neurology, University of Alabama at Birmingham
{ jmars@uab.edu  fskidmore@uab.edu }

ABSTRACT

Tensor-based Morphometry (TBM) is a widely-used class of algorithms for measuring anatomic differences in Magnetic Resonance Imaging (MRI) of the human brain. This paper proposes a new metric within the field of TBM called Radial Fiber Atrophy (RFA). RFA combines information from a deformation map, also known as a warp, with information from Diffusion Tensor Imaging (DTI). The core idea of this measurement is to compute stretch of the deformation field in the plane perpendicular to the predominant direction of diffusion as defined by DTI. Since the principle direction of the diffusion tensor in DTI is thought to be related to the direction of axons, the RFA metric is theoretically related to changes in the macrostructural cross-sectional fiber caliber of nerve fibers in the brain. It is anticipated that this new metric will provide significant insight and information for the study of neurodegenerative diseases.

INTRODUCTION

Dramatic differences in the size and shape of brain structures exist between healthy individuals. In addition to these normal anatomical variations, there are significant anatomical changes that can be observed in brains that are affected by neurodegenerative diseases. A variety of methods are used to match brain structures, including nonlinear warping. A full discussion of this complicated field is beyond the scope of this paper, but has been reviewed elsewhere (Klein et al 2009).

For clarity, it is important to note that the term “Tensor” is used in this paper describes two different, but related concepts. Tensor flow is a mathematical metric used to describe the characteristics of the deformation field when matching one image to another. A diffusion tensor mathematically describes Brownian motion (diffusion) of water by a Magnetic Resonance Image. These two concepts are further discussed below.

Tensor-based morphometry

Tensor-based morphometry (TBM) is a method used to capture brain variability, allowing differences in brain structures to be measured. (Ashburner & Friston, 2004) In the case where brain structures are well matched, TBM has been used to characterize changes in the size of matched structure (Ashburner, Good, & Friston, 2000) (Hua, et al., 2008). Continuing advances have been made in applying nonlinear warping techniques to brain imaging (Jin et al, 2015). TBM in the past has previously been applied to anatomic images.

Diffusion Tensor Imaging

Diffusion Weighted Imaging (DWI) is a method of magnetic resonance imaging that obtains measure of the diffusion of polar water molecules (e.g. Brownian motion) in tissues (Bammer et al, 2003). In the brain, DWI is hypothesized to measure diffusion in water along neural tracts. The DWI can be collapsed into a tensor called the Diffusion Tensor Image (DTI) (Assaf et al, 2008). The principle assumption of brain DWI is that diffusion measures microstructural diffusion. In fluid filled structures, diffusion is expected to be equal in all directions. However, in brain tissue, fluid diffusion is restricted by cell membranes. Water is expected to diffuses along neural tracts more effectively than it is expected to diffuse across membranes. The degree of directionality (anisotropy) of the diffusion is referred to as Fractional Anisotropy (FA) (Assaf et al, 2008). The principle direction of the DTI tensor would represent the principle direction of a neural tract. Diffusion along this principal direction can be derived from the DTI, and is called Axial Diffusivity (AD). Diffusion radial to this principle direction can also be derived from the DTI, and is called radial diffusivity (RD). The DTI tensor and these related metrics can be represented as an ellipsoid, and the associated metrics are illustrated in Fig. 1.

The brain can be represented as a three dimensional FA map, which is considered to be a representation of brain regions in which nerve fibers are highly directional. Figure 2 shows a mean FA template derived from 151 healthy young adults (Neuroimaging Informatics Tools and Resources Clearinghouse).

Computation of the RFA

The sequence of computations discussed in this paper is as follows:

A nonlinear deformation mapping is performed to align the brain structures of a subject brain image to a reference brain. This mapping may occur as a sequence of steps, with
the full deformation map consisting of the combination of all mapping steps. In the present analysis, the FA imaging modality is used to generate this deformation map.

TBM is computed from the deformation map at every point in the brain image. Information from the DTI is combined with the TBM to determine a plane perpendicular to the principal direction of nerve fibers. The RFA metric is computed from this combined information. This metric is computed at every point in the brain image, producing a new, 3D image of the brain. The RFA is presumed to measure a macrostructure characteristic of nerve fiber bundle—specifically the change in the projected regional macrostructural cross-sectional, radial area of the matched fiber bundles, after controlling for the longitudinal component of the fiber bundle.

Illustrations

The non-linear deformation field could be generated by a variety of programs. In the present research, the deformation has been computed by the 3dQwarp program, which is part of the Analysis of NeuroImaging (AFNI) software suite, discussed below. A graphical illustration of the results of this computation as applied to an actual FA brain image is shown in Fig. 3.

The simplified, two-dimensional illustration in Fig. 4 may help clarify the warping process. In the figure, the grey voxels represent nerve axons, bounded by the red lines. Note that the edges of the nerve tracts are likely to overlap voxel boundaries. The blue ellipses represent the DTI tensors in each voxel. To form the final, warped image, interpolation of voxels is needed. It is problematic to interpolate tensor values, but scalar values such as FA can easily be interpolated to produce a warped image.

The “ground truth” is only imperfectly measured, and does not exist in the final warped image. But the approximation provided by the warping information allowing us to compare microstructural diffusion and macrostructural atrophy or hypertrophy of tracts. This simple cartoon is provided for insight only. In practice, all operations must be performed in 3 dimensions with varying orientations, and the full metric calculations incorporate 3-D stretch, skew, and rotation, as described by this technical report.
MATHEMATICAL THEORY

Terminology notes

The analysis presented here is a slight modification of traditional TBM, which is a well understood and validated technique. (Ashburner, Good, & Friston, 2000) (Hua, et al., 2008) A traditional approach within TBM is to compute the determinant of the Jacobian of the deformation field. The approach presented in this paper performs a different calculation based on the deformation field. It is also typical to use a T1 MRI image in TBM. In the present analysis, deformation maps have been computed based on the Fractional Anisotropy (FA) images, which are derived from Diffusion Tensor Imaging (DTI).

In order to understand the process, it is important, briefly, to disentangle the two uses of the word “tensor”. The word tensor as used in “tensor-based morphometry” refers to the math involved in translating a set of coordinates with rotation, skew, and stretch from one location to another, and retaining a measurement of the change in voxel dimensions. The word tensor in “Diffusion Tensor Imaging” refers to a static measure of diffusion of water molecules in a particular voxel, as detected by an MRI. When performing a 3-D “tensor based morphometry” to fractional anisotropy, one obtains a measure of how the FA measure “moved in space” using tensor flow. This “tensor flow” can then be applied to the original diffusion tensor, to determine how much of the change occurred in a longitudinal direction along the principle axis of the diffusion tensor, and how much of the change occurred along the radial axis. To ensure mathematically valid results, it is vital that each step of the mapping operations is invertible, so that the composite mapping is also invertible.

Overall process

The first step in the process presented here involves a sequence of manual, rigid-body rotations and translations to align brain structures as well as possible in a global manner. The next step is to perform the non-linear warping. One can consider the total mapping transformation as a 3-dimensional, “rubber sheet” transformation of the FA signal. This deformation is analyzed in an attempt to measure the localized “stretching” of this rubber sheet in a plane perpendicular to the axis of diffusion, which is theoretically perpendicular to the plan of the principle direction in the tissue of the axonal bundles. The RFA measure becomes a point-wise value that will provide a numerical estimate of atrophy (or hypertrophy) in the particular location.

The following is a detailed discussion of numerical considerations with this approach. Note that some of the mathematical terminology used in the original discussion of tensor-based morphometry is replicated, because a full treatment is necessary to fully appreciate the method.

Manual isomorphic transformation

The first step of the mapping process is alignment of the brain image to a good first approximation of the atlas space. This transformation is a rigid-body transformation, allowing only rotation and translation. As such, it is a linear operation and can be represented by the follow matrix equation:

\[ x_M = A_{MN} x_N + b_{MN} \]  

(1)

where

\[ x_M \text{ is a 3x1 column vector in the manually aligned space,} \]
\[ x_N \text{ is a 3x1 column vector in the native space,} \]
\[ A_{MN} \text{ is a 3x3 orthogonal, rotation matrix, and} \]
\[ b_{MN} \text{ is a 3x1 translation vector.} \]

This operation is clearly invertible through elementary matrix operations. In practice, a sequence of isomorphic transformations may be performed to achieve the desired alignment. However, it is possible to mathematically combine this sequence of transformations into the single equation, shown above.

Warping transformation

The second step of the mapping process applies a warping process to provide a non-linear mapping to align individual voxels in the subject brain to a reference atlas. This non-linear mapping serves to more closely align individual structures. The warping process applies a 3-dimensional “rubber sheet” approach. The position of each voxel is locally stretched to more closely match the atlas image. This non-linear stretching produces a 3-dimensional vector map describing the warping transformation. Each voxel of the mapped image has a corresponding 3-dimensional vector indicating the original location of that voxel. It is a requirement that this transformation map must be “smooth” to a certain degree so that first order derivatives are continuous at all points. The technical term for these requirements is that the warping map must be a “bijective, C1 diffeomorphism.”

Mathematically, the warp can be represented as:

\[ x_A = W( x_M ) \]  

(2a)

\[ x_M = W^{-1}( x_A ) \]  

(2b)

where

\[ W( ) \text{ represents the forward warping function, and} \]
\[ x_A \text{ is a 3x1 column vector in the atlas space.} \]

The warping software produces a sampled representation of the forward and inverse of the warping transformation as described in the present discussion.

Combined mapping transformation

When applied successively, a composite, combined mapping function is produced:

\[ x_A = W( A_{MN} x_N + b_{MN} ) \]  

(3)

which must be invertible:

\[ x_N = A_{MN}^{-1} ( W^{-1}( x_A ) - b_{MN} ) \]  

(4)
It is advantageous to consider this combined transformation as a vector field. A mapping vector is associated with each point in the source space.

**MRI modalities and mapping transformations**

The native data taken by the MRI scanner in this case is a multiple-direction Diffusion Weighted Image (DWI). The DWI data cannot be manipulated by these transformations because it is not invariant under rotation, zoom nor shear operations. Diffusion Tensor Imaging (DTI) is derived from the full DWI data set. The DTI information is a symmetric, 3x3 tensor which is produced in the native space, and can be represented by six numbers. DTI can also be represented as a 3-dimensional ellipsoid. The 3-dimensional ellipsoid is represented by three axis lengths, and three, 3-dimensional vectors pointing along each of the semi-principal axes of the ellipse. The three vectors are representing the semi-principal axes of the ellipsoid are designated v1, v2, and v3, with v1 representing the longest of the three axes. The DTI information only makes full sense in the native space.

Fractional Anisotropy (FA) is a single-valued measure derived from the DTI information. As a scalar value, FA is invariant under transformations, and can be mapped by the warping function.

**Idealized measurement approach**

Attempting to measure the atrophy of the cross-sectional area of a tract of neural axons only makes sense in regions where a dominant direction of the nerve tract can be determined. These regions will correspond to a high value of FA. In regions with a low FA value, it is likely that these numerical results will be less meaningful. In the case of this analysis, an arbitrary FA cutoff value of 0.20 is used to define regions with high directionality.

Specifically, in these regions of high directionality, the 3-dimensional ellipsoid representation will correspond more closely to a "cigar" shape, and it should be possible to clearly determine a plane perpendicular to the direction of the tract. The v1 vector will lie in the direction in which the axons are aligned, along the long axis of the "cigar." The v2 and v3 vectors will describe the plane that is perpendicular to this axis. Recall that the v1, v2, and v3, vectors are defined in terms of the native, unmapped space.

For each point in these regions of high directionality, one can project the forward mapping vector field onto the plane formed by v2, and v3 at that point. One can then compute the "stretch" that occurred in the vector field as projected onto the plane as a measure of radial hypertrophy vs. atrophy.

One mathematical measure of the "stretch" in a vector field is the divergence of that vector field. Another commonly used metric within TBM is the determinant of the Jacobian matrix. The elements of the Jacobian matrix consist of partial derivatives of all vector field components taken with respect to all spatial directions. This paper presents an additional means of extracting meaningful information from the Jacobian matrix.

**Numerical issues and implementation**

There are some significant computational problems with the idealized approach. The mapping function produced by the warping operation \( W( ) \) is theoretically a smooth vector field, but it is stored numerically as points on a regular grid in the atlas space, not the original native space. So, while it is a fairly direct operation to numerically estimate the partial derivatives for the inverse mapping function in terms of the atlas space, it is more difficult to approximate the partial derivatives of the forward mapping function in terms of the native space.

Actual computations are performed at points in the atlas space. Working in this space will facilitate comparison between subjects and will also make the computations more manageable. Quantities referenced in the native space are always to be considered at the specific point in native space that is mapped to the point under consideration in the atlas space.

To compute the Jacobian of this vector function, first compute the Jacobian of the inverse warping function in the atlas space.

\[
J_W(x_A) = \text{Jacobian } \{ W^{-1}(x_A) \} \quad (5)
\]

Note that this Jacobian is a function of the position in atlas space. The values of the Jacobian tensor are computed at each point by taking the numerical partial derivatives of the Inverse Warping function.

Now, it is necessary to represent this quantity into the orthonormal space using v1, v2, and v3 from the DTI ellipsoid as basis vectors. Let \( x_V \) represent a vector (direction only) in this space.

\[ x_V \] is a function of each voxel in the brain image. So, for each point in the atlas space, a corresponding, nearest voxel in the native space must be located, and the values for v1, v2, and v3 from that voxel are to be used. The use of a nearest voxel is sub-optimal, because in general the inverse mapping from a voxel in atlas space will not land on an exact voxel in native space. However, no information exists about the values of v1, v2, and v3 in between voxels, so the nearest voxel is used.

Define a rotation matrix \( A_{VN} \) formed by the adjoining the column vectors v1, v2, and v3:

\[
A_{VN} = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} \quad (6)
\]

Next, the warping Jacobian can be "rotated" into the space of v1, v2, and v3 through the following equation:

\[
J_V(x_A) = A_{VN}^{-1} A_{MN}^{-1} J_W(x_A) A_{MN} A_{VN} \quad (7)
\]

Now, it is desired to measure the amount of "stretch" only in the cross-sectional, radial plane defined by v2 and v3. To do so, this new Jacobian is projected onto the v2 and v3 plane. This projection is done by simply dropping all terms in the matrix that involve v1. This reduces the rank of the matrix to 2x2. To measure the stretch within this v2,v3 plane, the
traditional divergence is computed, which is the trace of the reduced Jacobian matrix:

$$RFA = J_V [2,2] + J_V [3,3]$$  \hspace{1cm} (8)

This is the desired metric value. It is computed at each point in $x_s$.

Note that once the RFA metric has been computed, it is also possible to further derive an additional metric, the Axial Fiber Atrophy (AFA) as follows:

$$AFA = J_V [1,1]$$  \hspace{1cm} (9)

The RFA is the desired measure of "stretch" or hypertrophy as measured perpendicularly to the direction of nerve tracts, which is the radial diffusion direction.

Because the inverse warp was used in the computations, the sign of the numerical results is somewhat counterintuitive. Positive values of the RFA will indicate that in order to be mapped to a template, an individual’s structure must contract, indicating the original structure was hypertrophic. Negative values indicate that an individual’s structures must expand relative to the template, indicating the original structure was atrophic.

In terms of clinical significance, there are some potential similarities to other metrics that appear to be promising areas for future investigations. RFA appears to have some similarities to RD, and AFA appears to have some similarities to AD.

It should finally be noted that these measures are most useful when comparing the relationship of two groups to a given template. It is expected that both individuals with disease and healthy older adults may uniformly have some degree of atrophy with respect to standard templates. However, it is hypothesized that individuals with disease have more atrophy, and that it may be of significant value to further decompose this atrophy into radial and longitudinal components, as presented here.

IMPLEMENTATION RESOURCES AND SOFTWARE

The algorithms described above have been implemented and tested. The implementation uses a large number of diverse image processing tools and programming languages. Several particular software suites are discussed in this section. These software tool sets were chosen for the high quality and flexibility of their image processing. Although these particular tools are integrated into the present workflow, the ideas presented here could be achieved with other software. The analyses described are not married to these particular packages, nor the particular algorithms used to achieve the initial steps of processing.

UAB Research Computing

The University of Alabama at Birmingham (UAB) provides a high-performance computational infrastructure through their Department of Research Computing. This collection of resources is known as Cheaha. Extensive use has been made of the hardware resources of UAB Research Computing, and also of the expertise of the support personnel.

Oak Ridge National Laboratories

This research has been performed as a collaboration between UAB and the computational scientist at the Joint Institute for Computational Science (JICS) at Oak Ridge National Laboratories (ORNL). Programming expertise provided by their computational scientists and use of their computing infrastructure has played a key role in the analysis of data involved in this research.

TORTOISE

The National Institute of Health (NIH) publishes the Tolerably Obsessive Registration and Tensor Optimization Indolent Software Ensemble (TORTOISE) software package. (Pierpaoli, et.al., 2017) The TORTOISE software accepts raw DWI images from a scanner and produces DTI and related images as output. The software performs extensive calculations on the data to produce the best possible output images.

In the analysis presented here, TORTOISE has been used to perform the initial image preparation, the computation of the DTI, and associated DTI metrics.

AFNI

The NIH also publishes the Analysis of Functional NeuroImages (AFNI) software package. (Cox, et.al., 2017) AFNI consists of a large number of independent software tools that can perform an amazing variety of operations on MRI data.

In the analysis presented here, the AFNI suite has been used extensively to perform processing on brain images. In particular, the tool 3dQwarp has been used to perform the brain warping. Extensive work has been performed in conjunction with JICS/ ORNL to optimize the accuracy and efficiency of the 3dQwarp program. (Yin, et al., 2016) (Yin, et al., 2015)

MATLAB

MATLAB is a programming language and development environment produced by the MathWorks corporation. The specific algorithms described in this paper are primarily implemented in custom MATLAB code. All computations of the RFA are performed in MATLAB, using data produced by TORTOISE and AFNI. Results of the MATLAB calculations are written out as brain images for further processing.

BASH scripting

The Scripting language for the UNIX BASH shell has been used extensively as a top-level programming language to control and coordinate operations of all of the other software tools used in the analyses performed. BASH scripts have also been used in conjunction with AFNI tools to perform statistical analyses on brain image data from the PPMI data set and other data sets.
RESULTS AND CONCLUSIONS

The RFA metric presented here is a measure of radial fiber change. Note that in describing the RFA metric, it is implicitly possible to also measure the related axial fiber atrophy (AFA). The RFA and AFA metrics described here can be used to characterize macrostructural fiber change in healthy adults and in brain disease states.

The graphical results of these computations are shown in Fig. 5. The figure shows a subject brain with pronounced atrophy. This subject brain has been warped to the reference, Atlas brain that is shown. The resulting RFA measured for this subject is shown.

Note that the raw RFA measurement produces non-zero values in the area outside of actual brain tissue. This result is expected, because the rubber-sheet nature of the warping field will produce deformations beyond the actual brain structures that are moved. These deformations are a consequence of the requirement that the deformation field must be continuously differentiable. Also, the DTI tensor field contains a random direction for the v1 vector, even in the absence of actual brain tissue. Thus, a RFA value will be calculated at all points. In order to derive meaningful results from further analysis, it is important to mask the RFA image to the actual brain area in the Atlas space.

The authors of this paper have provisionally applied this metric to a comparison of individuals with Parkinson’s disease (a neurodegenerative disease) and healthy adults; a description of these findings will be discussed at a related SDPS conference, and are being developed for publication in relevant research and clinical journals.
REFERENCES


LUNG SEGMENTATION USING A CONVOLUTIONAL NEURAL NETWORK

David Odaibo¹ and Solomon Hatcher¹, Zheng Zhang¹, Chandra Yarlagadda², Frank Skidmore³,³
¹Department of Electrical and Computer Engineering, University of Alabama at Birmingham,
²Department of Biomedical Engineering, University of Alabama at Birmingham,
³Department of Neurology, University of Alabama at Birmingham
{godaibo@uab.edu solomonh@uab.edu fskidmor@uab.edu}

ABSTRACT

In this paper, we present an automated approach for segmenting lung tissue in low-dose computed tomography scans of the chest cavity. The method consists of two phases. The first phase consists of a series of image warps to align all samples to an anatomical template. The second phase involves applying a thresholding algorithm to segment lung tissue from non-lung tissue. High quality segmented samples from this initial phase are used as ground truth and training data for the second phase which forms the inputs to a convolutional neural network. The results on accuracy prove that convolutional neural networks are a promising tool for lung segmentation. Our method achieves a Jaccard similarity coefficient score of 0.9485 on 19380 2-D axial slices of the chest cavity.

INTRODUCTION

Cancer is one of the leading causes of death in the United States and computed tomography (CT) scans, Fig. 1, are an effective way of screening patients for lung cancer. Lung segmentation is often a critical step in various automated analysis techniques; yet, many lung segmentation approaches fall short in terms of accuracy of segmentation, segmentation time, and susceptibility to anomalies. In literature, various techniques have been proposed for lung segmentation. Brown, et al., 1997 presents an anatomy knowledge-based method for lung segmentation where anatomical knowledge is used to guide the segmentation process. There are obvious challenges with such an approach, a few of which include: the human engineering involved in capturing the anatomical knowledge; the effectiveness and completeness of the knowledge representation; and the susceptibility of such an approach to anomalies in structure. Sun, Bauer, and Beichel, 2012 also presents an anatomy knowledge-based approach to segmenting lungs. Initially, they use a shape matching method to locate the outline of the lungs and then utilize a surface finding method to adapt the initial segmentation. In recent years, convolutional neural networks (CNN) have outperformed other methods in visual recognition tasks (Sun, Bauer, & Beichel, 2012; Szegedy, et al., 2015). In this paper, we apply CNNs to lung segmentation. Our intuition suggests that CNNs will outperform prior approaches because of the ability of neural networks to automatically learn relevant features from the raw data, thus eliminating the need for human knowledge/feature generation that would be susceptible to human error and incompleteness.

Traditional CNN architectures lose the ability to localize objects as information propagates through them. The pooling operations often performed by the various pooling layers progressively reduce the spatial resolution as seen in Fig. 1, thereby limiting the networks ability to identify the location of the learned features in the original higher resolution image. Recently two neural network architectures have been developed to address the problem of segmentation: fully convolutional network (Long, Shelhamer, & Darrell, 2015) and convolutional autoencoder-like network or u-net (Fig. 2) (Ronneberger, Fischer, & Brox, 2015) as it is commonly called.

Both the architectures use up-convolution (i.e. up-sampling) in deeper layers of the CNN to upscale the feature maps. The original u-net architecture was developed for biomedical image segmentation and took a single channel greyscale image tile as input and produced a segmentation map.

For our experiments, we have modified this original u-net architecture to improve its performance.

Fig. 1 Convolutional neural network—pooling layers reduce spatial resolution
Our modification entails increasing the number of convolutional layers and adding batch normalization layers to reduce the internal covariate shift in the network. The modified architecture also differs from the original u-net in the concatenation/merge layers. The original u-net merged a smaller crop of the convolutional layers with the up-convolution layers; however, we maintained the dimensions of the merged layers.

**Fig. 2 Original u-net architecture**

**Fig. 3 Axial slice of chest cavity from CT scan**

### PRIOR WORK IN LUNG SEGMENTATION

There are numerous studies presenting methods for lung segmentation in literature, we present summaries of a couple here. In the paper “Adaptive Border Marching Algorithm: Automatic Lung Segmentation on Chest CT Images” (Pu, et al., 2008) the authors present a lung segmentation algorithm called adaptive border marching (ABM) that smooths the lung border in a geometric way and can be used to reliably include juxtapleural nodules while minimizing oversegmentation of adjacent regions such as the abdomen and mediastinum. As compared to other available methods, the authors claim ABM is more robust, more efficient, and more straightforward to implement. An obvious disadvantage of their method is the potential for over-segmentation into the abdomen and mediastinum which they try to minimize. This problem of over-segmentation and under-segmentation is a recurring theme in most non-neural network methods in the literature. In “Automated Lung Segmentation in X-Ray Computed Tomography: Development and Evaluation of a Heuristic Threshold-Based Scheme” (Leader, et al., 2003), the authors describe an automated scheme that was heuristically developed using a slice based, pixel-value threshold and two sets of classification rules. Features used in the rules include size, circularity, and location. The segmentation scheme operated slice-by-slice and performed three key operations: (1) image preprocessing to remove background pixels, (2) computation and application of a pixel-value threshold to identify lung tissue, and (3) refinement of the initial segmented regions to prune incorrectly detected airways and separate fused right and left lungs.

### THE DATASET

The original data consisted of 1600 low-dose CT scans of various dimensions, with each scan being in DICOM format. A CT scan is a volumetric image where each voxel has been assigned an integer value representing the attenuation of x-rays at each location throughout a substance. Through a series of linear transformations, the standardized measure of the x-ray attenuation coefficients was computed.

### DATA PREPROCESSING

**Creating the Training Data**

The creation of the training dataset required that the dimensions of the original dataset be standardized. The standardization process consisted of three stages. The first stage required converting the DICOMs into a NIfTI file format and then registering the images to an anatomical template (Yin, et al., 2016). A sample of 634 CT volumes from the registered dataset was selected for annotation by a physician. The second stage of the process generated masks for each CT volume; quality assurance was performed on the annotations which reduced the training dataset size to 208 CT volumes. The third stage required the images, and their masks, be resized from 425 x 380 x 410 voxels to 380 x 320 x 320 voxels with a slice thickness of 0.7 mm. The data was split into several “buckets” each containing approximately 30 full lung scans. Dividing the data between buckets mimics a multiple holdout validation scheme. Each bucket was further sub-divided into a training and validation set with a split of 75% to 25%. The volumes that failed quality assurance were used for testing.

**Data Synthesis**

The standardization of the CT data required a non-standard preprocessing methodology. The processing involved use of an NIH image processing software suite known as Analysis of Functional Neuroimages (AFNI) (Cox, 1996). AFNI primarily uses optimized C and C++ code to manipulate, resample, and register data. The training data
was created via two primary methods: warping (Anthony, et al., 2017) and semi-automatic annotation. Using a series of AFNI subprograms, the chest CTs were registered to the same coordinate space. Masks of the lungs were then generated on the warped data. The combination of the preprocessing methods allowed for a uniform approach to the development of a deep learning model for the segmentation of CT scans.

Initially all the CT scans were centered within a defined coordinate space using AFNI 3drefit. The data was then down-sampled to a voxel size of 4mm x 4mm x 4mm, and a median filter was applied. The median filter had the effect of zeroing out low signal data such as lung tissue. The remaining signal in the dataset comprised mostly of chest wall contours and internal structures such as the heart and major blood vessels. A single chest CT was selected for generating a uniform template at 4mm x 4mm x 4mm resolution. Using an optimized version of AFNI's 3dQwarp (Yin, et al., 2015; Yin, et al., 2016), a nonlinear tissue warp was applied to match each CT scan to the uniform template space. The 3dQwarp produced a bijective three-dimensional warp that was applied to match a given CT image to the template. A normalized mutual information optimization was applied in conjunction with a 6mm Gaussian blur, to both the input and template image. A within-voxel quadratic matching function was applied for contour matching.

Separately, the CT volumes were also resampled to a uniform resolution of 1mm x 1mm x 1mm voxels, as depicted in Fig. 4, using AFNI's 3dresample. The 4mm x 4mm x 4mm nonlinear warp produced by 3dQwarp was then up-sampled to a 1mm x 1mm x 1mm resolution and applied to the resampled CT data.

![Preprocessing workflow: image registration](image)

Non-essential data (e.g. not lung) was cropped using AFNI 3dZeropad to create a uniform, conservative version of the lungs in which all volumes were non-linearly warped to a uniform sample space. This uniform sample space was then used for generating the masks needed for training a deep learning model.

Generating masks for the training data required the application of various rule-of-thumb assessments by a physician. A k-means clustering algorithm was used to enumerate regions of interest by size. The result of the algorithm was a measure of the likelihood that a cluster represented lung tissue or not. This determination by the algorithm was used to generate a mask for each CT volume. Lungs were determined to be accurately identified if a region overlapped a group minimal estimate of space that was definitively lung for the entire sample. Generally, regions that overlapped with the canonical minimal lung space were chosen and those that did not were excluded.

**TRAINING PROCEDURE**

**Composite Loss Function**

We trained the neural network with a combination of the binary cross entropy loss function and dice coefficient loss function. Equations 1 and 2 are binary cross entropy loss function and Dice loss function respectively:

\[ H_p(y, \hat{y}) = \frac{1}{N} \sum_{n=1}^{N} [y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)] \]  

\[ \text{Dice} = \frac{2(y_n \cap \hat{y}_n)}{|y_n| + |\hat{y}_n|} \]  

Where \( y \) and \( \hat{y} \) are the expected and predicted pixel probabilities. Equation 3 is the composite loss function used to train the u-net.

\[ L(w) = H_p(y, \hat{y}) + (1 - \text{Dice}) \]  

The dice loss is combined with the binary cross entropy to improve the neural network training dynamics, enabling us to optimize the network on our primary evaluation metric directly. Also, the Jaccard similarity coefficient was calculated using the Dice score, as shown in equation 4.

\[ J_{acc} = \frac{\text{Dice}}{2 - \text{Dice}} \]  

**Data Augmentation**

Data augmentation included random applications of histogram equalization, translation, and rotation.

**RESULTS AND CONCLUSIONS**

We have presented a method for automated segmentation of lungs using a convolutional neural network. Utilizing a preprocessing workflow developed using the AFNI software suite and a modified u-net framework, a Jaccard similarity coefficient of 0.9485 was achieved after 18
hours of training. A sample of our results are presented in Fig. 5.

![Figure 5](image)

Fig. 5 A and B represent the lungs and the predicted masks respectively.

We are currently investigating extending the types of random augmentation performed on the training data to include adaptive histogram equalization, contrast stretching, and data/feature-space warping. Also, future work may entail modification of the u-net architecture to improve edge detection with the goal of improving accuracy of volumetric calculations on the CT lung scans.

REFERENCES


ABSTRACT
The Maximum Independent Set (MIS) is one of the fundamental techniques for solving optimization problem on graph. Optimization today is essential research tool in every areas of design, process, and science. The successful application of optimization processes in a broad range of practical problems, resulting from almost every domain of human activities that includes biomedicine, energy management, and telecommunications, are made on the basis of decision making tools.

Therefore, the objective of this research paper is to implement the MIS in modeling wireless ad hoc networks to elect cluster heads of wireless network and analyze their performance. In this paper, we will compare the performance of MIS algorithm with existing algorithms in clustering wireless ad hoc networks. These algorithms are Lowest-ID and Highest-Degree algorithms that are considered basic clustering algorithms similar to the proposed one in classification.

Keywords: Wireless ad hoc networks, Maximum Independent Set, Clustering algorithms, System modeling and analysis.

INTRODUCTION
Increase in wireless communication reliability and throughput is the result of modern improvements in wireless communication, considering the fact that portability, mobility, and accessibility are the driving force in wireless communication (Jafarkhani, 2005). In spite of the fact that stability, better performance, and higher reliability are now made available by wired communication, there is still that requirement that restricting users to a particular location or a limited environment. Wireless communication thus, has superiority over wired communication as a result of freedom that is made available by wireless communication.

In this research, graph is taken into consideration as a result of it being an essential tool in modeling of various kinds of relations and processes in several systems. Such system’s description can be derived by way of visualization that makes it possible for researchers to make predictions with regards to the situation of the system at any given period. This application of visualization is mostly possible when it comes to the systems of wireless communication, where the use of graph is possible in representing wireless nodes networks. The visualization of wireless system is first made by utilizing Computer Aided Design software prior to the commencement of the problem solving process.

WIRELESS AD HOC NETWORK
A decentralized type of wireless network made up of a number of wireless nodes fortified with a transmitter and a receiver is known as wireless ad hoc network (WANET). A network becomes an ad hoc only when it is dynamically constructed “on the fly” with no reliance on any pre-existing infrastructure like routers that are present in wired networks or access points that are present in managed (infrastructure) wireless networks. Instead, each node contributes in routing by having data forwarded for other nodes (Toh, Wireless ATM and Ad-Hoc Networks: Protocols and Architectures, 1997). Therefore, determining the nodes to forward the data is made dynamically based on network connectivity and the routing algorithm being utilized.

Wireless mobile ad hoc networks can be described as dynamic networks that are self-configuring, where the nodes have freedom of movement. Such Wireless network don't have infrastructure setup and administration complications, making the creation and joining of networks on the fly by devices possible at anywhere and anytime. Thus, there is high mobility of every component of an ad hoc network, thereby making the network’s topology to change dynamically as a result of this mobility. In wireless networks, the base station corresponds to the cluster-head in ad hoc networks. Nonetheless, it differs on the fact of the base stations being stationary while the cluster-head themselves have mobility as well. Unlike fixed wireless networks that are based in a master slave relationship form, the nodes only rely on each other in
WANETs for the establishment of communication links, and in acting as routers for the conveyance of data packets between source and destination pairs. The ad hoc network is also known as “multi-hop networks” because the data packet might need to go through a set of intermediate nodes in its way from a source to a destination node (Toh, 2002).

WIRELESS AD HOC NETWORK CLUSTERING

Clustering is being utilized in the partitioning of an ad hoc network into some smaller groups, but still the functioning of all the partitioned clusters acts as a whole. In every cluster, there is a cluster-head, some gateway nodes, and other ordinary nodes. If there is a need to transmit data to a destination node outside of a cluster, the cluster-head may pass it through or authorizes one of cluster’s gateways to do so, which results in more efficient way for data transmission. Moreover, clustering can be used in managing transmission, forming backbone, as well as routing efficiency. In short, network clustering can provide more efficient use of resources for large dynamic networks.

In (Basagni, 1999), Basagni clarifies that clustering mainly has two phases, which are the set up phase and the maintenance phase. Due to the equal and critical nature of the two phases, a brief description of each is provided blow.

Clustering setup phase

A setup process has to take place prior to the formation and partitioning of a network into small groups. Initially, every node in the network has to find the location of its own neighbors. Then, the clustering algorithm classifies network nodes into three types: Cluster-heads, gateways, and ordinary nodes based on their role in the network at a particular instant. There are many criteria such as unique ID, current location, and energy level that result in dynamically changing of node’s role in the network. Upon finishing the setup phase, a set of cluster-heads is elected as a dominant set that connect the whole network together. There will be dynamical change of the dominant set from time to time based on the network topology at any given moment (Basagni, 1999).

Clustering maintenance phase

Upon the completion of the setup phase, the algorithm will now move to the maintenance phase. Maintaining a network is a dynamic and critical task since the network spends most of its life time in this phase. The maintenance phase insure that all network’s nodes are connected to at least one of its neighbors. When there is no longer an active link between a node and its neighbor as a result of network topology change, the node will make attempt in reconnecting itself to another cluster-head. During this reconnection period, all active nodes update their neighbors list within their clusters by eliminating nodes that are no longer active as a neighbor and adding nodes that just join the cluster as a neighbor. In case a node failed to reconnect itself to an existing cluster-head within the dominant set, the clustering algorithm performs a network partitioning to new clusters and new dominant set will be elected (Basagni, 1999).

MAXIMUM INDEPENDENT SET

The MIS problem in graphs can be described as a basic optimization problem of graph that has different applications in numerous domains like coding theory and scheduling. In this research, MIS is adopted in order to come up with a visual definition of the problem and utilize the graph in inferring the answer. The MIS problem depends on finding the largest possible subset of vertices of a graph in a manner that no vertices are adjacent to one another by an edge; in other words, all vertices are independent of one another (Back & Khuri, 1994). The MIS modeling was studied to minimize required storage space in antenna systems (Güldal, Bilgin, & Tanik, Nov 2015). There is possibility of MIS becoming a useful tool in analyzing certain wireless ad hoc network properties. Thus, using MIS in modeling and analyzing certain properties of wireless ad hoc network can be a useful tool.

To demonstrate the MIS algorithm concept in graph theory, consider the example in Fig. 1 in which there are 6 vertices that are connected randomly. In this graph, the MIS is the subset or known as independent set that have three elements {B, D, E}. Since there is no independent set that has more than three elements, the size of MIS which know as maximum independency number, \( \alpha(G) \), is 3. One of the maximum independent set of this example is highlighted in red as shown in Fig. 1. Similarly, we can find all maximum independent sets of the graph by considering other possibilities. Another maximum independent set of this example with \( \alpha(G) = 3 \) is \{A, F, C\}. Thus, to reach all vertices of the graph, there are two sets of cluster-heads elected by MIS algorithm that can be used alternately without affecting the performance. In addition, MIS model seeks for providing highest redundant connectivity to cluster-heads nodes with rival performance.

![Fig. 1 MIS algorithm concept](image-url)
NETWORK CLUSTERING ALGORITHMS

In this study, we investigate other clustering algorithms used in clustering wireless ad hoc networks in order to compare them with the proposed MIS clustering algorithm in terms of the performance. These algorithms are Lowest-ID algorithm and Highest-Degree algorithm that are considered basic clustering algorithms similar to the proposed one in classification. These clustering algorithms are briefly explained in below.

The Lowest-ID algorithm

Also called “identifier-based clustering” as it classifies network’s nodes into three different roles: Cluster-head, gateway, and original nodes. This algorithm starts by assigning a unique ID to every networks’ node in the setup phase in order to elect a cluster-head. The node that becomes the cluster-head of the group is the one that is assigned the lowest ID in its group. Thus, all the other nodes in the group will have higher ID compared to that of the cluster head. A node lying in two different clusters will be set as a gateway, acting as a bridge to connect two or more clusters within the network and assist the cluster-head in passing data between clusters. In this algorithm, no cluster-head can be directly connected to each other, and a gateway has to be in-between to establish an inter-cluster session for data transmitting (Baker & Ephremides, April 1981; Baker & Ephremides, 1981).

Highest-Degree algorithm

Also called “connectivity-based clustering” which is among the first developed clustering algorithms utilized in ad hoc networks. Compared to the Lowest-ID algorithm, the Highest-Degree algorithm classifies network’s nodes into two different roles: Cluster-head and ordinary nodes. The main function of this algorithm is to control the nodes’ local traffic in the cluster. This algorithm begins with assigning a unique ID to every networks’ node in the setup phase in order to elect a cluster-head. The ID of the node is then broadcasted to other nodes lying within its transmission range as neighbors. As a result, each node will build its own neighbors list based on the received broadcasts. Then the algorithm will elect the group’s cluster-head based on the number of neighbors that a node has. In case of a tie in the number of maximum neighbor nodes, the node to be elected as that group’s cluster head is the one with lower ID. Similar to the Lowest-ID algorithm, it is not possible for two or more cluster-heads to simultaneously be neighbors to each other. The electing process is repeated until all the nodes that remain in the network turned into a cluster-head or join the cluster-head (Gerla & Tsai, 1995).

CASE STUDY

As a case study in this paper, we are considering a wireless network that are constructed randomly of 100 wireless nodes with 20 units coverage distributed randomly in an area 100x100 units². In the case study, we calculate the efficiency of clustering algorithm and compare it with similar algorithms. The clustering efficiency is calculated using below formula:

\[
\text{Efficiency} = \frac{(\text{Number of Nodes} - \text{Cluster heads}) \times 100}{\text{Number of Nodes}}
\]

Another, key performance indicator that we consider in the case study is the average redundancy that the clustering algorithm provides in connecting to cluster-heads. The average redundancy calculations as well as the clustering efficiency are carried out using Wolfram Mathematica package (See Appendix A).

The wireless network coverage that we consider in the case study is shown in Fig. 2. Each node coverage is represented in circle shape with 20 units radius where the node is located at the center. When there are two nodes close to each other with distance less than the node’s radius, this means there is connectivity between these two nodes. Such connectivity is represented in Fig. 3 for the entire wireless network. Since the graph is constructed, we can now use our tool to analyze the wireless ad hoc network for determining the cluster-heads nodes.

Fig. 2 Nodes’ coverage
Using MIS algorithm in clustering the wireless network results to identify 20 nodes as cluster-heads shown in Fig. 4. This means that these cluster-heads connect the entire network together. The efficiency of MIS algorithm is 80%, while the average redundancy is 2.16.

Also, we have implemented the Lowest-ID algorithm in clustering the same wireless network that we considered as a case study. Based on using this algorithm, there are 16 nodes are identified as cluster-heads; in other words, the efficiency of clustering algorithm is 84% as shown in Fig. 5. On the other hand, the average redundancy is 1.64, which means for each gateway or ordinary node there is only 1.64 available connectivity to cluster-heads in average.

Moreover, we have implemented the Highest-Degree algorithm to cluster the same wireless network we considered earlier for comparison purposes. The outcome of using this algorithm is identifying 15 nodes as cluster-heads that lead the efficiency in clustering to be 85% as shown in Fig. 6. The average redundancy when using this algorithm is 1.80 connectivity to cluster-heads.

The case study shows that all three clustering algorithms have comparable efficiency with advantage to MIS algorithm as it provides better network redundancy and alternately features. Such key features increase network availability.
Fig. 6 Cluster-head using Highest-Degree algorithm

SIMULATION AND COMPARISON

In this part of the study, we simulated each clustering algorithm to measure their performance over 1,000 randomly constructed wireless networks (See Mathematica implementation in Appendix B). The goal of this simulation is to measure efficiency frequency in general and with respect to change in number of nodes, nodes’ range, and area of the wireless network. Such simulation helps to build a solid understanding to the performance of the proposed clustering algorithm, MIS algorithm, compared to the existing clustering algorithms.

In general, the efficiency frequency of MIS algorithm is comparable to the Lowest-ID algorithm as well as the Highest-Degree algorithm. The efficiency range of MIS algorithm in this simulation is between 77% and 82%, while the most frequent efficiency is 80% as shown in Fig. 7. In addition, MIS algorithm generates different alternate sets that provides similar efficiency. Such alternate sets can take turn in order to increase network life. On the other hand, 84% is the most frequent efficiency recorded for the Lowest-ID algorithm in this simulation, while efficiency range is between 80% and 88% (See Fig. 8). The Highest-Degree algorithm shows efficiency range between 81% and 90%, and the most frequent efficiency is 85% as shown in Fig. 9. As a result, the MIS can be a rival algorithm in network clustering.

Finally, the simulation is conducted to measure efficiency performance algorithms with respect to change in number of nodes as well as nodes’ range and area of the wireless network. Overall, the performance of the MIS algorithm shows reactions similar to the Lowest-ID algorithm and the Highest-Degree algorithm with respect to the changes in number of nodes, nodes’ range, and area of the wireless network (See Fig. 10, Fig. 11, and Fig. 12).
CONCLUSION

We explored the potential of MIS algorithm for modeling and analysis wireless ad hoc network to improve system performance and increase network availability. Also, we went through some clustering algorithms used in wireless ad hoc network based on basic criteria, unique ID and current location, that the we considered in the propose clustering algorithm. We relied on studying theses algorithms using graph through converting a wireless network into graph representation to find the cluster-head based on each algorithm technique. Finally, we ended this study with simulation analysis and comparison to investigate the performance of the proposed clustering algorithm. This study shows that MIS can be a rival algorithm in network clustering, especially with high redundancy and alternately key features that increase network availability.

REFERENCES


**APPENDIX**

A. Wireless Ad Hoc Network Modeling and Analysis

The following code implements investigated clustering algorithms in wireless ad hoc network consists of 100 nodes with 20 units coverage distributed randomly in an area 100×100 unit², and calculates efficiency of clustering algorithms as well as average redundancy:

(*Define Variables*)

```plaintext
node = 100;
range = 20;
area = 50;
(*Distribute Nodes Randomly*)
x = Table[RandomReal[{-area, area}], {node}];
y = Table[RandomReal[{-area, area}], {node}];
center = Table[{x[[i]], y[[i]]}, {i, node}];
l = Table[{}, {i, node}];
con = {};
disk = Graphics[
  Table[{Style[Text[i, center[[i]]], FontSize -> 15], RandomColor[],
    Opacity[.5], Disk[{x[[i]], y[[i]]}, range]}, {i, node}]]
(*Define Nodes Connectivity in Graph*)
graph = {};
For[i = 1, i <= node, i++,
  For[j = i + 1, j <= node, j++,
    If[Sqrt[(x[[j]] - x[[i]])^2 + (y[[j]] - y[[i]])^2] <= range,
      AppendTo[graph, i <-> j]; AppendTo[l[[i]], j];
      AppendTo[l[[j]], i]; AppendTo[con, Length[l[[i]]]]];
  ];
map = Graph[graph];
(*Add isolated nodes that are not connected to other node*)
map = VertexAdd[map, Range[node]];
(*Represent nodes' coverage*)
index = VertexList[map];
map = Graph[map, VertexCoordinates -> Table[i -> center[[i]], {i, index}],
  VertexLabels -> "Name", VertexSize -> 5, VertexLabelStyle -> Directive[Bold, 10, Blue], ImagePadding -> 20]
(*Find Lowest-ID Nodes*)
id = Table[i, {i, node}];
clid = {};
For[i = 1, i <= node, i++,
  If[Total[Unitize[id]] != 0,
    AppendTo[clid, i];
    id = Table[i, {i, node}]
  ];
```
\[\min = \text{Min}[\text{DeleteCases}[\text{id}, 0]];\]
\[\text{AppendTo}[\text{lid}, \min];\]
\[\text{For}[j = 1, j \leq \text{Length}[\text{l[[min]]}], j++,\]
\[\text{id}[[1[[\text{min}, j]]]] = 0];];\]
\[\text{lidcacount} = \text{Length}[\text{lid}];\]
\[\text{lidca} = \text{HighlightGraph}[\text{map}, \text{lid}, \text{GraphHighlightStyle} \rightarrow "Thick"]\]
(*Find Highest Degree Nodes*)
\[\text{hd} = {};\]
\[\text{For}[i = 1, i \leq \text{node}, i++,\]
\[\text{max} = \text{Max}[\text{con}];\]
\[\text{If}[\text{max} \neq 0,\]
\[\text{p} = \text{Position}[\text{con}, \text{max}][[1, 1]];\]
\[\text{AppendTo}[\text{hd}, \text{p}];\]
\[\text{For}[j = 1, j \leq \text{max}, j++,\]
\[\text{con}[[1[[\text{p}, j]]]] = 0];];\]
\[\text{hdcacount} = \text{Length}[\text{hd}];\]
\[\text{hdca} = \text{HighlightGraph}[\text{map}, \text{hd}, \text{GraphHighlightStyle} \rightarrow "Thick"]\]
(*Find MIS Nodes*)
\[\text{mis} = \text{FindIndependentVertexSet}[\text{map}];\]
\[\text{miscount} = \text{Length}[\text{mis}[[1]]];\]
\[\text{miss} = \text{HighlightGraph}[\text{map}, \text{mis}, \text{GraphHighlightStyle} \rightarrow "Thick"]\]
(*Calculate the efficiency of clustering algorithms for wireless network*)
\[\text{lidcaefficiency} = (\text{node} - \text{lidcacount})*100/\text{node} /.\]
\[\text{x\_Rational} :> \text{N}[\text{x}, 2];\]
\[\text{hdcaefficiency} = (\text{node} - \text{hdcacount})*100/\text{node} /. \text{x\_Rational} :> \text{N}[\text{x}, 2];\]
\[\text{misefficiency} = (\text{node} - \text{miscount})*100/\text{node} /. \text{x\_Rational} :> \text{N}[\text{x}, 2];\]
\[\text{Grid}[\{\{"Efficiency of Lowest-ID Algorithm:" , \text{lidcaefficiency} "\%"\}\}
\[\text{Grid}[\{\{"Efficiency of Highest Degree Algorithm:" , \text{hdcaefficiency} "\%"\}\}]
\[\text{Grid}[\{\{"Efficiency of MIS Algorithm:" , \text{misefficiency} "\%"\}\}]
(*Calculate average redundancy*)
\[\text{lm} = \text{Table}[\{\text{i}\}, \{\text{i}, \text{node}\}];\]
\[\text{ccount} = 0;\]
\[\text{For}[\text{i} = 1, \text{i} \leq \text{node}, \text{i}++,\]
\[\text{count} = 0;\]
\[\text{For}[\text{j} = 1, \text{j} \leq \text{miscount}, \text{j}++,\]
\[\text{count} = \text{count} + \text{Count}[\text{l}[[\text{i}]], \text{mis}[[1, \text{j}]]];]\]
\[\text{AppendTo}[\text{lm}[[\text{i}]], \text{count}];\]
\[\text{ccount} = \text{ccount} + \text{count};]\]
\[\text{avm} = \text{N}[(\text{ccount} - \text{miscount})/(100 - \text{miscount}), 3]\]
\[\text{ll} = \text{Table}[\{\text{i}\}, \{\text{i}, \text{node}\}];\]
\[\text{ccount} = 0;\]
\[\text{For}[\text{i} = 1, \text{i} \leq \text{node}, \text{i}++,\]
\[\text{count} = 0;\]
\[\text{For}[\text{j} = 1, \text{j} \leq \text{lidcacount}, \text{j}++,\]
\[\text{count} = \text{count} + \text{Count}[\text{l}[[\text{i}]], \text{lid}[[\text{j}]]];]\]
\[\text{AppendTo}[\text{ll}[[\text{i}]], \text{count}];\]
\[\text{ccount} = \text{ccount} + \text{count};]\]
\[\text{avl} = \text{N}[(\text{ccount} - \text{lidcacount})/(100 - \text{lidcacount}), 3]\]
\[\text{lh} = \text{Table}[\{\text{i}\}, \{\text{i}, \text{node}\}];\]
\[\text{ccount} = 0;\]
\[\text{For}[\text{i} = 1, \text{i} \leq \text{node}, \text{i}++,\]
\[\text{count} = 0;\]
\[\text{For}[\text{j} = 1, \text{j} \leq \text{hdcacount}, \text{j}++,\]
\[\text{count} = \text{count} + \text{Count}[\text{l}[[\text{i}]], \text{hd}[[\text{j}]]];]\]
\[\text{AppendTo}[\text{lh}[[\text{i}]], \text{count}];\]
\[\text{ccount} = \text{ccount} + \text{count};]\]
\[\text{avh} = \text{N}[(\text{ccount} - \text{hdcacount})/(100 - \text{hdcacount}), 3]\]
B. Algorithms Simulation and Analysis

The following code simulates investigated clustering algorithms over 1,000 wireless ad hoc networks generated randomly to calculate efficiency frequency in general:

(*Define Variables*)
node = 100;
range = 20;
area = 50;
lidList = {};
hdList = {};

(*Distribute Nodes Randomly*)
misList = Table[x = Table[RandomReal[{-area, area}], {node}];
y = Table[RandomReal[{-area, area}], {node}];
center = Table[{x[[i]], y[[i]]}, {i, node}];
l = Table[{i}, {i, node}];
con = {};

(*Define Nodes Connectivity in Graph*)
graph = {};
For[i = 1, i <= node, i++,
  For[j = i + 1, j <= node, j++,
    If[Sqrt[(x[[j]] - x[[i]])^2 + (y[[j]] - y[[i]])^2] <= range,
      AppendTo[graph, i <-> j];
      AppendTo[l[[i]], j];
      AppendTo[l[[j]], i];];];
AppendTo[con, Length[l[[1]]]];]

(*Add isolated nodes that are not connected to other node*)
map = Graph[graph];
(*Find Lowest-ID Nodes*)
id = Table[i, {i, node}];
lid = {};
For[i = 1, i <= node, i++,
  If[Total[Unitize[id]] != 0,
    min = Min[DeleteCases[id, 0]];
    AppendTo[lid, min];
    For[j = 1, j <= Length[l[[min]]], j++,
      id[[l[[min, j]]]] = 0];];
AppendTo[lidList, Length[lid]];]

(*Find Highest Degree Nodes*)
hd = {};
For[i = 1, i <= node, i++,
  max = Max[con];
  If[max != 0,
    p = Position[con, max][[1, 1]];
    AppendTo[hd, p];
    For[j = 1, j <= max, j++,
      con[[l[[p, j]]]] = 0];];
AppendTo[hdList, Length[hd]];]

(*Find MIS Nodes and Repeat analysis for 1,000 trial*)
mis = Length@FindIndependentVertexSet[map][[1]], {k, 1000}];

(*Represent Results and Export Outputs*)
lidHistogram = Table[Count[lidList, i], {i, node}];
hdHistogram = Table[Count[hdList, i], {i, node}];
misHistogram = Table[Count[misList, i], {i, node}];

(*For MS Word*)
Export["L-ID node=" <> ToString@node <> "range=" <> ToString@range <> 
  "area=" <> ToString@area <> ".txt", lidList, "CSV"];
Export["L-ID node=" <> ToString@node <> "range=" <> ToString@range <> 
  "area=" <> ToString@area <> ". histogram" <> ".txt", lidHistogram,
"CSV";
Export["HD node=" <> ToString@node <> " range=" <> ToString@range <> 
"area=" <> ToString@area <> ".txt", hdList, "CSV"];
Export["HD node=" <> ToString@node <> " range=" <> ToString@range <> 
"area=" <> ToString@area <> " histogram" <> ".txt", hdHistogram, 
"CSV"];
Export["MIS node=" <> ToString@node <> " range=" <> ToString@range <> 
"area=" <> ToString@area <> " histogram" <> ".txt", misHistogram, 
"CSV"];
lidFreq = 0;
hdFreq = 0;
misFreq = 0;
For[i = 1, i <= node, i++,
    lidFreq = lidFreq + i*lidHistogram[[i]]; 
    hdFreq = hdFreq + i*hdHistogram[[i]]; 
    misFreq = misFreq + i*misHistogram[[i]]];
lidAverageFreq = (lidFreq/1000);
hdAverageFreq = (hdFreq/1000);
misAverageFreq = (misFreq/1000);
lidAverageEfficiency = N[(node - lidAverageFreq)*100/node, 2]; 
hdAverageEfficiency = N[(node - hdAverageFreq)*100/node, 2]; 
misAverageEfficiency = N[(node - misAverageFreq)*100/node, 2];
Grid[{{"Efficiency frequency over 1,000 trial"}}]
Histogram[lidList]
Grid[{{"Overall Average Efficiency of Lowest-ID Algorithm: ", 
    lidAverageEfficiency "\%"}}]
Histogram[hdList]
Grid[{{"Overall Average Efficiency of Highest Degree Algorithm: ", 
    hdAverageEfficiency "\%"}}]
Histogram[misList]
Grid[{{"Overall Average Efficiency of MIS Algorithm: ", 
    misAverageEfficiency "\%"}}]
SOFTWARE ARCHITECTURES FOR SMART APPLICATIONS WHICH MERGE ONTOLOGICAL REASONING WITH BIG DATA ANALYTICS

Radmila Juric
Department of Science and Industry systems,
University College of Southeast-Norway, Norway
Inhwa Kim
SAMSUNG, SDSE, UK
{ rju@usn.no , nina.k@samsung.com }

ABSTRACT

We propose a software architecture for smart applications, which utilizes two different types of computing and makes ad-hoc decisions about various situations we encounter in urban areas. The emphasis is on ontological reasoning using SWRL enabled OWL ontologies, in order to make quick decisions for an individual who uses the application. The data essential for the reasoning would be supplemented with instant collection and processing of live and user generated data with Big Data technologies and sourced from social media. We tested the architecture by collecting Twitter data with IBM Watson Analytics and processing it by using SQL like queries. By having an option of summarizing information available from social media with SQL queries upon Non-SQL repositories, and asserting or inferring them into OWL ontologies, we could infer OWL concepts and secure reasoning about various situations we encounter in urban areas. The proposed architecture generates applications which can be developed in either Java component based environments or Android.

INTRODUCTION

Software Architectures (SA) and SA styles with software patterns, have brought clarity to software development and reusability of their models (Bass et al., 1998), (Juric et al., 2004), (Gamma at al., 1994). After more than 20 years of focusing on SA, we are very much aware of their role and power in creating software applications, which answer business needs and technology demands. Whether we develop software applications at the enterprise level, create processing paradigms with new software technologies, or designing Apps for Android and iOS, we tend to exploit SA styles. We focus on the separation of concerns: from user interfaces and computationally powerful software components to enormous number, and varieties of data repositories (Granatir et al., 2007), (Bravhar and Juric, 2017), (Tarabi and Juric, 2018), (Shojanoori, 2014). The latter added another dimension to SA by taking into account that we generate data as we live, which puts Big Data (BD) technologies firmly on our agenda when designing SA. In other words, SA must make provision for processing the abundance of data and acknowledge that SA components should be deployable with Big Data technologies.

It is quite remarkable how BD technologies keep their configurations according to some software patterns and layered SA styles, which prove to be extremely efficient when developing software solutions at a more generic level (Williams et al., 2005), (Shojanoori et al., 2014). The significance of modern SA is in accommodating both new computational solutions and data sharing within traditional architectural styles (Bass et al., 1998), but without making restrictions on the
- type of data we use in computations,
- owners of such data (we do not have to compute with our own data) and
- specifications on which, when and how the data will be generated.

Therefore, the prerequisite in modern software development has remained the same, as in the last 20 years: we pay attention to SA style(s) which generate software application(s), with emphasis on
- (a) their reusability,
- (b) accommodating constant changes dictated by technology advances and
- (c) introducing new computational paradigms.

In this paper, we would like to focus on the creation of one type of software application, called smart, because it involves reasoning. It also accommodates constant changes in the environment where the applications run. We search for the best possible SA solution for such applications, which addresses (a) – (c) above. However, the term smart may trigger a separate discussion because of various interpretations of its meaning, particularly when ubiquity and pervasiveness start dominating our cyber and physical spaces (Shojanoori 2013), (Shojanoori et al., 2012). Smartness in software applications may be defined using different criteria, but in this research we rely on two important factors. Smart applications are either characterized by
- the abundance of data, which overflow our computational environments and bring various semantics, or
- new algorithms, which enrich existing computational paradigms, and have roots in the Artificial Intelligence (AI) and Semantic Web Technologies (SWT) (SWT, 2004), (Schwartz, 2013).
It has been very difficult to find related work for this type of research. The main problem is that we have not found any SA solution for merging Big Data and SWT. They either do not exist in peer reviewed publications or are buried in a variety of proprietary Big Data solutions which offer user-friendly interfaces, and hide data processing. However, we have found no evidence that SWRL enabled OWL ontologies and reasoning upon them have been incorporated into Big Data processing. From this perspective, our solution might be the first step towards merging different technologies for the purpose of creating new types of software applications.

We applied our proposal in the domain of urban emergencies. We have not found any evidence that SWT is used for decision making in urban emergencies, but software solutions which started dominating in the field of humanitarian crises in general, have embraced Big Data technology to a certain extent and use the SWT. There were no examples where both technologies were used in the same software application aimed at managing data in humanitarian crises.

**RESEARCH GOALS**

Our research goals are to:

(i) Create SA which generate smart applications for decision making in constantly changing environments;

(ii) Evaluate if BD technologies, which can collect and manipulate data generated in such environments, would work within the SA proposal;

(iii) Define a potential computational solution with SWT, which manipulates the semantics in such environments and support decision making.

Our goals above are tested in the management of the semantics in an urban incident.

Therefore, the smartness in our solution comes from the computational model with SWT, which directly depends on BD technologies to secure relevant data to feed the computations. These two are interwoven.

This paper is organized as follows.

In the next section we look at technologies which are essential for delivering (a) – (c), and achieving (i) – (iii). We justify the use of IBM Bluemix and Watson Analytics technologies (Bluemix and Watson) accompanied with SWRL (SWRL) enabled OWL (OWL 2) ontologies for proving the main concepts of this research. In the Scenario we explain an urban incident in London, UK, which was a case study for evaluating our proposal. We look at the potential of collecting live and user generated data and decision making during the incident. In the next section, we propose and explain the SA style which would generate software applications according to the goals from (i) – (iii). The SA is illustrated in the section which follows, and in Conclusions we elaborate on the contribution from this research, look at the scope for improvement and lessons learned, which would determine future work.

**CHOICE OF TECHNOLOGIES**

In this research, the technologies, which serve us well in incident identification and response, are related to two different branches of software engineering, which have not yet been placed together as one software solution. We need BD technologies, which collects and filters data relevant to the incident detection, and SWT technologies which would take the filtered data and make decisions on an adequate response to the incident.

The choice of Big Data technologies may be problematic (Panneerselvam and Juric, 2016), (Panneerselvam et al., 2016), because it dictates the way we collect and manipulate BD.

When comparing the Hadoop echo system and powerful components of BD management in IBM, through either Bluemix or Watson Analytics, we have chosen the latter for its user-friendly interfaces and assistance in their software configuration (Fotso and Juric, 2017). However, one of the most important aspects of our proposal is that

- We use BD technologies to collect efficiently relevant information and, at the same time, understand the semantics of the environments in which incidents happen, BUT
- We “transfer” the processing of collected BD to another computational environment in which we can compute by defining semantics in OWL and reason upon it with SWRL in order to define the response to the detected incident.

In the past, we have experimented with SWRL enabled OWL ontologies in urban traffic management decision making (Breznica and Juric, 2013), where the data relevant for decision making was collected through either user input or generated by sensors (Kataria et al., 2010). However, in this research, we focus on the collection and filtering of live and user generated data, which feeds the computational model based on SWRL enabled OWL ontologies, in order to make decisions and formulate responses for the incident detected through Watson Analytics.

In order to address the constant changes in such environments, we have to bear in mind that their semantic and reasoning may change from one moment to another. Therefore, the dynamism of such environments will be addressed with SWRL enabled OWL ontologies which will not accumulate vast knowledge, but will secure correct decision making in the moment we have relevant semantics about the incident. It will also allow us to change decisions in a matter of minutes if the semantics of the incident change.

This means that SWT and its languages are NOT used for creating shared vocabulary and formal ontologies. They enable us to create a sleek and portable software solution in which OWL concepts are supported by SWRL reasoning for defining and manipulating the semantics relevant to the incident. Reasoning with SWRL will be used for making decisions on the best possible response to the incident at the moment when it is detected.
If we wish to place these two disparate technologies BD and SWT together, within one software application, then we will have to bear in mind that

- SA style for such applications must focus on the philosophy and purpose of these technologies,
- An Integrated Development Environment (IDE) for the deployment of the SA should have plug-ins to use these technologies seamlessly and
- The technologies should allow us to maintain separation of concerns (Gamma et al., 1994), as defined in the SA.

**THE SCENARIO**

In our scenario, we looked at two incidents, which happened in Central London in March and June 2017. We collected live and user generated data from both incidents, as it was practiced in our earlier works, which justified the use of twitter platform in situations where incidents and crisis occur (Petai et al., 2015), (Juric et al., 2017), (Everiss et al., 2015).

We place the scenario in surroundings of ordinary citizens who happened to be either associated with the incidents or in their proximity. Our software solution is expected to collect relevant data from the twitter platform, in live situations, and create alerts in order to either protect ordinary citizens or help them to understand changeable situations within the incident. In our preliminary analysis, we had to focus on the collection and filtering of tweets, and either create and run SQL like queries with Bluemix or use Watson Analytics to answer queries and understand the content of the tweets, as experimented in earlier works (Juric et al., 2017), (Paneneeserlvam et a., 2016). In Table 1 we summarize the analysis of tweets by

- defining “questions” we wish to answer through the analysis (column 1)
- formulating expectations we may have when trying to answer these questions (column 2)
- itemizing questions from column 1 and defining actions which should be taken in order to answer them (column 3).

Therefore, Table 1 had a double purpose. It allows a manual content analysis of the tweets and, consequently, prepares us for filtering and processing tweets with Watson Analytics. Table 1 brought us important keywords and signaled the possible severity of the incident found in the content of tweets. The rightmost column, which describes actions becomes also very important: we could perform analysis with Watson Analytics according to the content of this column.

We performed the actions listed in Table 1 using Watson Analytics, connected to the Twitter platform, from 6pm to 11:59PM on 3 June, 2017, which collected 24,809 tweets for London Bridge attack. The same mechanism collected 11,460 tweets from 4 - 30 June, 2017. We were able to use Watson’s prepared questions and answer some of our questions from Table 1, which generated interesting outcome and could be used when issuing alerts within our application.

During the time of the attack, numerous tweets brought valuable information by

a) Explaining the current situation, specifying which areas are affected /closed including train/bus/tube stations and road routes in operation,

b) Informing if the incident is still ongoing,

c) Offering help to affected people (free lift or a place to stay, offering food, water, and medical help).

**Table 1 Process for Analyzing Tweets**

<table>
<thead>
<tr>
<th>Question</th>
<th>What to expect to find?</th>
<th>Broken down question and Actions taken for answering</th>
</tr>
</thead>
<tbody>
<tr>
<td>What was tweeted during attacks?</td>
<td>Are people asking for help? Explaining situations, warnings, describing attackers?</td>
<td>Which “word(s)” are used the most when the attack happened? - Examine words in tweets.</td>
</tr>
<tr>
<td>Who tweeted the most during each attack?</td>
<td>Who are the major tweetians, from which country? Are they authoritative professionals or individual citizens?</td>
<td>How many authorities (govnm., police or media), expected to be involved in disseminating information and guidance, were tweeting? Did non-authorities tweet? Did they disseminate the same info? - Examine owners of tweets/countries - Examine owners of tweets which were retweeted the most - Examine the subject of tweeting by counting words.</td>
</tr>
<tr>
<td>What was tweeted after each attack?</td>
<td>Did people intend to advise on the cause of attack, survival, terrorism, suspect, casualties, victims, condolences</td>
<td>Which ‘word(s)’ are used the most amongst tweets after the attack? - Examine words in tweets.</td>
</tr>
<tr>
<td>Who tweeted the most during attacks?</td>
<td>Who are the major tweetian: authoritative sector or individual tweetian.</td>
<td>How many authorities are involved in disseminating information and guidance, were tweeting? Did non-authorities tweet? Did they disseminate the same inform? - Examine owners of tweets/countries - Examine owners of tweets which were retweeted the most - Examine the subject of tweeting by counting words.</td>
</tr>
<tr>
<td>When tweeting peaked: on the date or after incident</td>
<td>Explain when public interest peaked: during, immediately after, in investigation, finding suspect</td>
<td>Examine the number of tweets on the date of incident and after the incident - Examine words in tweets</td>
</tr>
</tbody>
</table>
In the second collection of tweets, after the attack, the content of tweets significantly changed and therefore their semantics could not contribute towards any type of alerts. They disseminated aftermath information, commented on consequences of the incident, and included stories of bravery and altruistic behavior of citizens and authorities.

**Which alerts we May Issue and Why**

Table 2 shows a snapshot of semantics collected from Watson Analytics. We focused on the keywords, which
- have been found in relation to “situations” during the attack,
- could be used if we wish to issue “guidelines”
- could be used if we wish to formulate “help” in our alerts.

Table 2 directly contributes towards design decisions in OWL modelling and the definition of SWRL rules, because the semantics stored within it could be transferred into OWL classes, individuals and could contribute towards constraints in the form of object properties.

Finally, if we wish to model a situation in OWL in which we must issue alerts, then we will have to focus on “WHEN something happened”, “to WHOM the alert has been issued”, and “WHAT information should be included in it.” This is in line with the well know UN scheme on reporting and managing information during Humanitarian Crisis (Shamoug et al., 2012), (Shamoug et al., 2012A).

<table>
<thead>
<tr>
<th>Situations</th>
<th>Guide</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incident happened,</td>
<td>Emergency call number,</td>
<td>Anyone needs free lift to closest station,</td>
</tr>
<tr>
<td>Terror attack,</td>
<td>Support emergency,</td>
<td>Let me know;</td>
</tr>
<tr>
<td>Area(s),</td>
<td>Stay safe,</td>
<td>Shelter;</td>
</tr>
<tr>
<td>Eyewitness,</td>
<td>Diverted route,</td>
<td>Place;</td>
</tr>
<tr>
<td>Stay away from,</td>
<td>Traffic news,</td>
<td>Need a place to stay,</td>
</tr>
<tr>
<td>Serious incident,</td>
<td>Attackers moving route,</td>
<td>Need help,</td>
</tr>
<tr>
<td>Terrorist(s),</td>
<td>Run away,</td>
<td>Come for free,</td>
</tr>
<tr>
<td>Stabbing,</td>
<td>Stay away from incident</td>
<td>A flat to stay,</td>
</tr>
<tr>
<td>Drive into pedestrians</td>
<td>Remain where you are</td>
<td>Need somewhere to stay,</td>
</tr>
<tr>
<td>Emergency services,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surroundings,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accurate injury count,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disaster,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Injured,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Casualties,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public transportation,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stuck in the area,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic news,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long queue,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Station remain closed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**THE PROPOSED SOFTWARE ARCHITECTURES**

We propose a layered and component based SA style for deriving software applications, which help to meet the goals of this research. The collection, filtering and processing of live and user generated data, described in the previous section, has to find its place within the SA. However, the core of the SA model should point out where we allow for addressing research goal (iii). This is because the power of reasoning in SWRL enabled OWL ontologies will take place instead of processing live and user generated data solely with BD technology. Therefore, the computational model in our SA should show exactly where computations happen: is it with Watson Analytics or with SWRL upon OWL, and which data is being used?

Figure 1 is an abstract model of our SA proposal. Software components are layered, in order to follow the Model View Controller pattern, and named according to their role within these layers. They are all self explanatory.

Therefore:
- data sharing in the SA model is explicit: a repository called T1/T2 is being shared. It contains the semantics extracted from filtering and analyzing twitter data using Watson Analytics and could contain information stored in Tables 1 and 2.
- user interfaces and application management software components conform to a chosen Integrated Development Environment, i.e. NetBeans in our solution. Servlet software components in NetBeans may have exactly that role and SQL commands should be placed within Java Beans classes (if Enterprise Java Beans are not used).
- OWL-API, as a plug-in, available in NetBeans, secures a seamless transfer of computational power from Watson Analytics and Servlet technologies, available in IDE, to perform the reasoning with SWRL. The right most side of Figure 1 shows the transition of computational model: form SQL towards reasoning with SWRL, upon OWL concepts. A big circle denotes an OWL ontology with its concepts accompanied with a set of reasoning rules.

The left most part of Figure 1 shows that the collection and filtering of tweets could be done under the management of BD technology. It would allow the replacement of Watson Analytics and Bluemix with Hadoop echo system, if required, and would not affect the SA. The left most part of Figure 1 also corresponds to our activities performed in Section 3: they helped to extract more semantics from filtered Twitter data, create (T1/T2) repository and prepare them for OWL models and its concepts.

We have to draw reader’s attention that we performed the analysis of filtered twitter data using all services provided by IBM, including Bluemix. For reasons of simplicity, Figure 1 shows the use of Watson Analytics. Therefore the purple WATSON SQL component may have corresponded to running numerous and prepared questions by Watson Analytics. It can also indicate that we could have used SQL in Bluemix. The SA is not affected by these choices.

The feasibility of deploying the SA from Figure 1 is not questionable. SA with similar principles, but deployed with different technologies, have been used in quite a few software applications which needed SWRL enabled OWL ontologies for performing the main computations (Kataria 2011), (Tarabi et al., 2018), (Patadia et al., 2011).

In the next section, we illustrate the way we deploy the SA given in Figure 1 and create a software application. One of the most important aspects of the application development from the SA is the design of OWL model and the way we
transfer the semantics depicted in Section 3/ Tables 1 and 2 into OWL concepts. Consequently, we should be able to run SWRL rules to create alerts in a situation when an incident has been detected with Watson Analytics.

THE ILLUSTRATION OF THE PROPOSED ARCHITECTURE

In this section, we focus on the right most part of the SA from Figure 1 and look at the computational model which issues alerts. The model consists of giving definitions of classes in an OWL ontology, individuals, which populate these classes, and constraints between these individuals, which add more to the collected semantics stored in T1/T2. These are typical requirements for creating SWRL enabled OWL ontologies.

![Fig. 1 The Proposed Software Architectures](image)

We could create various OWL models from the semantics given in Section 3 and Tables 1/2 in particular. However, in this paper we stick to the content of Table 2 and convert column names into OWL classes and the contents of the columns into OWL individuals for one reason. We need to prove the concept that the deployment of our SA model and merging two different technologies within one application are feasible. The complexity of OWL model, and the number of individuals within it, does not have significant impact on the application and its performance, which houses the reasoning (Juric 2016).

Figure 2 shows basic OWL classes for our application, but there are numerous possibilities of creating different OWL models if we wish to perform reasoning differently (Almami et al., 2015), (Kataria and Juric, 20110), (Shojanoori and Juric, 2015).

**OWL Constraints**

The power of the OWL model from Figure 2 is in the definition of object properties, which exist between individuals of OWL classes. Table 3 shows just a selection of these object properties in order to illustrate the semantics they model. In a full scale commercialization of our proposal, this would be the most important part of the application in which various interested parties should be involved, including Police, Fire-brigades, government bodies and all involved in securing public safety/security. The semantics stored in object properties is a corner-stone of good reasoning with SWRL and creating meaningful alerts. Object properties can also

- be inferred, i.e. they do not have to be defined in advance,
- correspond to changes detected in the analysis of Tweets.

If the content of Table 2 changes, the update of object properties would be sufficient for accommodating these changes and Figures 1 and 2 would remain the same.

**Table 3 Object Properties for Reasoning Process**

<table>
<thead>
<tr>
<th>SITUATION (domain)</th>
<th>Object Property</th>
<th>GUIDE/HELP (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>is_good_for_S3</td>
<td>Stay where you are</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Remain calm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Call emergency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number XXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Do not get involved</td>
</tr>
<tr>
<td>S2</td>
<td>is_good_for_S2</td>
<td>Free taxi is in 50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>yards</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accommodation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>available at XXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Access to</td>
</tr>
<tr>
<td></td>
<td></td>
<td>information at</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XXXX</td>
</tr>
<tr>
<td>S1</td>
<td>is_good_for_S1</td>
<td>Incident at XXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shooting and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stabbing involved</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Casualties XXXX</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ambulance Directed</td>
</tr>
</tbody>
</table>

In Table 3 we illustrate object properties by categorizing individuals of domain class (SITUATION) as S1, S2 and S3.
Situation described as S1 would correspond to the stage when incident has JUST happened, and the first information about the incident has to be disseminated, through alerts.

Situation described as S2 would be after the incident was either resolved or under control, and assistance or help to affected people has to be disseminated.

Situation described as S3 would correspond to the “peak” of the incident where citizens are guided and instructed of their actions in order to protect themselves and contain the incident.

Table 3 is just one of many ways of defining object properties. In full scale commercial implementations we may replace the grouping of various factors which describe a “situation” (S1, S2 and S3) into definitions of properties between individuals of SITUATION class directly. We kept S1, S2 and S3, when testing the application, for one simple reason: we created these categories of SITUATIONS by refining the available data from Table 2, which helped us to differentiate between different types of alerts we may issue. Also S1, S2 and S3 directly trigger the reasoning with SWRL by being retrieved from the results of Watson Analytics and transferred into OWL ontological concepts (individuals and object properties).

The Reasoning Process

Reasoning Process in Figure 3 and SWRL rule in Figure 4 shows how a particular alert has been created. It is interesting to note that the rule is reusable, because it lists alerts according to the severity of situations: S1, S2 and S3. It relies on basic OWL classes and object properties imposed on their individuals. Therefore, as long as we adhere to OWL classes from Figure 2 and object properties from Table 3, we can run SWRL rule whenever we wish. We can issue alerts, triggered by a particular result, obtained through Watson Analytics.

The illustration of the reasoning process shows that we perform the inference upon OWL individuals. This means that individuals from Guide/Help OWL class are moved into the ALERT class if they satisfy the conditions specified in SWRL rule and in object properties.

OWL concepts and SWRL rules have been created in Protégé ontology editing tool. They are accessed from the application through OWL-API. The final result of the reasoning with SWRL is displayed within the UI Alerts software component shown in Figure 1.

Yellow line in Figure 3 shows the participation of individuals in object properties (Table 3) and the blue line shows participation of classes in the SWRL rule (Figure 4).

CONCLUSIONS

This paper illustrates the generic SA which allows the deployment of two different technologies within one software application. We illustrate the feasibility of merging Watson Analytics and SWRL reasoning upon OWL concepts, from a single Java application generated in NetBeans. Watson Analytics was performed, OWL concepts defined and reasoning ran, upon twitter data collected during one urban incident in London, UK. The outcome of reasoning is in the form of issued alerts according to the semantics of situations we detect in urban areas. In this experiment we relied on live and user generated data from the social media, but our proposed SA model would work in any other environment where we are able to collect and analyze any type of BD.

The Contribution

The novelty of this work is twofold.

Firstly, we tried to place two diverse technologies together within one SA but without any integration or changes in the way these technologies operate. IBM BD technologies were used for collecting live and user generated data from the Tweeter platform and extracting the semantics available in the collected tweets.

Secondly, the SWTs are used within the same SA in order to create OWL concepts from the collected semantics of tweets through Watson Analytics. We perform SWRL reasoning upon them and create inference, which ultimately issues alerts. Both technologies, BD and SWT, co-habit together within our application and the transition of computations from one technology to another is seamless. BD technologies work on the enormous amounts of data, but our selection of SWT use a small number, but semantically rich OWL concepts derived from the collected tweets and perform reasoning upon them with a relatively small set of SWRL rules.

In addition to the above, SWRL enabled OWL ontologies give an important feature to software applications which host them. We know that we can create OWL concepts by using assertions and manual entries in the ontology. However, we can infer anything we wish into OWL: from classes and individuals to object properties. A careful balance between assertions and inference would produce the best possible management of applications derived from Figure 1 for two reasons. We can
• instantly “absorb” the changes in the semantics of data collected during the incident and issue prompt alerts and 
• ensure that we reason on the most accurate data we can have.

Our computation, which issues alerts, will not depend on “the abundance of data”, as it happens in BD processing. It will be a rather small, but reliable set of OWL data and concepts, which would be sufficient for instant reasoning and creating alerts.

Scope for Improvement

Readers might not see a clear thread from Section 3 (The Scenario) to Section 5 (Implementation of the OWL model). This means that we should elaborate on the process, which automates the conversion of the results of Watson Analytics into OWL concepts and constraints. This part of the work has been done manually for one reason: we had to manipulate Watson Analytics according to the questions we expect to be answered by Watson Analytics (Fotso and Juric, 2017)) and it is unlikely that this type of data processing with BD technologies will change in near future. Therefore, Tables 1 and 2 are as close as we could get to bridge the gap between the results of running Watson Analytics and the OWL model we need to create.

More work has to be done to see if Hadoop echo system, which gives us more flexibility in processing BD could generate more suitable data for SWRL enabled OWL ontologies, as elaborated in (Panneerselvam and Juric, 2016), (Panneerselvam et al., 2016).

The commercialization of this type of applications does not require significant investment (Juric 2016), (Tarabi and Juric, 2018), (Shojanoori and Juric, 2013). The application created in NetBeans and run in Windows environments proved that our proposal is feasible. The conversion of Figure 1 into an Android application is also trivial because of the existence of OWL-API in Android Studio. However, if this type of applications are to be run in environments, which do not change often, but are exposed to significant amount of data, then we might not need OWL concepts and SWRL reasoning. We could be better off with any type of BD processing with BD technologies and by including SQL like queries upon NoSQL repositories in order to analyze the content of the tweets and reason upon it. Consequently, there should be a clear understanding on what we expect from BD and SWT technologies.

Our proposal will be equally reusable in any type of incidents where instant reaction of the software application is required, and alerts are inferred. However, Figure 1 must be revisited for reusing it in large-scale humanitarian crises.

Lessons learned

As we have already mentioned, it was difficult to find publications, which propose or use SA, for housing disparate technologies within one umbrella and, at the same time, use SWT outside the Web, solely for software engineering purpose. However, the lesson learned from this research is twofold.

Firstly, there are discrepancies between DB technology configurations and the way we extract the semantics from collected BD, which may affect the performance of our solution. Therefore our choice of using Watson Analytics in this particular study was unavoidable if we wished to stretch outside the BD environments and move the computations outside BD technology, towards SWRL reasoning. Watson Analytics proved to be a stable and user-friendly environment but did not give us the flexibility of extracting relevant semantics for OWL modelling from the collected tweets the way we wanted (Fotso and Juric, 2017).

The second lesson is that there is a lack of awareness of and interest in mature and stable SWT outside the web and formal ontologies (Juric, 2017). It has always been difficult to convince readers that our proposal will not generate big ontologies based on shared vocabulary. Our solution is actually quite the opposite.

REFERENCES


SWT Road Map, available at W3C website https://www.w3.org/2001/03-swv-1/slide7-0.html.


LEARNING SPACE MODEL WHICH ADDRESSES DIFFERENCES IN LEARNING

Eiman Almami, Radmila Juric1 Mohammed Zaki Ahmed
School of Computing and Mathematics, Plymouth University, UK 3HSN,
1University College of South East Norway(USN), Faculty of Technology,
Natural Sciences and Maritime Sciences, Kongsberg, Norway
{Eiman.almami@plymouth.ac.uk rju@usn.no M.Ahmed@plymouth.ac.uk }

ABSTRACT

This paper illustrates design decisions for creating Learning Spaces, which address differences in learning. These Learning Spaces are derived from the Reference Model ManSemLeS, which uses SWRL enabled OWL ontologies for managing the semantics of the learning spaces and meet individual learner’s needs. We explain and illustrate how our rich OWL domain model has been reduced to a manageable set of OWL classes and properties in order to address learning goals and differences in learning, at the same time. Our process of creating such learning spaces is reusable. We can take it into any learning environment where we need to model OWL concepts which secure reasoning with SWRL for inferring teaching practices for both learners and instructors.

INTRODUCTION

This paper is a continuation of our published work on the creation of Learning Spaces, which address differences in learning. The essential part of creating these learning spaces is the creation of the OWL ontological model design, which have been in the focus of our research interest for many years. Therefore, readers interested in OWL modelling and its role in creating Learning Spaces, are advised to read all our previous publications which appear in the reference list. In this particular paper, we elaborate on the rationale behind OWL ontology design decisions, which is a very important part of the core computational and reference model, used for creating Learning Spaces. The motivation for writing this paper is threefold:

a) There are no formal or any other type of OWL ontologies, in the published literature, suitable for determining which semantics is relevant when creating Learning Spaces, which address differences in learning.

b) There is no known consensus on how to grasp relevant semantics in any problem domain and model it as a SWRL enabled OWL ontology, for the purpose of addressing a very specific nature of Learning Spaces: they are likely to be cyber physical spaces which are to be created with the support of software intensive solutions.

c) There have been changes in our OWL modelling design decisions throughout the years, in order to determine which semantics of the problem domain would be modelled as OWL classes, individuals or properties.

Therefore, our final OWL design decision had to be made in order to create a final solution: a software application which creates a particular Learning Space.

d) The implementation of the software architectural model, which generated the application for creating a particular Learning Space, has been successful and therefore it proved that our OWL design decision could be placed firmly within the core computational model based on reasoning upon SWRL enabled OWL ontologies.

The paper is organized as follows.

In the next section we describe the reference model ManSemLeS for configuring a Learning Space (LeS), which was introduced in our earlier work, but have matured in order to accommodate our final OWL modelling design. In the next few sections we describe the ManSemLeS constituent modules: The Domain Model module (DMM), The Learning Space Model Module (LeSMM) and the Instructional Model Module (IMM). Obviously, the rationale behind the OWL modelling is placed within the DMM section and our illustration of LeSMM and IMM summarizes which exact publications form our previous work have to consulted for evaluating the power of OWL modelling presented in this paper. We finish the paper by specifying in Conclusions, which exact research breakthrough we achieved and what can be expected from this research in future.

THE REFERENCE MODEL FOR CONFIGURING LES

Figure 1 is an illustration of the reference model, which has been adopted from our previous publications Almami, Juric, Everiss, and Ahmed (2013), (Almami, Juric, & Zaki Ahmed) and illustrates the role and the content of the computational model named Managing Semantics of Learning Spaces (ManSemLeS), which secures an ad-hoc configuration of a LeS.

The ManSemLeS, is divided into three distinctive parts:

- the GUI-APPLICATION components which contain user interfaces which may exist when creating LeS,
- the MODEL components, which store domain information relevant for the creation of LeS and the INSTRUCTIONAL model (IMM)
- and the MANAGEMENT components. which manage the semantic stored in various model components, carry out the configuration of a LeS and
prescribe teaching and learning practices which become parts of a particular LeS.

GUI-APPLICATION is essential for communicating and formulating demands upon the ManSemLeS and for displaying the computational output from the ManSemLeS.

It is important to note that the core of the reference model form Figure 1 is its layered computational core (LeSMM), which shows the way of defining, storing, and manipulating the semantics of LEs when configuring a particular LeS.

Therefore, it is very likely that the set of LeS, within the IMM model, at the bottom of the proposed referenced model (highlighted in yellow), is a cyber/physical space, which is possible to configure through the computational core LeSMM. This means that the model allows for all possible situations: from creating a physical, and traditional classroom to purely virtual LeS which have been configured strictly according to the user’s needs and expectations (learner’s and instructor’s demands).

In spite of changes in the the ManSemLeS model from its first publication, we have not change the main purpose of creating LeS. Therefore, Figure 1 distinguishes between two different types of ManSemLeS users: Learners and Instructors. Learner enter LE because they have their own goals and demands in terms of

(a) willingness and need to participate in a particular learning session
(b) need to express their personalized needs as learners (academic and personal profile)
(c) need to articulate any DiffInL they may have.

On the other hand, instructor, enter LEs because there is a request for them to participate in a particular session which will then include

(1) instructional design of a learning process, including learning sources/materials,
(2) assessing a learner’s performance and learning outcomes.

One of the most important part of the ManSemLeS is a domain model (DMM). It contains a complete semantics essential for computations and creation of a particular LeS and forms the foundation for understanding the role of Learning Space Model (LeSMM).

THE DOMAIN MODEL MODULE (DMM)

The Domain Model Module (DMM) Figure 2 contains ontological concepts and their hierarchies which belong to any LE (Almami & Juric, 2011b).

These concepts are likely to contain the basic semantics of any LE and should include a personal learner’s profiling with emphasis on DiffInL and their impact on the learning processes. Therefore, the DMM should store concepts and their relationships, which describe any possible situation that may be encountered in the LE. The semantics of DiffInL and its role in creating an LeS, must be understood and interpreted.

It is expected that DiffInL be modelled from various perspectives: the individual learner, types of differences he/she may exhibit in learning, goals of a learning session and the sequence of activities that must be performed in order to claim that this is an effective LeS for a particular learner.

It is important to note that the ontological model from the DMM is not a formal domain ontology, which has already been elaborated in (Almami, Juric, & Ahmed, 2016; Almami et al., 2015a; Almami, Juric, Zaki Ahmed, & Dabbour, 2015b; Almami, Almami, & Juric, 2016 ). In the literature, when the term “domain ontology” is used, a formal definition of all concepts and relationships between them is meant.

Fig. 1 The Reference Model for Creating LeS
Whereby application-specific ontology is referenced as domain independent model that in collaboration with domain model serves a particular problem. In this research the domain ontology is somewhere in between. It is not a formal ontology and is not application-specific ontology. It is generic enough to address learning needs of students with learning difficulties, whilst not comprehensive in terms of its axioms to represent the entire LE concepts.

On this basis the domain model this research represents a type of knowledge base (KB) that is comprised of both T-Box T, set of axioms about the concepts and relationships between them, and A-Box A, set of individuals that are member of some concept. This is to say that the KB = < T, A > If the LE domain is represented as LEΔ then it can be said that all concepts C ⊆ LEΔ, that are the domain individuals i ∈ LEΔ, and that all relationships r ⊆ LEΔ × LEΔ. This refers to a binary relationship r that a class per and class i of C have with each other:

\[
\text{per} \rightarrow i : r \subseteq \text{LEΔ}
\]

For example, PERSON is a concept within C, that is PERSON ∈ C. An individual i of PERSON is denoted by PERSON (i). And where the individual is shown as a variable x it is denoted as PERSON (x). In the LEΔ every individual of PERSON has some type of impairments. That is to say that

\[
\forall \text{per}(\text{PERSON}(< \text{per}>) \rightarrow \exists i(\text{IMPAIRMENT}(i) \land \text{has i}))
\]

This is shown in Figure 3. In Description Logic terminology, that is to say that for all per, that is a PERSON, definitely there must exist some IMPAIRMENT i and that the individual per has i. The expression is not of course a universal one that applies to every single human being. The expression is only true within the LEΔ : T ⊆ LEΔ. In LEΔ we are only concerned with concepts which are sufficient for describing basic semantics of the environment. The representation of LEΔ can be strengthened through various other OWL concepts by introducing sub-hierarchies of existing classes, and relationships. For example, as shown in Figure 3

\[
\text{PERSONAL_PREFERENCES} \subseteq \text{PREFERENCES} \subseteq \text{PERSON}
\]

In OWL terminology it reads as PERSON subsumes PREFERENCES which in turn subsumes PERSONAL_PREFERENCES.
Figure 4 shows an improved initial ontological model form Figure 2, with the choices of basic ontological classes, still self-explanatory, but debatable. Figure 4 follows the earlier experience of modeling the semantics of pervasive environments, (Almami & Juric, 2011b) (Almami, Patel, Koay, Juric, & Everiss, 2012) (Almami, Juric, & Ahmed, 2014) (Almami et al., 2015a; Almami et al., 2015b; 2016) where people, devices, technology and services delivered in them are interwoven. It is obvious that there will be semantic overlapping between sub-hierarchies of LEARNING ENVIRONMENT, PERSON and TECHNOLOGY classes, i.e. when describing the LE∆ it must be taken into account that its semantics can be juxtaposed to:

i) learners/instructors profiles and their expectations from LE and
ii) technologies which are an important part of any LE.

Semantic overlapping secures inferences, i.e. it can trigger reasoning, which will determine exactly which characteristics a configured LeS should have if it is expected that learners and instructors will participate in it.

It is important to note that it was not possible to find any suitable domain ontologies, which could help to describe the semantics of LE and LeS. In other words, it was not plausible to look at any existing attempts to formulate the formal semantics of LE with formal languages like OWL. However, this is not a particular concern of the researchers for two reasons (Almami et al., 2016; Almami et al., 2016):

(i) Domain specific and formal ontologies are difficult (if not impossible) to use in modern software applications, which require certain performance characteristics in terms of their ad-hoc creation, responsiveness to the environments where they reside (LE and LeS) and their implementations on wireless/mobile devices,

(ii) we should be aiming at creating our own domain specific ontology for any LE and it is not guaranteed that we will carry on inferring knowledge which will become persistent, because we may be in a position where storing and reusing knowledge is not desirable.

These to bullets above show that our intention is not to build an expert system, nor is to create any persistence, which grows proportionally to the level of inference mechanism defined upon the researcher’s designed ontology when configuring LeSs.

In contrast, the researcher’s domain ontology is broad and only the required classes, which are extracted from Figure 4, with only a fraction of individuals asserted and all results of inference deleted, after running the application created upon the ontology, i.e. after a particular LeS is configured.

**The Ontological Model**

The discussion in the previous section makes an impact on the ontology design. We will look at the horizontal hierarchies of Figure 4 and try to explain semantics behind their design decisions.

Figure 4a, is an excerpt from a full scale set of ontological concepts which have been elaborated in (Almami & Juric, 2011a; Almami & Juric, 2011b). Each of the leaves of the ontological structure from Figure 4 has been extended into sub-hierarchies, according to various domains of interests and the purpose of the ontologies when resolving problems in learning environments. The ontology has been created for the
purpose of satisfying the objectives of this paper and therefore the choice of axioms and inference which accompanies the reference model is dictated by the three main conditions, when determining how to convert the semantics of LE into OWL concepts:

a) There has to be sufficient set of classes and their hierarchies to describe the basic semantics of a LE to achieve LEA,

b) It should be possible to extend the model from a) and assert/infer more concepts (sub- hierarchies, properties and individuals) in order to strengthen the semantics of the ontology according to ‘demands’ imposed by learners and instructors,

c) There should be a way to secure a reasoning mechanism upon the ontological concepts from b) in order to create inference and configure LeS.

It is important to note that the model shown in Figure 4 has been re-used in several experiments for decision making in learning environments. Therefore, the main ontological concepts and their hierarchies should contain overlapping semantics in order to enable the reasoning with OWL and SWRL. It is possible to describe the semantics of any learning environments if one uses sub-hierarchies of PERSON, LEARNING_ENVIRONMENTS and TECHNOLOGY and enrich them with constraints. However, the exact choice of sub-hierarchies, their further extensions and choice of their classes from Figure 4, which will be involved in reasoning, depends on the need to perform decision making in order to create a particular learning environment.

In other words, the extensions of ontological leaves from Figure 4 becomes essential. For exploiting the ontology in a particular decision making process and within a particular domain of interest. For example, in (Almami et al., 2012) it became necessary to extend horizontal hierarchies of some of the existing classes of Figure 4 as shown in Figure 4a - Figure 4c.

The reason why only DISABILITIES, LEARNING, ASSISTIVE TECHNOLOGIES and TYPE_OF_ACTIVITY classes were chosen was due to the fact that they carry enough semantics to address the given scenario (Almami et al., 2012). The DISABILITY class from Fig. 4 Extensions of classes

Figure 4a has a set of subclasses, but TYPE and its subclass CONSEQUENCES store the exact information on any possible manifestation of Autism Spectrum Disorder (ASD) which might be found in pupils, which are relevant to the teaching and learning process and tasks performed in the classrooms.

The TYPE_OF_ACTIVITY class from Figure 4b has a set of subclasses, but the LEARNING_IN_THE_CLASS and CLASROOM_TASKS store the exact information on all possible activities that may run in the class for the purpose of ‘learning’, i.e. t. the instructor needs to de ne which particular academic task he/she would like to perform in the classroom as a part of learning activities.

The LEARNING class from Figure 4c has a set of subclasses, but INSTRUCTIONS and its subclass PROPOSED PRACTICES store the exact information on how to perform the academic task in terms of having a set of instructions which would be proposed as practices in order to secure learning outcomes.

The instructor will adopt instructions, advice, recommendations and suggestions in the classroom as ‘proposed practices’ in order to perform a desired academic task.

While the intensity was to enrich the T-Box T as opposed to the A-Box A, it became evident that further horizontal extension of ASSISTIVE TECHNOLOGIES would unnecessarily complicate the ontological model. Therefore, all possible assistive technologies were added to the A-Box as the class individuals.

In another experiment in a particular situation in LE, the researcher wanted to see if the materials were ready for a particular class and if they would serve students with LDif. In OWL/SWRL terms the aim was to capture the semantics of such LE in ontological classes from Figure 2 and reason upon it in order to answer a particular competency question:

“What materials, in terms of their content and format, are ready for accommodating students with dyslexia and Attention Deficit Hyperactivity Disorder (ADHD), for the purpose of running a session on Social Intensive Systems, taking into account that these students have particular interests in Social Network Privacy and Security (SNPS)”?

The premises of the above competency question in Description Logic format would be:

PERSON ⊓ (∃ has Dyslexia. DISABILITIES) ⊓ (∃ has ADHD.DISABILITIES) ⊓ (∃ interested In SNPS.PREFERENCES)

To answer the question, the following steps were taken:

- firstly, to extract which classes from the ontology will be sufficient for answering the question and are involved in reasoning;
- Secondly, to define constraints, value and existential constraints (∃ and ∀) on the selected classes in order to strengthen the semantics of the ontology and prepare it for reasoning;
thirdly, to run SWRL rules upon these selected classes to find the answer to the competency question.

The selection of classes which is solely based on the competency question above is shown in Figures 5.

The possibility of horizontal extension of DMM according to specific needs has been shown here. As far as OWL and SWRL are concerned each instance of an ontological model answers a specific competency question through inference. Thus, the issue of competency questions is probably the main rationale behind deciding exactly which OWL concepts are going to represent which part of the semantics of LE. It has been decided that the term DISABILITY should stay within the ontology as a vehicle which can then allow situations to be modelled when:

a) ‘disability’ causes LDif, or
b) ‘disability’ can be connected to DiffInL, or
c) ‘disability’ is NOT connected to DiffInL, LDif etc.

This will help to mark which ontological concepts will carry the semantics of disabilities in general, and which will be related to LDif and/or DiffInL by either extending classes from Figure 4, through their sub-hierarchies or imposing constraints on the ontology.

The rationale behind modeling the PERSON sub-hierarchy is dictated by the purpose of the ontology: it should be possible to match individuals from the sub-hierarchy of the DISABILITIES subclass with any other individuals in sub-hierarchies from the Learning Environment and services delivered within them.

Therefore, whichever set of subclasses are created when describing disabilities in the ontology, it should be borne in mind that they have to be addressed in LEs through various activities/practices, which may have a form of service delivered within a personalized LeS.

In other words, a ‘space’ must be found for dealing with DiffInL and instructional design, and the author must also must be able to manipulate the semantics of the ontology in order to secure a) c) above.

**Extending the DMM Classes**

In order to strengthen the semantics of the DMM ontology for the purpose of addressing the ‘demand’ when configuring a LeS, it is sometimes essential to extend leaves of classes of the ontology in Figure 4 with sub-hierarchies. In this subsection one example of creating a deep sub-hierarchy for the DISABILITY class, is given in Figure 6.

The rationale behind modelling the PERSON sub-hierarchy is dictated by the purpose of the ontology: it should be possible to match individuals from the sub-hierarchy of the DISABILITIES subclass with any other individuals in sub-hierarchies from the Learning Environment and services delivered within them.

Therefore, whichever set of subclasses are created when describing disabilities in the ontology, it should be borne in mind that they have to be addressed in LEA through various activities/practices, which may have a form of a service delivered within a personalized LeS. In other words, the ontology should contain the semantics which will enable all three objectives of this research to be delivered.

A ‘space’ must be found for dealing with DiffInL and instructional design, we also must also be able to manipulate the semantics of the ontology in order to secure a computational model from (ii) and (iii) in the Introduction.

There are seven main subclasses of the DISABILITY class as shown in Figure 6 above:

- TYPE ⊆ DISABILITY,
- CAUSE ⊆ DISABILITY,
- SUPPORT ⊆ DISABILITY,
- VISIBLE SIGNS ⊆ DISABILITY,
- COGNITIVE SIGNS ⊆ DISABILITY,
- REACTION_AND_FEELING ⊆ DISABILITY,
- LEARNING_DIFFICULTIES ⊆ DISABILITY.

These subclasses are disjoint classes: Disjoint Classes (TYPE, CAUSE, SUPPORT, VISIBLE_SIGNS, COGNITIVE_SIGNS, REACTION_AND_FEELING, LEARNING_DIFFICULTIES)

They are chosen for the purpose of having a provision within the ontology where we can store the semantics of DiffInL.

The term LEARNING_DIFFICULTIES was left as a subclass of the DISABILITY class within the ontology for one important reason: it will enable the author to decide, when creating a particular LeS, whether DiffInL will be modeled as an individual of the LEARNING_DIFFICULTIES class, a constraint between the PERSON and LEARNING_DIFFICULTIES classes or a simple subclass of the LEARNING_DIFFICULTIES class.
This decision can be made only after a particular ‘demand’ imposed by learners and instructors is analyzed. Furthermore, we might need all these subclasses of the DISABILITY class for ontological matching, i.e. the inference mechanism, which ultimately configures a LeS. The flexibility of OWL allows for a choice of classes from Figure 4 to be involved in the reasoning.

When describing the semantics of LDiff/DiffInL within the DMM ontology, more emphasis might be put on constraints instead of sub-hierarchies. The same applies to any of the ontological design decisions: the exact modeling decision on ‘which OWL concept will model what’ is always specific to a particular configuration of LeS (‘demand’) (Almami et al., 2012). The deepest hierarchies in the DMM ontology have been created for types of disabilities and their support. Relatively shallow hierarchies for CAUSES needed necessary.

In terms of modeling cognitive signs, these can be easily categorized as individuals in OWL (from the available literature) and therefore a deeper hierarchy for them was not felt necessary.

The richest part of the DISABILITY sub-hierarchy is a sub-hierarchy of the class COM-MON_TYPE_OF_LEARNING_DISABILITIES. COMMON_TYPE_OF_LEARNING_DISABILITIES ⊆ TYPE ⊆ DISABILITY

It should be noted, as it will be seen shortly, that it is always possible to balance these deep horizontal hierarchical subclasses against object properties that were referred to earlier as binary relationship between two, which can be defined on the ontological classes and their sub-hierarchies.

THE LEARNING SPACE MODEL MODULE (LESMM)

The Learning Space Model Module (LeSMM) is derived from the DMM, which is created as a result of a particular ‘demand’ upon LE, issued by a learner. This means that we extract from the ontological model stored in the DMM only class hierarchies, T-Box T, and their possible individuals, A-Box A, relevant to the learner’s demand, and infer sub-classes if required by the ‘demand’.

The ontological model in LeSMM becomes learner specific and contains only the semantics which is relevant in a particular configuration of a LeS. However, it is not enough just to extract relevant concepts from the DMM; constraints need to be added, or any other means of improving the semantics of the LeSMM as a prerequisite for defining instructional and learning concepts, as part of the configuration of a LeS.

THE INSTRUCTIONAL MODEL MODULE (IMM)

The Instructional Model Module (IMM) uses the ontological model from LeSMM and runs reasoning rules upon LeSMM concepts in order to specify exactly: which instructions are needed for creating LeS, what are the exact learning/teaching activity sequences for that particular LeS, where the learning sources and materials are, what their content and format is and how they can be supported.
The Instructional Model Module (IMM) should have the power of managing repositories and reasoning according to their purpose. Semantics of LE are managed by creating repositories and constraints within the LeSMM through assertion and inference and finally, the process of teaching and learning is managed by generating instructional learning design in IMM as the results of inferences, which in turn configures a particular LeS.

**ILLUSTRATIONS OF LESMM AND IMM**

The LeSMM module is responsible for interpreting the semantics of the ‘demand’ and extract relevant classes from the DMM ontological model and enrich them with additional sub-hierarchies, properties and individuals, which may be inferred or asserted. The IMM module will take the semantically rich ontology prepared by the LeSMM module and perform reasoning. It is difficult to illustrate these two tasks without being ‘demand’ and learner specific. The only possibility in this research is to illustrate types of demand, which may be issued, and which will then gear the configuration of a LeS. There may be a number of ‘demands’, which include the following:

- Which measures need to be taken to address the problem of avoiding “eye contact” when teaching autistic students?
- What needs to be done to ensure that ADHD students are following instructions in a particular classroom?”
- As an instructor in a higher education institute, are materials ready in terms of their content formats and availability on hand-held devices, for a debate on Social Networking in Healthcare, and scheduled for all students regardless of their DiffInL such as Asperger, dyslexia, or visual impairments?
- What would be the exact set of instructions (including choices of technologies) for children who are learning ‘reading’ in a mixed class of pupils with different mother tongues, DiffInL and different exposure to spelling?
• How can a correct selection of all possible hardware be made, (e.g. devices) and software solutions, which assist in a LE, which make provision for a clearly declared DiffNL?
• Some of students for a class at 3 PM are autistic – are the environments ready to guarantee their participation?

Each of these ‘demands’, whether taken in groups or individually, carries semantics which help us to choose excerpts of ontological classes from the DMM ontology, assert either individuals and properties upon them in LeSMM and then infer whatever is needed to answer the ‘demand’ through the IMM. For a full illustration of examples, how different demands create different ontological models based on Figure 4, and how OWL concepts are asserted and inferred in that model for preparing it for reasoning, please refer to to (Almami & Juric, 2011a) (Almami & Juric, 2011b) (Almami et al., 2012) (Almami et al., 2013) (Almami et al., 2014) (Almami et al., 2015a) (Almami et al., 2015a; Almami et al., 2015b).

CONCLUSIONS
Research Breakthrough

The core of this research is the LeSMM of the Computational Model. It is this layer of the software architecture that sets the foundation for any reasoning that will take place to support, in the most appropriate way, a user with LD. Therefore, it was felt not only worthwhile but also necessary to spend some time identifying the alternatives.

A good software engineering solution is often determined by the extent of interactions required from a user. Having this in mind, led us to revisit the structure of data-property intensive ontology required for the representation of learning spaces. The more ontology concepts entail eventually the more input from user. Considering also that the ultimate users are pupils with LD and instructors in an educational institution, and that the duration of any teaching session is limited, there is not enough time to be spared for extensive interaction with a system only to prepare the environment for teaching and learning.

In the author’s earlier experience, ontological structures were heavily dependent on data properties (Almami & Juric, 2011b). Considering the fact that broad use of data properties in ontological classes would require extensive user input to insert values for data properties that are dynamically defined, it was necessary to move away from data property insertions.

This observation has resulted switching to object properties as a dominant factor besides classes of the ontology. When data properties are used, the emphasize of the model is on individuals of classes, but when the attention is more towards the relationships between individuals of classes then data type properties should give way to object properties.

The author has briefly pointed out what object properties are. However, in view of the reliance of the rest of this paper on object properties, it is worth mentioning a few words about these properties. Unlike data properties that define individuals of a class, object properties define relationships between individuals of two classes.

It is the OP that determines who/what should be the object and who/what should be the subject of the relationship. In other words, once an OP is defined, for any instance of the OP, it can be inferred without any hesitation what class the object and the subject of the relationship are member of.

The universality of OP that once it is used the class membership of individuals at each side of the relationship is known, and that object properties are unique within the ontology, it makes models more reusable comparing with data-property oriented ontologies. Or formally expressed as:

\[ \forall R.C : \{per | \forall iR(per,i) \rightarrow i \in C\} \]

Where R is an object property, C is a concept at the end side of the relationship (or OP), per is an individual at one side of the object property, and i is an individual at the other side of the object property.

In addition to moving from data property to OP, ways of reducing number of classes required for representing any learning space need to be found.

This could not be possibly overlooked as it would have negatively impacted the end result, i.e. users would have to input endless amount of information to start off the system. Therefore, the attention to OP and less ontology classes whilst not sacrificing necessary data was a breakthrough.

New Design Foundation Classes

We had to agree on a set of minimum number of classes without which the representation of the learning space would be impossible. At the centre, of course, is the user, or in this case LD pupils who have been represented in a more generic term PERSON. Therefore, PERSON is the first class in the foundation set of classes in the LEA:

\[ \text{CA} = \{ \text{PERSON} \} \]

The driving motivation behind this research has been serving pupils with impairments. Therefore, there must be a class to accommodate different types of impairment, hence the class IMPAIRMENT:

\[ \text{CA} = \{ \text{PERSON}, \text{IMPAIRMENT} \} \]

There are specifically defined practices associated with each kind of impairment. There might be several practices that suit a particular impairment, or a specific practice might serve several impairments. This, therefore, suggests another class, PRACTICE.

\[ \text{CA} = \{ \text{PERSON}, \text{IMPAIRMENT}, \text{PRACTICE} \} \]

At any given time a pupil with LD may require different need. In other words, they may have different goals to attend to in different sessions. The model should address this requirement. It was named as GOAL as the fourth and last class in the foundation set. Therefore, the complete foundation set C for the LEA contains four classes:

\[ \text{CA} = \{ \text{PERSON}, \text{IMPAIRMENT}, \text{PRACTICE}, \text{GOAL} \} \]

It has to be added that the above classes of CA individuals are disjoint classes with respect to the T-Box T. That is the intersection of each two pairs results in an empty set as shown in Table 1 below.
Table 1 Empty Sets of Disjoints Classes of C\(A\)

| PERSON II IMPAIRMENT \(= \emptyset\) | PERSON II PRACTICE \(= \emptyset\) | PERSON II GOAL \(= \emptyset\) | IMPAIRMENT II PRACTICE \(= \emptyset\) | IMPAIRMENT II GOAL \(= \emptyset\) | PRACTICE II GOAL \(= \emptyset\) |

In summary, following the implementation of several models addressing different competency questions, it became evident that rearranging some of the high level classes and asserting all possible individuals inside these classes, the above foundation, will represent the environment as well as the earlier models. Considering that the new model is much simpler than the earlier model and believing that good solutions are usually the simple ones, the decision was made.

As a result of this decision, as will be shown later, the set of rules imposed upon ontological classes became shorter.

Future Work

The material presented in this paper shows a very stable OWL model which allows reasoning with SWRL and in turn creates the semantics of a particular learning space in which students achieve their learning goals, but their differences in learning have been addressed, at the same time. The OWL model is rich enough to accommodate a verity of situations in learning. We would like to find out if the OWL model presented here is reusable across the spectrum of situations in which differences in learning play a major role in everyday lives, and not solely in education. Our Breakthrough section is clear that the OWL model is reusable and will be long lived, as long as we have a freedom of using object properties for strengthening the semantics, which describes connections between differences in learning and teaching practices. We are currently looking at instructions which can be generated to create cyber physical spaces in sport and music performance, for individuals who learn sport and music “differently”.

Finally, excerpts from this ontological model have already been used outside teaching and learning domains, which is an unexpected outcome of this research. Our SWRL enabled OWL ontology appears to be more powerful than anticipated (Juric, 2015), (Tarabi & Juric, 2018).

REFERENCE


DEVELOPING AN ONTOLOGICAL MODEL FOR ADDRESSING LIFESTYLE CHANGES FOR REVERSIBILITY OF DIABETES 2

Mehdi Tarabi, Radmila Juric, Steinar Larsen
Department of Science and Industry Systems
University of South East Norway

ABSTRACT

This paper illustrates design principles of a SWRL enabled OWL ontology used as the main data repository for creating a software application which can advise the population on the reversibility of diabetes 2. The OWL model is not a formal ontology, which contains extensive knowledge on suggested changes in the life style of population in general. It is rather a model which addresses individuals and contributes towards a personalized approach to reversing this chronic condition. The novelty of the model is twofold. Firstly it stores asserted and inferred knowledge, at the same time, in order to address the individual needs of a patient. Secondly, it is SWRL enabled, which secures reasoning upon OWL concepts at any time and generate inference whenever required. The proposed model creates tailor-made meals and exercise regimes for every individual who supplies his/her glucose level readings. The application built upon the proposed ontological model has been implemented with JEE technologies, using NetBeans, Servlets and OWL-API to access the ontology and perform reasoning.

INTRODUCTION

Semantic Web Technologies (SWT) were standardized more than a decade ago and since then they have been used for manipulating meaning on the Web through standards and proposals available on W3C websites. Their Semantic Web Layered Cake is a rich collection of standards and languages with emphasis on the unification of various inference mechanisms (Juric, 2016), (Menemencioglu and Orak, 2014), (Kumar and Dimakaran, 2015), (Svartz, 2013). One of the most popular and used languages is Ontology Web Language (OWL) (Patel-Schneider et al., 2004), which has origins in Description Logic (DL), follows its powerful knowledge representation (Drummond, 2010) and allows the execution of reasoning (Glimm et al., 2010), for the purpose of decision making. It uses DL philosophy, extended with reasoning rules (Motic et al., 2009), (Horrocks and Patel-Schneider, 2004) (Mei and Bontas, 2004). The power of OWL is unquestionable and its application in creating OWL ontologies has been remarkable. We have witnessed the creation of numerous OWL formal ontologies, particularly in biomedical informatics, medicine and healthcare, which systemize and explore knowledge stored in OWL concepts. There are ontologies for specifying human anatomy (Golbreich et al., 2013), classifying clinical research (Sim et al., 2014), annotating objects for medical image simulations (Gibaud et al., 2014) and sharing neuroimaging data (Gibaud et al., 2011), ontologies for documenting genetic susceptibility to diseases (Lyn and Sakamoto, 2009), pharmacological drug representation ((Zhu and Tao, 2014), classification and integration of drug discovery data (Lin et al., 2017), deriving clinical practice guidelines (Jafarpour et al., 2014) and creating chronic disease dietary consultations (Chin et al., 2015).

This list of applied OWL ontologies is just illustrative. There are more of them not mentioned here, which have been created as we speak. All of them store and manage excessive information, i.e. domain knowledge usually scattered around main OWL concepts: OWL, individuals and literals, properties, data ranges, axioms and annotations (https://www.w3.org/TR/owl2-quick-reference/). In order to address a trade-off between expressivity and decidability in OWL, reasoning rules have been discussed (Mei and Bontas, 2004) which rely on predicate logic and its inference mechanism. In 2005 Semantic Web Rule Language (SWRL) language has been proposed which combines OWL and RuleML (https://www.w3.org/Submission/SWRL/). It enables the strengthening of the semantic stored in OWL concepts, i.e. in OWL ontologies, creating inference and modelling of OWL ontologies across any domain. We talk today about SWRL enabled OWL ontologies (Shamoung and Juric, 2017), (Almami et al., 2016), (Almami et al., 2015) (Kataria and Juric, 2013), (Waqar et al., 2011), (Oppong et al., 2012), (Kataria and Juric, 2010) (Golbreich 2004), which enable inference from OWL concepts through SWRL rules.

In this paper we illustrate the modelling of a SWRL enabled OWL ontology in the domain of personalized medicine. We created a software application, based on SWRL/OWL for helping patients to address the reversibility of diabetes 2 through lifestyle changes (Tarabi and Juric, 2018). It is very important to note that the use of OWL and SWRL in this research is not focused on cumulating all knowledge within OWL and its potential growth, as we run any type of inference. In other words, we do not build formal ontologies. Our SWRL enabled OWL ontologies play a role of a software engineering mechanism, which enable reasoning upon the modelling concepts, in order to understand situations in which these software application reside (Kataria, 2012) (Suh et al., 2011), (Shojanoori, 2013), (Saaidi et al., 2011). We are able to interpret the semantic of such situations through OWL and SWRL and reason on ad-hoc basis (Shonajoori et al., 2012), (Juric, 2016). Obviously, this is a
slightly different computational paradigm compared to software applications, which use formal ontologies. However, we have been able, in numerous problem domains, to understand and manipulate semantics of almost any environment in which reasoning is needed. SWRL enabled OWL ontologies are here used outside semantic web, but the W3C initiative, on the deployment of vertical applications through the SWT, does encourage research similar to ours.

This paper is also a continuation of our previous work (Tarabi and Juric, 2018) in which we illustrate the role of SWRL enabled OWL ontologies in creating smart software applications which can personalize the reversibility of diabetes 2.

The paper is organized as follows.

In the Related Work we talk about a handful of papers, which use OWL ontologies in the management of diabetes 2. We single out a few of them which are closer to our work in terms of the defining OWL concepts in the same problem domain.

In the section which follows, we define our OWL modelling principles, which are based on the premises that
(a) we do not build formal ontologies;
(b) SWRL reasoning addresses a particular moment in which users (patients) interact with the software applications and require assistance for supporting their efforts to reverse diabetes 2;
(c) The computational model which hosts our SWRL enabled OWL ontology is reusable and SWRL does not contain values and literals,
(d) Interference is allowed to either move OWL individuals or define new properties.

The modeling section deals with OWL concepts, which in this case, range from OWL classes and their individuals to object properties. It is important to note that we show only one example of SWRL rule in this section for one reason. SWRL rules are parts of our computational model, which can run, i.e. be executed only if our OWL ontology is SWRL enabled. Therefore this is a core section of the paper and illustrates the way a SWRL enabled OWL ontology could be modelled.

In Conclusions we comment on the OWL model, its reusability within the same problem domain and across personalized healthcare delivery.

RELATED WORK

The initiative of W3C to establish the Semantic Web for the Healthcare and Life Sciences Interest Group (https://www.w3.org/2001/sw/hcls/), in order to help organizations in the adaptation of SWT, has been illustrated in the Briefings in Bioinformatics, published by the Oxford Academic Press (https://academic.oup.com/bib/issue/10/2) in 2009. This special issue debates on state of the art in Semantic Web for Healthcare and Life Sciences, and looks at a selection of papers, which focuses on creating ‘data webs’, knowledge discovery and semantic/context awareness. This initiative did not create considerable interest for a complete deployment of SWT in Healthcare, as it has been on the Web. Instead, we have witnessed interest in individual SWT languages and standards available within the SWT layerCake. Therefore, Related Work in this paper looks at the latter, which means that we are interested in the role of OWL ontologies, their models and the reasoning with SWRL, when managing and manipulating semantics in the healthcare problem domain.

We initially looked at the papers, which use OWL in the management of diabetes in general, and have found a variety of interesting publications. In (El-Sapaghi and Ali, 2016) we can read about OWL 2 Diabetes Diagnosis Ontology (DDO) which represents entities in the domain of diabetes and follows design principles recommended by the Open Biomedical Ontology Foundry. This is obviously a serious attempt to systemize knowledge related to diabetes and its diagnosis and place it within a formal ontology, which might have a positive impact on the way the knowledge is disseminated and shared.

A similar domain ontology and SWRL rules have been defined in (Chen et al., 2012) in order to extract the best possible anti-diabetic drug. The drug selection is performed though ontological matching between drug effectiveness and side effects, and patient’s symptoms. Their ontology is formal and contains knowledge from numerous existing sources and databases, but SWRL is used for performing reasoning as the only means of carrying out ontological matching. The latter part of their work is interesting: (a) a potential drug prescription attempts to be personalized and (b) the OWL ontology is probably SWRL enabled, to a certain extent. However, the ontological matching is solely based on semantic overlapping between the efficacy of drugs and patient’s symptoms. This leaves many other aspects, which may influence the choice of appropriate drug selection, outside scope, when treating diabetes 2.

In (Alharbi et al., 2015) we learn about another type of formal ontology which is based on a clinical decision support system for diabetes diagnostics. In their solution, a treatment plan is prescribed for patients based on their symptoms, risk factors and laboratory test. This is an example where SWRL is used upon OWL to strengthen the semantics of OWL model and store results of SWRL reasoning within the ontology. These results, i.e. a ‘treatment plan’ for patients are persistent and could be retrieved when necessary. These solutions are very close to ideas explored in knowledge management where OWL and SWRL rules build a knowledge-base. In (Mekruksavanich, 2016) we can see that their diabetes OWL ontology has been developed for a similar purpose. However, an expert system, implemented as a web application, interprets the meaning of patient’s symptoms through the ontology and may diagnose diabetes, very early, and in situations when patient symptoms are confusing.

The work of (Rahimi et al., 2014) focuses on algorithms based on an ontology, which can read patient electronic records and identify potential patients with type 2 diabetes. This is very interesting work, but their ontology has been influenced by (a) formal SNOMED-CT-AU ontology, which may have issues with its completeness and applicability in automating the process of diagnosing diabetes and (b) software applications, which would host algorithms in an
It is semantically rich and uses OWL classes, properties and decade ago, as a part of an e-Health integrated project in EU. which apply to most of them.

before the meal. It only advices patients on suitable meals blood sugar a particular patient may have on that day or just low. This application does not address the current level of (Patadia et al., 2011), but the level of personalization is rather similar to (Tarabi and Juric, 2018), (Juric and Kim 2017), (Sun et al., 2017), (Yumertz von Schwartzzenberg, 2015), (Sarathi et al., 2017). This has opened a completely new chapter in the deployment of OWL and potential of reasoning with SWRL in order to manage and disseminate knowledge and information on the reversibility of diabetes 2.

Classifying food and looking at its correlation with various diets has already been attempted (Snae and Bruckner, 2008). If OWL ontologies are being used, then they have a role of either annotating information available about food and nutrition in order to personalize diets in web retrievals as in (Albukhaitan and Helmy, 2013), or creating a co-relation between ingredients of food which could trigger serious health conditions in (Celik, 2015) (taking just these two examples as illustrations).

However, publications which explore the semantic correlation between the food and reversibility of diabetes 2 are in their infancy. At the time of writing we found only a few papers which come slightly closer to our area of research.

In (Arwan et al., 2013) an automated system which could replace a nutrition expert has been developed. Information relevant for dietary control is placed within the OWL ontology and patients are able to choose their own food based on information on calorie needs. Both food and calorie concepts are OWL ontologies and ontological matching is used to create the best possible combination of both. The ontologies have been enhanced with SWRL rules for the purpose of automatic classification of food. SPARQL is used for retrievals of the most suitable food for patients in which a choice can be made by an individual patent. In other words, the system will propose suitable menus for a patient who can then decide about his/her exact food intake. Their application has been deployed using Java libraries, OWL-API and JENA, similar to (Tarabi and Juric, 2018), (Juric and Kim 2017), (Patadia et al., 2011), but the level of personalization is rather low. This application does not address the current level of blood sugar a particular patient may have on that day or just before the meal. It only advises patients on suitable meals which apply to most of them.

In (Cantais et al., 2005) a formal OWL ontology which contains nutritional and healthcare concepts was developed a decade ago, as a part of an e-Health integrated project in EU. It is semantically rich and uses OWL classes, properties and individuals to classify and systemize information relevant for diabetes control. There is no evidence that the ontology has been used in any further research or software application development.

The ontology available in (Kumar and Latha, 2015) is used for food recommendation, but its role is to assist in the collaborative filtering of the food according to patient’s requirements.

The issue of physical activities for diabetic patients has also been investigated through either finding barriers to exercising as in (Alfaifi et al., 2017) or giving recommendations on physical activity during medical treatments as in (Pramono, 2015). In both cases, the proposed ontologies build knowledge on patients and physical activities and perform semantic searches upon both of them in order to create “recommendations” which in turn can be obtained through SPQRQL. In (Faiz et al., 2014) an integrated approach of diet and exercise regime has been proposed in which four different ontologies (Personal heath profile ontology, Disease, Food and Exercise domain ontologies) have been developed. They are integrated in order to give patients the most suitable recommendations related to their life style. However, the paper provides no explanation on how the knowledge from these ontologies was integrated and which semantic reasoning rules have been used. This solution seems comprehensive, but represents a heavy knowledge based implementation, which would require significant hardware and software resources. This might not be suitable for an individual patient, who should have access to the ontological retrievals on his/her home or hand held devices.

All the examples in this section have something in common: they do address the issue of lifestyle changes in the battle against the complexity of diabetes, as a disease, but none of them:

a) Give serious consideration to the fact that diabetes 2 is reversible and

b) Convince us that the commercialization of their solution is feasible.

Software applications built from any of the solutions elaborated in this paper would be difficult to implement in real life. We talk here about the environments where individual patients run the programs based on ontological reasoning. Considering the complexity and size of ontologies and their concepts, these will not be simple software applications and it is unlikely that they can run outside applications servers which store persistent repositories and allow complex web application to deliver results from ontological retrievals.

It is disappointing that none of these authors look at the role of both: food and exercise in the reversibility of this chronic disease.

**OWL ONTOLOGY MODELLING PRINCIPLES**

The modeling principles from the introductions are to be defined before the modelling of OWL concepts start.

We begin with two initial premises ((a) and (b) from the Introduction) that we do not build formal ontologies and that SWRL reasoning addresses a particular moment in which users (patients) interact with the software applications and
require assistance in reversing diabetes 2. These two are very important requirements, which go hand-in-hand. If we wish to address a moment at which decision making happens then OWL ontology can not become formal and can not accumulate knowledge. Each moment is characterized with a specific semantic, which may change. For example, users may have different glucose level readings from one moment to another, even if they were taken within a short period. The reasoning results should also change accordingly and it would not be advisable to make them persistent. These results would not be applicable to the same user if the glucose level, and some other factors, which impact the choice of food which should be taken, change. What is “true” at one moment might not be “true” at another and our OWL model should acknowledge this. Consequently, our basic ontological model should be simple and contain a minimum number of OWL classes and properties to accommodate the semantic of the moment, when the reasoning, in order to select the best meal at that moment, is to be performed.

Thinking about a simple OWL model, as advised in (Shojanoori, 2013) and (Juric, 2016) would not mean losing the semantic. It would mean that we collect the semantic of that moment which is sufficient for the reasoning upon it and allowing instant, new reasoning to happen if the semantic changes. No formal OWL ontology can make provisions for these requirements, because they are not meant to address the dynamic of software applications built upon them.

The third premise of our OWL modelling principle looks at the role an OWL ontology may play in software engineering. It is very important that we maintain reusability of the computational model, which hosts our SWRL enabled OWL ontology. This means that the OWL model is not “user” (patient) specific and computations with SWRL do not contain values and literals. This might NOT seem to be a requirement typical of OWL conceptualization and formulation of SWRL rules, but it is extremely important in software engineering. Obviously, moving a portion of OWL semantics into properties, and using them in SWRL rules is the best and safest way to generate reusability to a certain extent and avoid programming with literals and values.

Finally, the inference in the OWL model is encouraged and may range from (a) moving individuals from one OWL class to another to (b) inferring properties on an ad-hoc basis. However, this requirement is closely related to the “computational” part of our solution in which we perform reasoning with SWRL. According to the discussion in the first three paragraphs of this section, it is unlikely that the results of such inference would be made persistent. There should be a balance between the efficacy and simplicity of OWL models and making results of reasoning persistent as discussed in (Kataria and Juric, 2010), (Shojanoori et al., 2012), (Shojanoori, 2013), (Juric, 2016).

The Process of Generating OWL Model

There are no known processes of generating OWL ontologies which are not intended to be formal. For readers interested in debating how to create OWL models in modern and pervasive computational environments, the reading from (Shojanoori, 2013), (Shojanoori, 2015), give an insight into the level of assertions and inference in OWL, which would produce a stable computational model based on reasoning. However, this will be according to software engineering principles. Therefore, in this particular case we would have

(A) A set of basic OWL classes which would store individuals essential for defining the semantics of the domain of interest (in our case reversibility of diabetes 2 co-related to life style changes)

(B) A set of object properties defined upon OWL concepts which enhance and strengthen the semantics stored in the basic OWL classes

(C) A set of SWRL rules which reason upon the semantics stored in (A) and (B) and give the result, which serves the user: “which food should I take at this moment when my glucose level is XXX”

Obviously, we will manage to make our OWL ontology SWRL enabled if (A)-(C) is feasible and if the individuals from (A) and properties from (B) are captured at the moment when we wish to perform the reasoning. Avoiding literals and any type of values to appear in SWRL rule would bring forward a few observations:

- We would be better off if we use object instead of data properties and

- We should pay attention to the role object properties may play in the formulation of SWRL rules.

We have to bear in mind that the power of SWRL enabled OWL ontologies is in the definition of semantic overlapping between OWL concepts and consequently exploiting possible semantic mapping (Kataria and Juric, 2013), (Shojanoori et al., 2012). This will probably be one of the main arbitrators which would determine on the basic OWL classes and type of object properties we may have within the basic OWL model. Please note that the word “basic” in this context would mean the essential OWL model, which would secure successful reasoning, if the OWL concepts are populated with individuals. Extensions to the “basic” OWL models are welcome, but its retraction not desirable.

It is important to emphasize that the basic OWL model should allow for a clear definition of the reasoning process in which SWRL plays the most important role (Juric, 2016). Therefore we need to clearly specify which OWL individuals would be connected through object properties (as individuals of domain and range classes) and consequently which OWL classes would be involved in the reasoning process. Obviously, the reasoning process should follow the same logic regardless of inferring individuals or properties.

The choice of the reasoning process might have some impact on the overall performance of the computational model, which hosts the OWL model and SWRL rules. Therefore a wise software engineering practice should be followed. It is important to exercise a balanced use of constraints, which are properties in OWL modelling for many reasons. By not overusing constraints we might contribute towards a flexible and quick-response software application
which can run at any time and probably on any hand-held device.

To summarize, our OWL modelling process should include (A)-(C) above, but the selection of basic OWL classes, their properties and the definition of reasoning process, in which all concepts play an important role, must be carefully balanced. There is no quick or universal solution for modelling concepts for SWRL enabled OWL ontology.

**OWL MODELLING**

**The Scenario**

Our OWL model and reasoning with SWRL should address the following problem domain. We aim to build a software solution for personalized management of the reversibility of diabetes 2 and assist patients in making decisions on personalized changes in their life style and diet, and therefore address the chronic condition.

We focus here on the final product, which is a lightweight software application, deployable on smart devices. It should give an instant, i.e. ad-hoc advice to a particular patient on how to personalize the reversibility of diabetes 2 at the moment when it suits the patient. This can happen when either the patient is motivated to take an action or when his/her clinical data “signals” that the rising level of glucose in blood needs urgent attention. In other words, advice on diet and life style changes must be articulated in real time, when either the patient current clinical data (available on an ad-hoc basis) or patient requests, trigger it.

In this paper we look at blood glucose levels only, as one of the main triggers for reasoning within our OWL ontology. This is sufficient to prove that our concepts work. We can extend the semantics related to blood glucose levels with the list of various other triggers (found in other patients’ clinical data), which can alert patients towards life style changes in future work.

Furthermore, for reasons of simplicity we show here how we approach the problem of advising on food changes by looking at the way we create a “meal” based on the glucose level readings. Therefore, the personalization in reversing diabetes 2 happens at any moment when we want to create a “meal” according to the glucose level readings. The software application does not create a uniform “meal” for various glucose levels: the creation of a “meal” is triggered by a particular glucose level reading at a particular patient. This could be automatic, by the application (whenever glucose level readings change), or by the patient who may require a “meal” after checking his glucose level reading.

**OWL Classes**

Figure 1 shows our basic, proposed OWL model. There are 5 essential OWL classes which model the semantic of the problem domain. Each of these classes in Figure 1 are shown as single OWL classes. However, if we wish to extend their structure by adding more semantics to each base OWL class, we can have additional horizontal OWL hierarchies from each of them. We can also distribute various individuals of each of these OWL classes across their horizontal hierarchies. Furthermore, by calling a class Glucose_Level we signal that the semantic, related to clinical data, which can trigger the action of “preparing a meal”, is stored in this particular class.

The class called Food and Physical Activity are base classes which store information on available food and various physical activities which may be relevant for life style changes in general, when trying to address the reversibility of diabetes 2. There is a sharp difference between the individuals stored in these two classes and individuals in the Glucose_Level class, which is very important for SWRL enabled OWL ontologies. The latter class must have asserted individuals which are directly sent by either sensors (which measure glucose levels) or are taken from any other software application which calculates glucose levels for a patient. Therefore this is asserted data potentially generated by sensors. The former classes are actually cradles for persistent (asserted) values of various Food and Physical Activity individuals which are also not likely to be inferred. Therefore all these three classes represent asserted knowledge which does not have to be changed frequently (such as Food and Physical Activity) and knowledge which is constantly fed as assertions to the OWL ontology (Glucose_Level).

![Figure: The Basic OWL Model According to the Scenario](image)

The remaining two classes from Figure 1 have a completely different purpose. They are a cradle for storing inference for a particular moment when we run the application and perform the reasoning with SWRL. Therefore, the role of Meal class is to store an inference of individuals in OWL, after running SWRL rule(s), which would take only individuals of Food class, deemed to be sufficient and safe for creating a Meal, after reading a particular Glucose_Level. The same logic applies to the Exercise class. We move to this class all individuals from the Physical Activity class which are suitable for a particular Glucose_Level readings and consequently create an exercise program for a particular patient, at the moment when the Glucose_Level readings trigger the reasoning. Consequently, these two classes Meal and Exercise are here to store the result of reasoning and hold the inference created by SWRL rules. It is important to note that we create inference by moving individuals, which proved to be an effective method of interpreting the semantics of the environment in which decisions on the most appropriate meal or exercise have to be taken, as previously practiced (Katari...
and Juric, 2013), (Kataria and Juric 2010), (Shojanoori, et al., 2012), (Almami et al., 2016). However readers interested in the role of semantic overlapping when creating such inference, are advised to read the same sources of information.

The classes from Figure 1 are easy to populate with asserted individuals and to make them persistent. However, classes populated with inferred individuals are stored within OWL ontology for a relatively short period of time. At the moment, when Glucose Level readings change, the content of Meal and Exercise classes may not be valid any more. They may even contain dangerous information. Therefore these inferred individuals should be deleted before we run the reasoning again, with a new individual in the Glucose_Level class, which triggers it. In other words, we perform the deletion of inference if we experience changes of the individuals in the Glucose_Level class.

**OWL Constraints**

Table 1 shows a small excerpt from the list of all object properties we defined between the individuals of the classes Food (domain) and Glucose_Level (range) from Figure 1. For reasons of brevity, we show just a snapshot of potential individuals for Food. In the case of individuals of the Glucose_Level class we used abbreviations gl1 and gl3 to associate them with a range of Glucose_Level readings for many reasons. We wish to avoid literals and values in SWRL rules, we wish to leave space for any other trigger found in patient clinical data, which would require similar reasoning, and finally, glucose level readings can be asserted in ranges if we find out which range of Glucose_Level reading would require which food.

**Table 1: An illustration of Constraints in the OWL Model**

<table>
<thead>
<tr>
<th>Glucose_Level (domain)</th>
<th>Object Property</th>
<th>Food (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gl1</td>
<td>is_good_for_GL1</td>
<td>Poultry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whole plant food</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Beverage W/O sugar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pineapple</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Farmed fish</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fish</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Organic meat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Game</td>
</tr>
<tr>
<td>gl3</td>
<td>is_good_for_GL3</td>
<td>Cheese</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Organic meat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Purple potatoes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chocolate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Berries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Broccoli sprouts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Beans</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Green tea</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fish</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lentils</td>
</tr>
</tbody>
</table>

One of the most important outcomes of defining object properties shown in Table 1 is the illustration of their power in this reasoning. The semantics we store within Table 1 are more powerful than the semantics shown in Figure 1 (Juric, 2016), (Shojanoori, 2013), (Kataria, 2013), (Almami et al., 2015), (Saaidi te l., 2012). Not only must object properties must appear in SWRL rules, they can be easily asserted or even inferred whenever necessary. This means that if we obtain new knowledge about the food classification and its applicability in reversing diabetes 2, we can assert or infer new object properties in Table 1 and leave the OWL model intact. Object properties must be a subject of changes because they glue the semantics stored in OWL concepts and provide the best possible means of explaining how individuals of ontological classes are related to each other. The most powerful concepts of inferring object properties may add to the dynamics of our final software solution and enhance the automation we perform in such environments, solely based on reasoning with SWRL enabled OWL ontologies.

**The Reasoning Process**

The reasoning process performed with SWRL has already been outlined in the previous 2 sections. This means that we have already modelled OWL concepts and their constraints by having a particular reasoning process in mind.

The purpose of Figure 2, which is named “The Reasoning process”, is to show where inference happens and what the role of object properties is. Furthermore, the individuals in the form of “triangle” in the Food class are the individuals which will be inferred through the reasoning to create a Meal and therefore moved to the Meal class.

![Figure 2 The Reasoning Process for Creating a Meal](image-url)

Obviously without clearly defined object properties, marked in Figure 2 as a blue bidirectional arrow, we would not have enough semantics to decide which individuals from the Food class are going to be moved to the Meal class.

The SWRL rule which perform this inference is relatively simple and can run regardless of which individuals we assert in the ontology.
CONCLUSIONS

This paper illustrates an excerpt from the ongoing research on the role of SWRL enabled OWL ontologies in delivering personalized healthcare. The paper is strongly related to one of our publications (Tarabi and Juric, 2018). The ideas and practices of creating SWRL enabled OWL ontologies for software engineering purposes across healthcare and some other domains of interest, can be found in publications available either in our reference list or upon request. We encourage readers to look at the Java based application (Tarabi and Juric, 2018), which uses our SWRL enable OWL ontology as a computational core, in order to deliver personalized meals to people who are interested in addressing the reversibility of diabetes 2, at the moment when either clinical indicators trigger the change in their life style or people adhere to it.

All these goals will require no major revision of the OWL model from Figure 1 and might require no changes in the application developed in (Tarabi and Juric, 2018). On the contrary, this application proved that it is feasible to create it, and it is just a matter of time and resources when it could be commercialized and moved to hand held personal devices. We can not expect that people interested in their health would run this solution as a web application.

REFERENCES

Alfaifi, Y., Grasso, F., Tamna V. Towards an Ontology to Identify Barriers to Physical Activity for Type 2 Diabetes, Proceedings of the DH ’17 ACM conference on Digital health, pp 16-20 , London UK


SWT Road Map, available at W3C website https://www.w3.org/2001/10/03-sww-1/slide7-0.html


SOUND ANALYSIS AND MACHINE LEARNING IN NONINVASIVE CLASSIFICATION OF NEUROLOGICAL CONDITIONS

Khaled Elzaafarany¹, Gyanendra Kumar², Moustafa H. Aly³, Arie Nakhmani¹
¹Department of Electrical and Computer Engineering, University of Alabama at Birmingham, AL, USA
²Department of Neurology, Mayo Clinic, Phoenix, AZ, USA
³College of Engineering, Arab Academy for Science and Technology, Egypt

{khaledz@uab.edu, Kumar.Gyanendra@mayo.edu, mosaly@aast.edu, anry@uab.edu}

ABSTRACT

Machine learning and signal analysis are widely used to assist medical practice. Modern medicine is in constant search for effective noninvasive methods for diagnostics. Unfortunately, many of the developed automatic methods are prone to noise and disturbances and have high computational complexity. A transcranial Doppler (TCD) is a noninvasive and reliable device that can monitor the blood flow rate in the brain and can help neurologists to diagnose many brain problems like edema, trauma, hemorrhage, and aneurysm. The proposed algorithm is a blend of statistical and machine learning tools that are used in Big Data analysis. The algorithm’s goal is monitoring the TCD signals in the real-time for detection of cerebral vasospasms, which produces enormous amounts of data. We handled the data by carefully selected time and frequency domain features which allowed designing classifiers with the desired sensitivity and specificity. In addition, the proposed convergence of digital sound analysis and medical fields could prove to be useful in the future modeling of various brain disorders.

INTRODUCTION

Nowadays, sound processing methods are well developed and widely used in our cell phones, smart watches, and computers. Though modern machine learning based algorithms could automatically transcribe music and recognize speech, those algorithms were not explored enough about their applicability in medical diagnostics. The physicians still rely on their experience and senses to diagnose various diseases instead of using automated techniques. However, sound analysis and machine learning start to take their place in medicine, and this article shows a successful attempt of using sound analysis framework in neurology.

One of the medical devices that is used for diagnostics is the Doppler apparatus, which is noninvasive, inexpensive and safe. It was first used by Satomura in 1959 and its popularity is growing since then (Wright, Gough, 1997).

The physical Doppler effect allows measuring object velocity based on the sound arriving from the object. In medical applications, blood flow velocity in human body could be measured based on the Doppler effect, but it requires sound emitter and receiver to penetrate the blood vessels and receive a variation of sound frequency due to scattering and reflection. Measurement in the brain arteries is especially challenging, because it requires advanced spectral resolution methods. That kind of measurement is provided by transcranial Doppler device (TCD). A TCD can monitor the increase and decrease of blood flow velocity, or change in blood flow resistivity. Based on TCD measurement, presently neurologists detect stenosis, edema, trauma, hemorrhage, and aneurysm, but additional disorders are researched (Li et al., 2014). Angiography and magnetic resonance could be used as alternative techniques, but they are expensive and more complicated.

TCD signal analysis is active field of research. Serhatlioglu et al. used Fast Fourier Transform (FFT) for feature extraction and used a neural network as a classifier for TCD signals (Serhatlioglu, Guler, 2003). Ozturk et al. used chaotic invariant features of TCD signal as the input of neuro-fuzzy classifier (Ozturk et al., 2008). Guler et al. used FFT and adaptive autoregressive moving average (A-ARMA) methods for a spectral analysis of TCD signals; the A-ARMA shows better spectral resolution than FFT (Guler et al., 2002). Other studies (Uguz et al., 2008), (Uguz et al., 2010) used fuzzy discrete hidden Markov model (FDHMM) and the Roccchio-based hidden Markov model (HMM) for enhancing TCD signals classifier.

As seen in many studies, integral transforms are used widely, but the techniques cannot display time information that would be sufficient for a good classification. This means that many features of interest will be lost during the processing due to poor spectrum resolution especially in non-stationary
TCD signals. Therefore, we tried to use both time and frequency features for classification due to its optimal time-frequency resolution in all ranges of frequencies.

In the following section, the framework for signal processing and analysis is presented for machine learning applications in medical diagnostics. In the Results and Discussion section, we demonstrate the application of the proposed framework to detection of cerebral vasospasms in stroke patients. Finally, the Conclusion section summarizes the findings and proposes future research directions.

**SIGNAL ANALYSIS FRAMEWORK**

There is a direct analogy between speech and music analysis and other kind of sound analysis (Doppler). Doppler sound could be explored using developed for speech and music tools/features. There is an infinite number of ways to define features; we have selected those which work the best through a rigorous selection process. We propose to apply speech, sound, and music analysis approaches to the Doppler sound to detect and classify various brain disorders or events. The audio signal is acquired with TCD and analyzed for further classification. The framework consists of three stages:

1) Preprocessing
2) Feature Extraction
3) Classification

In the first stage, the signals are recorded, or processed directly from the audio output of the TCD. They are processed with standard filtering techniques to minimize noise. Fig. 1 shows one example of the cerebral vasospasm (bottom) signal and normal (top) TCD signal at the same time and voltage scale. It is visually obvious that those two signals are different, but it is not obvious how this difference could be quantified. In addition, in many cases normal and abnormal signals are very similar in their appearance. We have carefully selected from the features used in audio research those that might be useful in TCD analysis and beyond. In the third stage, the trained classifier arrives to the decision about the problem. The framework is providing supervised training of the classifier offline. Concerning the classifier, we used the decision tree classifier. The architecture of the proposed signal processing and classification algorithm is shown in Fig. 2.

A. TCD Signal Collection

In this study, 160 wave files of 3-15 seconds duration were recorded from TCD. Medical experts classified them as normal or cerebral vasospasm.

B. TCD-Signal Classification

We apply two essential steps: feature extraction and classifier for a condition classification.

**B.1 Feature Extraction**

This step is used in machine learning to compress the big dataset size into a small size and at the same time to increase the classifier efficiency. We used features from time and frequency domains. The time domain features are zero-cross rate (ZCR), energy, and energy entropy. The frequency features that were used are Mel-frequency cepstral coefficients (MFCC), Chroma coefficients, spectral centroid, spread, entropy, spectral roll-off, spectral flux and harmonic ratio.

**Fig. 1 TCD recording from control (top) and cerebral vasospasm (bottom).**

**Fig. 2 The architecture of the proposed classification algorithm.**

**B.1.1 Zero-Crossing Rate and Harmonic Ratio**

ZCR is used as a feature in the audio analysis (Tzanetakis & Cook, 2002). It counts the number of times the signal changes its sign per second. Concerning the harmonic ratio, we can assume the TCD signals as quasiperiodic to calculate the fundamental period by applying autocorrelation
function (by shifting the signal and calculating the correlation of the original signal with the shifted one).

B.1.2 Energy and Energy Entropy

Energy and Energy Entropy are used in the audio content analysis (Zhang & Kuo, 2001), where the energy is calculated by dividing the signals into frames and each frame energy to samples. Energy-entropy can reflect the sudden variation of energy (Lartillot & Toivainen, 2007). It can be computed by dividing energy of each frame to energies of sub-frames.

B.1.3 Mel-Frequency Cepstral Coefficients and Chroma

MFCC is used in speech processing field (Theodoridis & Koutroubmas, 2001). It measures the signal cepstral, where we apply the first 13 coefficients for CV detection. On the other hand, moreover, it can be used in music (Dan-Ning Jiang et al., 2002). It calculated by transforming the DFT coefficients to bins.

B.1.4 Spectral Centroid, Spread, and Entropy

Spectral Centroid, Spread, and Entropy are used for music, speech discrimination and watermarking (Scheirer & Slaney, 1997), (Kirovski & Malvar, 2003). Centroid means spectrum mass center and the spread is the second moment of the spectrum. With regards to spectral entropy, it is calculated like energy entropy, but the frame spectrum is divided into sub-bands.

B.1.5 Spectral Roll-off and Flux

Spectral roll-off can be used for differentiating between unvoiced and voiced signal (Kim et al., 2005). It is calculated by power spectral distribution. Spectral flux measures the spectral change for two sequential frames.

B.2 Classifiers

First, we apply feature scaling and mean normalization as a preprocessing step before the classifier to standardize the features values (Frohlich et al., 2003). We tried various classifying methods like a decision tree, K-nearest neighbors, support vector machine, and logistic regression. We found that decision tree gives the highest detection rate.

RESULTS AND DISCUSSION

In this study, we used the machine learning for classifying TCD signals and the obtained results are compared to a manual expert classification. A Supervised learning approach was used to train the classifier. We can enhance the efficiency of the algorithm by increasing the number of training information and the number of extracted features. The goal of supervised machine learning is to extract the model that makes an estimation based on evidence in the known input data and known output. Then, this model can predict the future output data. It is complicated to extract the correct model which is based on trial and error technique, for example, when the model has too many parameters to train, the training data will generate a sensitive model that will model minor variation which can be noise. On the other, hand too simplistic model will have low classification accuracy. Therefore, the right algorithm is a tradeoff between the training data, the number of features, accuracy, model speed, and complexity. The following schematic in Fig. 3 helps to overcome some of the machine learning challenges.
If the sensitivity of the model is low, so we should take more training dataset in the beginning. This loop should be repeated until we reach high accuracy and sensitivity model for the detection.

In the example of cerebral vasospasm detection, we used 10 time-frequency features to extract a high sensitivity model with using a decision tree classifier. The decision tree contains branches and leaves, where the decision rule can be made by an if conditional statement. Figure. 4 shows an example of two levels out of 34 levels of our decision tree classifier.

After extracting high accuracy and high sensitivity model, we tested the model to classify the cerebral vasospasm using real-time proceeding from the recorded TCD signals. On the tested signals the sensitivity and specificity were over 78%, which is an improvement over state-of-the-art results (Kumar et al. 2016), (Kumar, Elzaafrani & Nakhmani, 2017).

CONCLUSION

We designed and applied machine learning framework based on audio and music features extraction to TCD signals for automatic diagnosis of cerebral vasospasms without manual assistance. Time and frequency features were extracted from the signal, and they were used as input to the classifier. Our diagnostic system detects cerebral vasospasms with high sensitivity and specificity, which can help in medical practice. In the future research, we could use the wavelets for expanding the feature set to enhance accuracy. Moreover, one could try to use the classifier decision output and design embedded real-time system for alarming or for treatment purposes.

ACKNOWLEDGEMENT

K. Elzaafranay thanks, Arab Academy of Science and Technology for supporting his research.

REFERENCES


A MODEL OF WIKIPEDIA KNOWLEDGE COMMUNITIES: LEARNING FROM USER BEHAVIOR

Chester Parrott, Doris Carver
Division of Electrical Engineering and Computer Science
Louisiana State University
Baton Rouge, Louisiana
{ cparro5@lsu.edu, dcarver@lsu.edu }

ABSTRACT

The Internet is now the primary mechanism through which human knowledge is produced, stored, and consumed. Wikipedia is a rich source of information for researchers who find value from not only the content, but also the structure and organization of the articles in this open-source repository. Previous work concerning data from Wikipedia.com showed how the data can be used to assist with the construction of ontologies, human language models (through natural language processing), and improvements to the process of requirements engineering (Mahmoud & Carver, 2015). We describe research that uses aggregated user behavioral clickstream data from Wikipedia to derive meaningful semantic clusters. We apply community detection algorithms and modified graph algorithms to extract semantic clusters from graphs of user behavior. We derive a set of connected components which represent related articles within a shared subject domain. These resulting knowledge communities have application in numerous domains where understanding user behavior would be relevant, including software engineering, healthcare, and security domains.

INTRODUCTION

The Internet has become the primary mechanism through which human knowledge is produced, stored, and consumed. We direct our attention to means through which users of this ‘system of systems’ extract knowledge. One particular source, the website ‘Wikipedia.com’, has become a rich source of information for users and researchers. As an open-source knowledge repository, Wikipedia hosts millions of articles on a broad range of topics. Researchers find value from not only the content but also the structure and organization of these articles. Since this site has been available to the public, the number of articles has increased from 0GB in 2001 to 12GB in 2014, cementing Wikipedia as one of the major repositories of human knowledge (see Fig. 1).

Previous work concerning data from Wikipedia.com has shown how the data can be used to construct ontologies, human language models (through natural language processing), and improvements to the process of requirements engineering (Mahmoud & Carver, 2015). The data used in these works have come from the content of the Wikipedia website itself: ontologies of domain knowledge based on how content is organized on the site (Suchanek, Kasneci, & Weikum, 2008), natural language models based on the text within the articles (Collobert & Weston, 2008), and extraction of verbiage for software requirements engineering based on the text within the articles (Mahmoud & Carver, 2015). The form, structure, content, and links present within a Wikipedia article reflect the authors’ understanding of the semantic relations between the authored content and other content on the site. In contrast, the aggregated clickstream data reveal how users interact with the links which are present within different articles on the site. We posit that research efforts using content from Wikipedia as source material can benefit from analysis of the aggregated clickstream data.

In particular, as researchers use information about how the content of the site is structured to form ontologies of domain knowledge, understanding how users interact with the links which are present can provide valuable perspective into how the content is related. This assertion applies to other knowledge stores as well (such as corporate wiki sites (Syed, Finin, & Joshi, 2008). Understanding how consumers of a site interact with its content is to understand the content from the eyes of those consumers. Collectively, the activity...
patterns of Wikipedia users can provide insight into the ‘relatedness’ of articles on the site. We loosely define relatedness as the degree to which two or more articles share similar content, subject matter, and domain knowledge. Relatedness is the heart of semantic structure in a knowledge community. The core of a topic listed on Wikipedia is usually not stored in just one article. Meaning must be extracted from several articles (which might be linked together by contributors to that article). When links between several articles are regularly traversed by readers, logically, traversed links are more germane to the shared topicality of those articles than are those links which might be present but minimally traversed (or not traversed at all). Links listed in an article echo how the author(s) intended the content to be consumed; however, user behavior shows how articles are actually related in a knowledge community.

Existing methods for analyzing these data rely on the content. We extend prior methods with new analysis that includes user behavior. Mahmoud and Carver state the necessity of Wikipedia-based knowledge extraction efforts to be able to focus on individual domains rather than the entirety of the site’s contents (Mahmoud & Carver, 2015).

We describe the clickstream data (a small portion of which is illustrated in Fig. 2) to be a weighted directed graph, the Wikipedia Behavioral Click-stream Graph, where the distinct articles are represented by vertices, the links between articles are represented by edges, and the amount of web traffic between two articles are represented by the edge weights. Based on the behavioral click-stream data however, one immediate observation is that not all links in each article are actually used or represented as edges. Thus, we can truncate the number of links/edges in the Wikipedia Link Graph (a small portion of which is illustrated in Fig. 3) to be only those links which have been traversed by the users (as observed in Fig. 2).

Classic link-based approaches to building ontologies from Wikipedia use the links in each article (which might be on the order of several hundred, depending on the article). We propose that when constructing a Wikipedia-based domain ontology, the linked articles to which users navigate are more relevant to the domain than those linked articles which are ignored. In addition, articles which are in strongly connected components of the Wikipedia behavioral clickstream graph are topically more representative of the encompassing domain than articles which are only weakly connected. Strongly connected components require components which are constructed of outbound directed edges only, where weakly connected components can be constructed using inbound edges as well. In effect, a weakly connected component can be constructed which contains a set of articles not satisfying the classic connected component algorithm (Tarjan, 1972) discussed in this paper.

We show a modified version of Tarjan’s STRONNGCONNECT algorithm as a means of detecting clusters of semantically related articles (knowledge communities) from aggregated Wikipedia Clickstream data. We demonstrate the usefulness of Depth Constraint in reducing the size of the output clusters.

The paper is organized as follows. The following section contains related work. Next is the Wikipedia Clickstream Graph description. This is followed by the methodology to extract semantic clusters. Subsequently presented are the results and discussion of the results, respectively. Finally, last section contains the summary.

RELATED WORK

We show relevant methods and ideas from process mining, user behavior analysis and modeling, web application modeling, and conceptual models of knowledge (ontology). Additionally, we show previous work using Wikipedia data (which is relevant to our methods and results).

Modeling Wikipedia

The term ‘ontology’ found its mathematical underpinnings with the work of Wille, who established foundations for what he termed ‘Concept Lattices’ (Wille, 1982), (Wille, 1983), (Wille, 1992). This series of works focused on providing an applications-centric bridge to the field of lattice theory with the key contribution being a formal definition of how concepts can be structured. The commonly used description of information-specific ontology came when Gruber listed (Gruber, 1995) a formal process for the design and subsequent care of knowledge representation in the form of ontology. Gruber described ontology as a specialization of a conceptualization: in other words, a formal means of representing how knowledge and concepts are interrelated. With the creation of Wikipedia as an online knowledge store (ca. 2001), the global user community of the site has generated large amounts of domain-specific content. While Voss (Voss, 2006) states that Wikipedia articles do not have a strict semantic structure as do computer scientific ontologies (as described by Gruber in (Gruber, 1995)), such structure can be inferred from the similarities between articles and categories. Since Wikipedia’s inception, efforts have focused on extracting such formal concept specifications from the Wikipedia content.
Punuru and Chen show methods for extracting concepts from domain-specific documentation (Punuru & Chen, 2006), and Schönhofen details how cross-referencing Wikipedia titles and categories with document words can be used to classify the domain of documents (Schönhofen, 2009). Similar efforts using natural language processing (Janik & Kochut, 2008), (Wu, Yang, Yan, 2013, (Wu & Weld, 2008) have focused on extracting ontological structure from the site by the text inside the articles; however, other approaches such as Nakayama, Har, and Nishio (Nakayama, Har, & Nishio, 2007) and Syed, Finin, and Joshi focus on the categories associated with each article, the resultant hierarchy of concepts on the site, and graphs of both the links and concepts extracted from the articles (Syed, Finin, & Joshi, 2008). One user study related to Wikipedia consumption by Yu, Thom, and Tam (Yu, Thom, &Tam, 2007) describes how users of the site tended to have exploratory patterns with consistent ‘backtracking’ (navigation from the current page to the previous one) when the navigational information of the site was not clear enough to complete various research-related tasks (assigned as part of the study). The results indicate that backtracking represents the single largest percentage of user behavior on the Wikipedia site: users locate a ‘gateway’ page (a topical anchor point for basing further searches) and iterate through the page links until the next ‘gateway’ can be located (Yu, Thom, & Tam, 2007).

Modeling User Behavior

With the release of Wikipedia clickstream data, how can deeper understanding of Wikipedia users’ behavior be used to improve upon previous Wikipedia-based ontological efforts? Rich proposes one of the original papers on using a data-driven approach to modeling multiple users for clustering (Rich, 1979). This knowledge-based approach results in a static model akin to a decision tree for user representation. Middleton, Shadbolt, and De Roure in (Middleton, Shadbolt, & De Roure) show a knowledge-based ontological method for profiling users in recommender systems. Anderson’s novel approach using the Markov representation of sequential actions (as discussed in this research) is used for classification of malware (Anderson, Quist, Neil, Storlie, & Lane, 2011). Instruction combinations are used as features for clustering malware families.

Static means of using such models of processes and usage have been used in software engineering practices such as Cleanroom Software Engineering (Whittaker & Poore, 1992). Schur, Roth, and Zeller show how their tool ‘ProCrawl’ can extract behavior models from web applications (Schur, Roth, & Zeller, 2015), and Cotroneo et al. show that their ‘RELAI’ framework is able to use a probabilistic model based on operational profiles to improve software reliability (Cotroneo, Pietrantuono, & Russo, 2015). Probabilistic methods proposed by Menasce in (Menasce, 2002), Barros et al. in (Barros et al., 2007), Kant, Tewari, and Iyer in (Kant, Tewari, & Iyer, 2002) use a Markov-based model for generating behavioral traffic. However, in these approaches, each user’s behavior is represented by a unique model: an approach which does not scale when taking into account the volume of traffic for Wikipedia.

Community Detection & Graph Clustering

In order to reduce the number of individual models being used (and thus unnecessary performance overhead), many methods have been shown to cluster user behavior. Notable works include Barth in (Barth, 2010) who used social network analysis to identify groups of users in wikis (such as Wikipedia); Xiong, Wang, Jiang, and Huang in (Xiong, Wang, Jiang, & Huang, 2011) who showed how to use Markov Models to cluster sequences of categorical data (such as user behavioral trace data); Melnykov in (Melnykov & others) who showed a method for clustering clickstream data; and Benson, Gleich, and Leskovec (Benson, Gleich, and Leskovec, 2015) describe a means of segmenting network structures (such as our clickstream behavioral graph) using a tensor-based representation of the data and spectral clustering. Lee, Ellis, and Loui in (Lee, Ellis, & Loui, 2010) show a Markovian based model for clustering which is capable of detecting semantic concepts in audio data. Ahlers, Dirk and Mehrpoor shows methods of semantically searching shared information in (Ahlers, Dirk, & Mehrpoor, 2015).

Community detection algorithms are another means of detecting subgroups or clusters inside of graphs. Yang et al. discuss different community detection algorithms and their performance in (Yang, Aagesheimer, & Tessone, 2016). These algorithms have been implemented and tested in the R programming platform within the ‘igraph’ package. We summarize the runtime performance of these algorithms in Table 1. As nomenclature for measuring algorithmic complexity, let E as the number of edges in the subject graph, and N to be the number of nodes, or vertices. The ‘Edge betweenness’ algorithm focuses on using Freeman’s betweenness centrality as a means of filtering out edges which are highly shared between different communities in order to extract those highly shared edges from the larger graph. The algorithm runs in O(E^2 N) (Yang, Aagesheimer, & Tessone, 2016). ‘Fastgreedy’ runs in O(N log^2 N) and starts...
cluster or community then proceeds to pairwise match communities until a ‘modularity’ metric is no longer maximal (Yang, Agesheimer, & Tessone, 2016). ‘Infomap’ runs in $O(E)$ and uses decoders to reconstruct a view of the graph based on information collected from random walks on the network graph (Yang, Agesheimer, & Tessone, 2016). ‘Label propagation’ also runs in $O(E)$ and relies upon a majority heuristic by assigning tokens to each node in the graph and iterating until each node has the same token as the majority of its neighbors (Yang, Agesheimer, & Tessone, 2016). ‘Leading eigenvector’ runs in $O(N^2)$ on sparse graphs (where the number of edges is much lower than the number of nodes), and $O(N(N + E))$ otherwise; it relies upon using eigenvalues and eigenvectors of the modularity matrix to maximize modularity iteratively until there is less than an $\epsilon$ level of increase in modularity (Yang, Agesheimer, & Tessone, 2016). The ‘Multilevel’ algorithm is similar to ‘Fastgreedy’, but runs in $O(N \log N)$ (Yang, Agesheimer, & Tessone, 2016). ‘Spinglass’ runs in $O(N^{3.2})$ and can use simulated annealing to find communities (Yang, Agesheimer, & Tessone, 2016). ‘Walktrap’ runs in $O(N^2 \log(N))$ for sparse graphs and $O(EN^2)$ otherwise; it also relies upon random walks of the network and works on edge degree distance between nodes to hierarchically merge communities pairwise (Yang, Agesheimer, & Tessone, 2016).

Another effective algorithm to segment a graph is the STRONGCONNECT algorithm described in (Tarjan, 1972); this algorithm works to extract strongly connected components from a directed graph, and weakly connected components from an undirected graph (all of which are further discussed in (Diestel, 1997). STRONGCONNECT runs in $O(N + E)$ and is deterministic in nature. Building on the approaches used in these algorithms, we show how the Wikipedia clickstream data are structured and how Markov-based user behavior modeling and graph clustering can be used to extract semantic clusters from the Wikipedia.com website.

---

### Table 1 Community Detection Algorithms

<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge betweenness</td>
<td>$O(E^2N)$</td>
</tr>
<tr>
<td>Fastgreedy</td>
<td>$O(N \log^2 N)$</td>
</tr>
<tr>
<td>Infomap</td>
<td>$O(E)$</td>
</tr>
<tr>
<td>Label propagation</td>
<td>$O(E)$</td>
</tr>
<tr>
<td>Leading eigenvector (sparse)</td>
<td>$O(N^2)$</td>
</tr>
<tr>
<td>Leading eigenvector (full)</td>
<td>$O(N(N + E))$</td>
</tr>
<tr>
<td>Multilevel</td>
<td>$O(N \log N)$</td>
</tr>
<tr>
<td>Spinglass</td>
<td>$O(N^{3.2})$</td>
</tr>
<tr>
<td>Walktrap (sparse)</td>
<td>$O(N^2 \log(N))$</td>
</tr>
<tr>
<td>Walktrap</td>
<td>$O(EN^2)$</td>
</tr>
<tr>
<td>STRONGCONNECT</td>
<td>$O(N + E)$</td>
</tr>
</tbody>
</table>

---

### Table 2 Example Wikipedia Clickstream Data

<table>
<thead>
<tr>
<th>Previous</th>
<th>Current</th>
<th>Type</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blacksmith</td>
<td>Hephaestus</td>
<td>link</td>
<td>1000</td>
</tr>
<tr>
<td>Hephaestus</td>
<td>Goldsmith</td>
<td>link</td>
<td>100</td>
</tr>
<tr>
<td>Goldsmith</td>
<td>Pewter</td>
<td>link</td>
<td>1000</td>
</tr>
<tr>
<td>Pewter</td>
<td>Tinsmith</td>
<td>link</td>
<td>1000</td>
</tr>
<tr>
<td>Tinsmith</td>
<td>Blacksmith</td>
<td>link</td>
<td>100</td>
</tr>
<tr>
<td>Blacksmith</td>
<td>Bladesmith</td>
<td>link</td>
<td>100</td>
</tr>
<tr>
<td>Bladesmith</td>
<td>Metallurgy</td>
<td>link</td>
<td>30000</td>
</tr>
<tr>
<td>Metallurgy</td>
<td>Tinsmith</td>
<td>link</td>
<td>100</td>
</tr>
<tr>
<td>Metallurgy</td>
<td>Metallurgy</td>
<td>link</td>
<td>100</td>
</tr>
</tbody>
</table>

---

WIKIPEDIA CLICKSTREAM GRAPH

The first clickstream dataset for Wikipedia was released by the Wikimedia Foundation for the month of January 2015 (Wikipedia.com). Sporadic monthly releases have since followed which represent the aggregated traffic of all articles for the observed month. Table 2 shows an example of how the data are arranged. These data contain a list of links as tuples containing information on the ‘Previous’ or ‘source’/’referrer’ article from which the browsing behavior is initiated, the ‘Current’ or ‘target’ article which is being navigated to, the volume of traffic associated with the browsing behavior (‘N’), and additional classification information on the link itself (‘Type’). Individual records of user behavior are obscured by a monthly aggregation of the traffic. The clickstream data provide more information about how the site is navigated other than internal link traffic (traffic from one article on Wikipedia to another); however, as this research focuses on internal traffic, the additional classification information is useful for filtering out external link traffic. The data provide one-hop frequency information. As shown in Fig. 1, each record in the data provides the previously observed page (e.g. Bladesmith) as well as the currently observed page (e.g. Metallurgy). By constructing a graph of the clickstream, additional analysis of the data provides multi-hop information for base level statistical analysis; moreover, such analysis provides the data to analyze the semantic relationships between the links on the site.

We define our clickstream graph $G = (V, E)$ where $V = \{v | v$ is a distinct article on Wikipedia$\}$ and $E = \{e = (v_1, v_2, w) | v_1, v_2 \in V, w \in \mathbb{N}, w > 0\}$ is the set of all edges between any two articles on Wikipedia (where users have navigated from $v_1$ to $v_2$, and where $w$ represents the directed volume of traffic flowing in the direction: $v_1 \rightarrow v_2$. The clickstream graph is similar to the data structure used in the PageRank algorithm employed by Google’s search engine framework (Page, Brin, Motwani, & Winograd, 1999). The clickstream graph as exemplified in Fig. 2 is a sub-graph of the link graph (as represented in Fig. 3; however, we see that

---

1 Traffic between Wikipedia and external sites is not immediately relevant to this research.
traffic in the clickstream graph represents search behavior on the site as well. A link to an article might not be present on the site (and thus in the link graph), yet it might be present in the clickstream graph. Conversely, links which are part of a source article might never be clicked by users of the site; thus, these links will not be present within the clickstream graph, but present within the link graph. A minor modification to the clickstream graph calculates total transition volume from each source article and represents all edge weights as probabilities of transitioning from the source article of the edge to the target article of the edge (by dividing edge weight by total source volume). The resultant data structure is a Markov Chain representing the internal traffic at Wikipedia as a memoryless stochastic process. However, this Markov Chain clickstream graph is large and represents the traffic for all users across all articles on the Wikipedia website. We now focus on how to extract semantic clusters of interrelated articles on specific domains from this clickstream graph.

METHODOLOGY TO EXTRA CT SEMANTIC CLUSTERS

Semantic clusters in the context of Wikipedia Clickstream Graph are knowledge communities of articles which are topically similar (and thus ‘semantically related’). The high level process we describe for extracting these communities is shown in Figure 4. The steps range from (I) collection and (II) analysis of user data, to (III) flattening and aggregation, and finally (IV) the extraction of semantic clusters. Many methods exist for extracting semantic clusters, communities, and connected components from graphs; however, as described in the Related Work, each algorithm comes with computational cost to consider. As a Markov Chain representation of the Wikipedia Clickstream Graph is simply a graph with additional data associated, we use efficient graph-based algorithms for extracting communities. Tarjan’s algorithm is a means of discovering connected components of nodes within a connected graph (Tarjan, 1972). A connected component is defined to be “weakly connected” if the algorithm is run on a directed graph and follows the edges disregarding the direction (thus as an undirected graph). Tarjan provides a Depth First Search (DFS)-based algorithm for discovering the components (shown in Fig. 5 and Fig. 6) and formally defines strongly connected components with his ‘Lemma 8’ (here quoted from (Tarjan, 1972)):

\[ \text{Let } G = (V, E) \text{ be a directed graph. We may define an equivalence relation on the set of vertices as follows: two vertices } v \text{ and } w \text{ are equivalent if there is a closed path } p: v \Rightarrow v \text{ which contains } w. \text{ Let the distinct equivalence classes under this relation be } V_i, 1 \leq i \leq n. \text{ Let } G_i = (V_i, E_i) \text{ where } E_i = \{(v, w) \in E | v, w \in V_i \}. \text{ Then each } G_i \text{ is strongly connected [...].} \]

Extracting strongly connected components from the clickstream graph is equivalent to finding sets of articles which are circularly linked together through behavioral traffic patterns. Thus, if link traffic in our clickstream graph (Fig. 1) shows that an article on ‘Blacksmith’ has traffic to an article on ‘Metallurgy’, the article on ‘Metallurgy’ has traffic to ‘Tinsmith’, and the article on ‘Tinsmith’ has traffic back to ‘Blacksmith’, we consider these articles (and any others in the path from the source article back to itself) as being in one strongly connected component. The weaker condition (weakly connected components) only requires an un-directed path between these articles. When the behavioral chain extends through the graph to connect topics which are unrelated, we want to constrain the depth of the search through these vertices in order to reduce the size of the resultant strongly connected components and increase the semantic relatedness of each component’s articles. An example (as illustrated in Fig. 7) is ‘Blacksmith’ \rightarrow ‘Hephaestus’ \rightarrow ‘Mythology’ \rightarrow ‘Zeus’ \rightarrow ... \rightarrow ‘Blacksmith’.

We use a Depth Constraint - the recursion through which the STRONGCONNECT algorithm performs its search is halted when the constraint has been met. Users of the Wikipedia site might not choose to browse hundreds of articles in a single reading or research session, it is necessary to prevent the creation of strongly connected components which might have a deep transitive relationship, but not as deep of a semantic connection between the content of the two articles. Using the original StrongConnect algorithm results in a set of articles represented by Fig. 7; the shaded nodes are the set of articles most related by their content to ‘Blacksmith’. Where the transitive relationship chain might be thousands of links deep and create a strongly connected component, the resultant connected component will also have thousands of articles (such as ‘Zeus’ and ‘Apollo’) which are not directly relevant to the content of the original article (‘Blacksmith’). Our goal is to reduce the number of articles which have this transitive relationship. We increase the depth of the constraint by powers of 10 to simulate the level of depth the entire knowledge community might browse in extending the content of an article.

\[ 2 \text{ It is of note that while the exact definition of } G, V, \text{ and } E \text{ are different here from how they are defined in Wikipedia Clickstream Graph section, our definition is an extension of the one being used in this Lemma.} \]
Constraining the maximum recursion depth to ten links deep reduces the inclusion of tertiary transitive articles in our semantic clusters which are not topically related to the domain being described by the other articles in the cluster.

We modify the STRONGCONNECT algorithm (as shown by the bold text in Fig. 8 and Fig. 9) by adding a recursion counter which increases with each level of recursion. If a max depth of recursion has been reached (meaning we have traversed out to a maximum distance from the current search node), the recursion is broken and the next child of the search node is processed. Extracting connected components from the clickstream graph with the additional Depth Constraint reduces the number of extraneous nodes in each of these connected components. By reducing extraneous nodes, we are able to increase the semantic relatedness of each connected component. We illustrate this via Fig. 7 through the subset of shaded nodes versus the set of all nodes in the Figure. The more semantically related nodes are shaded where all nodes in Fig. 7 are only transitively related. Again, semantic relatedness is analogous to cohesion in that the articles are cohesive in their meaning. The articles discuss content which

StronConnect:

begin
  index < - 0
  indices < - {}
  on_stack < - {}
  S < - Stack()
  G < - Graph()

  for node in G.nodes:
    if node not in indices:
      ConnectComponent(G.nodes[node])

end

Fig. 5 Tarjan’s StrongConnect Algorithm (Tarjan, 1972)

ConnectComponent (n):

begin
  v < - n.value
  indices[v] < - index
  lowlink[v] < - index
  index < - index + 1
  S.append(v)
  on_stack[v] < - True

  for e in n.edges:
    v < - e.left
    w < - e.right

    if w not in indices:
      ConnectComponent(G.nodes[w])
    elseif on_stack[e.right]:
      lowlink[v] < - min(lowlink[v], indices[w])
    else:
      lowlink[v] < - min(lowlink[v], lowlink[w])

  if lowlink[v] == indices[v]:
    cc < - Component()
    while len(S) > 0:
      w < - S.pop()
      on_stack[w] < - False
      cc.add_node(G.nodes[w])
      if w is v:
        break
    G.components.append(cc)

end
ConstrainedConnectComponent (n, depth):

\begin{algorithm}
\begin{algorithmic}
\State $v \leftarrow n$.value
\State $\text{indices}[v] \leftarrow \text{index}$
\State $\text{index} \leftarrow \text{index} + 1$
\State $S$.append($v$)
\State $\text{on_stack}[v] \leftarrow \text{True}$

\For {$e$ in $n$.edges:}
\State $v \leftarrow e$.left
\State $w \leftarrow e$.right
\If {$\text{depth} \geq \text{max_depth}$}
\State \textbf{break}
\EndIf
\If {$w$ not in $\text{indices}$:}
\State $\text{ConnectComponent}(G\.\text{nodes}[w], \text{depth} + 1)$
\State $\text{lowlink}[v] \leftarrow \min(\text{lowlink}[v], \text{lowlink}[w])$
\EndIf
\If {$\text{on_stack}[e$.right$]:$
\State $\text{lowlink}[v] \leftarrow \min(\text{lowlink}[v], \text{indices}[w])$
\EndIf
\EndFor

\If {$\text{lowlink}[v] = \text{indices}[v]$}:
\State $\text{cc} \leftarrow \text{Component}()$
\State \While {$\text{len}(S) > 0$:}
\State $w \leftarrow S$.pop()
\State $\text{on_stack}[w] \leftarrow \text{False}$
\State \text{cc}.add_node($G\.\text{nodes}[w]$)
\If {$w$ is $v$}:
\State \textbf{break}
\EndIf
\State $G\.\text{components}$.append($\text{cc}$)
\EndWhile
\EndIf
\end{algorithmic}
\caption{Constrained ConnectComponent SubRoutine (Depth Constraint Logic in Bold)}
\end{algorithm}

which is similar in nature and topicality; without depth constraint, the number of unrelated articles produced by the original STRONGCONNECT algorithm is untenable as a model for representing knowledge communities.

RESULTS

The behavioral clickstream data are structured such that the graphical approach (subject to the depth constraint) is able to reveal information about the articles on the Wikipedia site which might not be readily apparent from the links listed in each article. Where the Wikipedia Links Graph in Fig. 3 might show several articles which are connected by links, these articles might only be partially connected by content. Specifically, Fig. 10 Fig. 11, and Fig. 12 illustrate how the depth constraint filters noisy traversals between articles. In Fig. 10, the maximum recursion is set to 10,000 nodes. For each search node, the DFS tree is traversed in the STRONGCONNECT algorithm for at most 10,000 hops until a cycle back to the search node is reached (represented by the dashed lines in Figure 7). To traverse this tree without optimization is computationally expensive; moreover, it can be problematic as resultant connected components include many articles which are not cohesively relevant to the domain of the DFS tree’s search root. The upper limit of 10,000 is used to illustrate the scale of larger, unconstrained executions of Tarjan’s STRONGCONNECT Algorithm. By example of what STRONGCONNECT can produce without suitable Depth Constraint, Fig. 10, represents the results of the algorithm running with a Depth Constraint of 10,000. There are components which have an inordinate amount of traffic across all the articles inside each component. Similarly, these components have an inordinate number of articles in each component. Each circle in Figures 10, 11, and 12 represents a connected component extracted using Depth Constraint. The connected component contains a set of articles which we expect to share content within a shared domain. The size of the circles represent the number of articles (in logarithmic scale) where their position on the x axis shows the total amount of traffic for each component. While we might not care as much about the volume of traffic being larger for one component than another, it is problematic to have that many pages in only a few components.

Similarly, for weakly connected components with a maximum recursion depth of 10,000 nodes (as in Fig. 11), there are even more of these components which have an overly large amount of pages. Both inbound edges and outbound edges are computed for the weakly connected condition to be satisfied, thus, the weakly connected components algorithm uses even more computational resources than the strongly connected version of the algorithm. By reducing the depth of recursion as in Fig. 12 to 10 nodes (Depth Constraint: 10), a drastic size reduction in the scale of the outliers occurred. We see in each progressive Figure 10,11, and 12 how reducing the Depth Level simultaneously reduces the average size of the resultant connected components. With the classical Tarjan algorithm, the output consists of many tiny connected components and few extremely large connected components (see Fig. 10 and Fig. 11. An example of a tiny connected component is the set containing the shaded nodes in Fig. 7; in contrast, a large, semantically unrelated component would be the set of all nodes in Fig. 7 such that the dashed lines represent hundreds or thousands more nodes. Inspection of the large connected components revealed that the articles represented did not share a holistic semantic relationship.
By constraining the depth to which maximum recursion was able to execute to a smaller number of traversed links (see Fig. 12), we reduced the size of the outlier components by orders of magnitude (observed as smaller circles in Figures 10, 11, and 12) and produce components of a more consistent size. The large connected components represent long chains of links in articles which are semantically related to each other but not to articles further down the chain. The chained relations between Wikipedia articles require any DFS-based community detection algorithm to make similar adjustments as we have shown with Tarjan’s algorithm. Such adjustments will increase the cohesiveness of the articles in each knowledge community.

**DISCUSSION**

We have shown that using community detection algorithms, and adapting graph algorithms toward the goal of detecting knowledge communities can be used to extract semantically related articles from Wikipedia. By constraining the depth to which such community detection algorithms execute, we are able to reduce the number of extraneously related articles in the resultant knowledge communities. Modeling these communities as individual behavioral Markov chains will enable researchers to extend previous efforts at constructing ontologies, requirements engineering, knowledge mapping, and semantic mapping of the articles on Wikipedia. Using Wikipedia’s behavioral data requires orders of magnitude less space and memory as compared to the raw article text. If we include the amount of data related to users editing pages, the costs of space can rise by an additional four or five orders of magnitude. For researchers on a constrained budget, these behavioral data could provide insight previously not available at a space/computational cost reduction. In addition, using Depth Constraint to extract knowledge communities from the data simplifies the process of analyzing the semantic relatedness of the Wikipedia.com content.

These results show how clickstream data can be used to extract semantic clusters from Wikipedia data. Previous methods using only the link graph to construct Wikipedia-based domain ontologies result in groups of articles which include unrelated pages (as index pages and other unrelated links appear in every article on the site). Inclusion of unrelated pages is due to extraneous transitive links; the ‘backtracking’ observed in the study by (Yu, Thom, & Tam, 2007) could provide additional insight into (or partially explain) how user behavior with the Wikipedia site could cause such results. The depth constraints presented reduce the inclusion of such unrelated transitive links in semantic clusters. The connected component approach is used by Janik and Kochut (Janik & Kochut, 2008) for text categorization and classification of non-Wikipedia documents; however, our usage for the connected components algorithm is focused on extracting semantic clusters from Wikipedia. Where Janik and Kochut create a graph, extract connected components, and use Wikipedia to classify documents according to a similar graph constructed from Wikipedia data, they do not use clickstream data. Rather, Janik and Kochut focus on using the natural language from the documents they seek to classify and the natural language from the articles on Wikipedia. While they use similar methods, the output and focus of their work is different from our research, in that they seek to classify documents, not create semantic clusters.
By reducing the recursive depth of the Depth First Search steps in the STRONGCONNECT algorithm, the nodes in the resulting components are more tightly connected than those which can be extracted from the links graph. This ‘tight-connectedness’ property filters extraneous edges from components of large graphs such as social networks or computer networks. We have presented the Depth Constraint extension of the STRONGCONNECT algorithm; however, an additional way in which the data can be constrained (by looking at the weight associated with the edges of the graph) is the Pareto Constraint. While the edge weight (traffic volume) is not immediately relevant for the construction of the Depth-Constrained connected components within our clickstream graph, these weights could be used for pruning the size of the connected components by removing links which only have transitive relatedness to the overall component. We define this procedure to be a Pareto Constraint; in other words, if the inbound and outbound traffic volume associated with a vertex is below an observed or pre-defined Pareto frontier, we disclude that vertex from the connected component. If the vertex is discluded from the connected component, all of that vertex’s outbound neighbors must thus be discluded as well (unless they can be visited by a transitive neighbor of the originating node of the STRONGCONNECT algorithm’s current iteration). We plan future inquiry on Pareto Constraint to restrict the size, breadth, and depth of connected components in large network graphs.

SUMMARY

We have shown a modified version of Tarjan’s STRONGCONNECT algorithm which we use to extract semantic clusters from a monthly aggregation of Wikipedia.com user browsing patterns. We represent the aggregated browsing patterns as a graph. Previous community detection algorithms do not run in the same order of efficiency as Tarjan’s STRONGCONNECT algorithm, and Tarjan’s STRONGCONNECT algorithm is not typically used for extracting communities from graphs, or network structures. The changes we have made to Tarjan’s algorithm capitalize upon the compact structure of Wikipedia browsing data, the efficient structure of Tarjan’s Algorithm, and the relatedness of the sites to extract meaning from data which previously has not been explored. Our modified version of the algorithm reduces the depth to which the embedded search algorithm in STRONGCONNECT traverses the graph. The output of the algorithm is a set of connected components which represent related articles within a shared subject domain. This depth constrained graph-based approach to extracting knowledge communities uses aggregated browsing data to capitalize on user behavior to produce representative clusters of semantically related articles on Wikipedia.com.

ACKNOWLEDGMENT

This work was funded in part by the Donald W. Clayton Ph.D. Assistantship.

REFERENCES


http://doi.org/10.1109/DSN.2007.102

Barth, F. J. (2010). Using Social Network Analysis and Hierarchical Clustering to Identify Groups in Wikis (pp. 8–11). Presented at the 2010 Brazilian Symposium of Collaborative Systems II - Simposio Brasileiro de Sistemas Colaborativos (SBSC-II), IEEE.


THE FRAMEWORK FOR COMPARING BIG DATA TECHNOLOGIES IN PROCESSING LIVE AND USER GENERATED DATA

Joseph Fotso, Radmila Juric
Department of Science and Industry Systems
University of South East Norway
joseph.fotso@student.hbv.no rju@usn.no

ABSTRACT

This paper explores the possibilities of choosing, configuring and running Big Data technologies for the purpose of answering strategic questions companies may have. The outcome is an evaluation framework which takes into account important functionalities of most frequently used Big Data technologies and issues associated with their use in commercial environments. Four Big Data operational environments have been configured and analyzed, and applications run within them: Hadoop's echo-system within its Cloudera's distributor, IBM Bluemix platform, Watson Analytics upon collected data and Watson Analytics for Social Media. The result of the research shows drawbacks and advantages of each of them, but also shows the exact criteria which have been used in the evaluation of the environments.

INTRODUCTION

The problem of deciding how to process Big Data, which Big Data technology should be chosen, and in which circumstances is difficult to address. The complexities of using Big Data technologies have been very well known in industry and academia and it will take more time and research to address the issue (Gandomi and Haider, 2015), (Manu and Anandakumar, 2015), (Gudivada et al., 2015), (Kaisler et l., 2012), Rajputohit, 2013). The complexity of these technologies will not decrease, and the amount and type of data we generate on a daily basis is rapidly growing.

This research addresses one important problem companies may face, while trying to collect and process Big Data. Most companies are very much aware that they have to process Big Data, for various reasons, but there is no known strategy on how to choose a particular Big Data technology and how to use/apply it. Will the choice of technology fulfill their expectation? Will the technology deliver exactly what companies expect? Which investments are needed before companies start using Big Data technology(ies)?

In this paper, we illustrate our own methodology of developing an evaluation framework, which can measure the efficiency of Big Data technologies. We chose to evaluate four environments, which have been dominating the Big Data scene in the last 5 years: Hadoop ecosystem (Loshin and Reifer, 2016), (HADOOP), Bluemix (Bluemix), Watson Analytics (WA) and Watson Analytics for Social media (WA - SM). Their comparison has been made through the proposed framework. The framework criteria are generic and could be used for any other Big Data technology. It is divided into three parts: choosing and configuring technology components, collecting and filtering relevant data and performing computations with the data.

However the proposed framework:

- Requires to formulate business questions which are expected to be answered by processing Big Data (DeAngelis, 2015), (Littlefield, 2016), (Panneerselvam et al., 2016),
- Pays attention to the source of Big Data, which would answer the question,
- Analyses configuration issues which accompany any technology of similar complexity,
- Examines connections of the technologies to social media, which often creates live and user generated data important for running businesses and
- Focuses on running SQL like queries upon NoSQL repositories generated by Big Data technologies.

We are not aware if a similar framework exists, which could help companies to choose an appropriate technology for their Big Data processing. We do not claim that our framework is universal and that it covers all possible situations, but it selects criteria which are applicable in evaluating any Big Data environment. It groups the evaluation criteria in order to make comparison for the evaluation clearer. It also raises awareness of two important issues which have always been hidden behind the marketing of such technologies and which actually triggered this research (Panneerselvam and Juric, 2016), (Fotso and Juric, 2018) Juric et al., 2017).

The first issue is that companies should always have clearly defined business questions ready, before they choose a suitable technological solution.

The second issue is the need to strike a balance between the complexity of the technologies, if we decide to go for open source solutions, and user-friendly environments which are offered by commercial distributors of the same technology. We have to pay attention to the fact that the complexity of the technology may have a negative impact on its successful deployment. However, commercial distributors, such as Cloudera for Hadoop, may reduce the problem, but would also reduce our power when we try to decide how and when to process Big Data (Panneerselvam and Juric, 2016).

We expect that the framework could point towards the similarities and differences between the technologies and open the door for improvements in all these four environments.

The paper is organized as follows.
In the next section we describe the methodology which has been used for developing the evaluation framework. The methodology comprises three stages and 23 steps. In the section which follows we define the framework and apply it to the four chosen Big Data platforms. In conclusions we analyze the result of the evaluation and look at the potential improvements of the framework by including Big Data technology services, which are not covered in this study.

THE METHODOLOGY

The methodology in this research has been specially formulated for achieving the research goal: How do companies choose which Big Data technologies should be deployed in order to answer their important business questions?

We focused on the applicability of Big Data technologies in healthcare and the publications which address chronic diseases and diabetes 2 in particular (Kumar et al., 2015), Cichosz et al., 2012), (Ramzan et al., 2016). Our business question, which was supposed to be answered by Big Data technologies, in order to achieve our research goal above is:

Do Big Data technologies have the power of processing live and socially generated data, for the purpose of educating the population that diabetes 2 is reversible through diet and lifestyle changes?

It is important to note that we had to define this methodology for specific reasons. If we wish to evaluate Big Data technologies, then we have to set the scene, decide about the problem domain which would require Big Data processing and finally formulate a business question which is supposed to be answered through Big Data processing. Without these three points we would not have been able to create and apply any evaluation framework. Consequently, the methodology has become very problem domain specific, which might not result in its high reusability. However, when creating three important stages within the methodology, and defining steps within these stages we actually had to make provision for semantic grouping of various criteria which would comprise the framework. Therefore not only should our methodology secure the generation of an evaluation framework, it should also guarantee that we would be able semantically to group various evaluation criteria in order to increase the reusability of the framework and make it applicable for the evaluation of some other Big Data platforms.

The steps of the methodology are itemized below and they are self-explanatory.

They are divided into 3 STAGES. In stage 1 we prepare the research and run the first set of experiments with the Hadoop ecosystem. In Stage 2 we partially repeated steps from Stage 1 in order to run the same experiment on 3 different platforms offered by IBM. In Stage 3 we create evaluation criteria and place them into the framework.

It is important to note that the definition of Stage1 has been influenced with our earlier experiments and their results when we investigated the efficiency of Hadoop ecosystem (Panneerselvam and Juric, 2016), (Fotsso and Juric, 2018)). In other words, steps in Stage 1 have already been tested and produced interesting results. Therefore, we are reusing them in Stage 2 when looking at the other three environments from the IBM corporation. Stage 3 summarizes the outcomes from the previous stages and creates a matrix which contains both: our evaluation criteria and the results of applying them in all four environments.

Steps of all three stages are listed below.

STAGE 1

1. Investigate and download tools available for managing Big Data Technologies
2. Learning Hadoop Echo-system, understanding its components and their applicability in processing of Big Data
3. Choosing Hadoop’s commercial distributor (Cloudera).
4. Downloading Cloudera and analyzing the level of configuration required when using Hadoop in Cloudera
5. Summarising possible advantages and drawbacks of using commercial distributor, such as Cloudera
6. Testing Hadoop’s configuration in Cloudera by choosing a Problem Domain (Processing live and user generated data, such as tweets, related to the Problem of the management of information on reversibility of Diabetes 2).
7. Defining the purpose and the role of Big Data Processing in terms of deciding WHICH business question will Hadoop answer (see the business question above). It is also important to define upon which data we need to run Hadoop’s components in order to answer the same question.
8. Running our Hadoop’s application with FLUME
9. Define SQL like queries which would answer the business question from Step 6.
10. Summarizing problems and results of running SQL like queries upon NoSQL repositories from Hadoop’s HDFS.

STAGE 2

11. Investigate and choose other Big Data platforms: investigate services offered by IBM
12. Join the IBM academic initiative, secure free access to their services and analyze all services they offer.
13. Choosing services, which would process live and user generated data, as in Step 5, but this time in IBM.
14. Learn Bluemix and Watson Analytics
15. Steps 13-15 are the same as Steps 7-9, but this time applied to BLUEMIX
16. Steps 16-18 are the same as Steps 7-9, but this time applied to Watson Analytics in IBM
17. Steps 19-20 are the same as Steps 7-9, but this time applied to Watson Analytics for Social Media.

STAGE 3

21. Collate the evaluation criteria derived from previous steps and group them in order to use them across all four environments in our problem domain (answering business question from Step 6).
22. Create a framework for measuring the efficiency of the chosen Big Data technologies in our problem domain by placing evaluation criteria within the framework.
Step 23 Create a matrix of the results of evaluation to see explicitly advantages and drawbacks of all four Big Data technologies applied in the chosen problem domain.

THE EVALUATION FRAMEWORK

<table>
<thead>
<tr>
<th>Table 1 Evaluation Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HADOOP</strong></td>
</tr>
<tr>
<td><strong>Question formulated by business company.</strong></td>
</tr>
<tr>
<td>Choose from: Watson Analytics IBM Watson Analytics for Social Media Analytics Exchange</td>
</tr>
<tr>
<td><strong>Find source of data</strong></td>
</tr>
<tr>
<td>Choose from: Twitter Forums Reviews Facebook pages (Beta) Videos Blogs News</td>
</tr>
<tr>
<td><strong>Choose Comm. Distributor</strong></td>
</tr>
<tr>
<td>Create Application Devop Env.</td>
</tr>
<tr>
<td><strong>Choose Hadoop comp:</strong></td>
</tr>
<tr>
<td><strong>Flume</strong></td>
</tr>
<tr>
<td><strong>Hue</strong></td>
</tr>
<tr>
<td><strong>HIVE</strong></td>
</tr>
<tr>
<td><strong>MapReduce</strong></td>
</tr>
<tr>
<td><strong>HDFS</strong></td>
</tr>
<tr>
<td><strong>Pig</strong></td>
</tr>
<tr>
<td><strong>Mahout</strong></td>
</tr>
</tbody>
</table>

Configure chosen component Flume | Launch: dashDB Insights for Twitter | Enter a topic | Enter keyword(s) |
| Connect Flume with Twitter data | Dashboard: Load Twitter data Connect to Application | Choose lang. of dataset | Choose lang. of dataset |
| Filter Twitter data -use the question) | Search for Twitter data: Enter keyword# | Enter keyword# | Enter keyword# |
| Live tweets are copied into HDFS | Live Tweets are copied in dashDB | Import data | Import data |
| Choose from HIVE LINUX | Choose from HIVE LINUX | Choose from HIVE LINUX | Choose from HIVE LINUX |
| Prepare HIVE External Schema (aligned to tweets) | Amend predefined external table | Discover data (showing schema) | Discover data (showing schema) |
| Itemize the question into smaller question | Itemize the question into smaller question | Itemize the question into smaller question | Itemize the question into smaller question |
| Connect to business intelligence /analytics-focused application | Matching Watson questions to smaller questions | Matching Watson questions to smaller questions | Matching Watson questions to smaller questions |
| Run SQL queries on one “small question” | Run SQL queries on live data in dashDB | Choose one of Watson proposed questions | Choose one of Watson proposed questions |
| Choose HIVE | View results of queries HIVE | View analysis | View analysis |
| View results of queries HIVE | View results of queries dashDB | View analysis | View analysis |
| Visualize results -CHART function | Visualize statistics results -CHARTS | Visualize results -CHARTS | Visualize results -CHARTS |
| Visualize results -CHARTS | Visualize results -CHARTS | Visualize results -CHARTS | Visualize results -CHARTS |

The proposed framework is in Table 1.
There are four columns in Table 1, which correspond to our four chosen Big Data environments: Hadoop, IBM
Bluemix, Watson Analytics for Social Media and Watson Analytics. The rows are our criteria in the evaluation. They are divided into three parts. Part I contains the first 10 rows in Table 1 and defines the Business Question, finds Source of Data relevant for answering the question and deals with the Configuration of the Big Data environment. Part II contains the next 4 rows in Table 1 and deals with collecting and filtering live and user generated data. Part III contains the last 8 rows of Table 1 and prepares and processes filtered and collected data, stored in a Big Data technology’s repositories, in order to answer the business question. In total we have 22 evaluation criteria.

Yellow shaded words within each criterion are related to the CHOICE we made when using each of these four Big Data technologies. This is an indication of what exactly we did within each of these four environments when we applied the evaluation criteria. The choices were listed by the technologies and were chosen according to the business question we formulated earlier. Therefore the choices we had to make have become an important aspect of our evaluation.

Part I Defining the Business Question, Finding Sources of Data and Configuring the Environment

Table 1 shows that in all four environments there were provisions for specifying a business question which is expected to be answered at the end of Big Data processing. Also, there was space for finding the source of data which could answer the question. In Hadoop, Bluemix and WA we could explicitly choose to work with TWITTER data, but not at the same stage of using the technology. In Hadoop we have to perform a special configuration (FLUME) before we connect to the social media. In Bluemix, the type of created application would determine that the we would be able to launch DashDB Insights for Twitter.

It is very interesting to notice that the left most column in Table 1 differ from the others: Hadoop is the only environment where configuration of components is compulsory, but Bluemix and WA SM would require either the creation of development environments or Social Media project. WA skips all of them and shows a smooth journey towards the selection of Twitter after choosing a NEW DATA option. It is obvious that WA focuses on data very early and creates a project around the chosen data. By contrast, Hadoop and Bluemix choose to deal with the selection of data sources (i.e. Twitter) much later in the process. Finally only Hadoop required a user managed configurations of some of its components, in spite of using its commercial distributors. In the IBM environment there was no need to look for such distributors because Bluemix, WA and WA SM are all proprietary software managed and sold by IBM (i.e. they are not open source).

Part II Collecting and Filtering Data

There were no significant differences in the way all these four environments deal with the collection and filtering of live and user generated data. In all of them we were able to enter keywords for filtering (#keywords). However in WA SM and WA we had an opportunity for experimenting with more precise filtering criteria and choice of time slots and languages used in the selection of tweets. It is interesting to note that we were aware of the existence of repositories in which collected and filtered data (Tweets) were stored in Hadoop and in Bluemix. This is not the case for WA and WA SM.

It is important to note that filtering consists of copying data in HDFS (for Hadoop), and in the DashBD for IBM Bluemix. In WA SM filtered data are stored within “the system” but we not have the flexibility of manipulating stored data, while in WA the corresponding step is just “import data”. We could not manage it as a repository of collected and filtered data.

Part III Preparing for and Processing Data with SQL

Preparation of collected data for their processing using SQL like queries upon NoSQL repositories, singled out WA SM as the only environment where we could not run our prepared queries. We could only view the result of the automated analysis. However in Hadoop and Bluemix we could fully exercise the power of running our own queries which would answer the business question a suitable way. In WA we were not able to run the same queries, but we were in a position to match them with WA prepared questions and their ready made answers.

Also in WA and WA SM we could not influence the SQL schema which we were supposed to generate before we ran SQL like queries. In Hadoop the schema was created manually, which was not the case for the other three environments. They created a needed schema automatically but allowed a different level of intervention.

In WA we were not given the platform of running SQL queries but could write/ask questions. WA will then automatically generate related questions and answers. The flexibility of visualizing the results of the questions, in the form of charts, is very impressive in WA.

It is important to note that the way we created SQL queries to answer the business question from step 6, is outside the scope of this paper. Readers interested on our process of creating SQL like queries upon NoSQL repositories should find more information in (11,12,13,19).

CONCLUSIONS

This paper illustrates the development of the framework for evaluating Big Data platforms and shoes the results of applying it in Hadoop, Bluemix, WA and WA SM.

There are a few interesting outcomes of this research.

We define the methodology which guaranteed the development of an evaluation framework, applied the proposed framework in 4 distinctive Big Data environments and show the result of the evaluation. The novelty of the research is in both: the way we created the framework and the results of the evaluation. They highlight the problems in Big Data processing which usually do not surface above marketing mechanisms of Big Data platforms and the hype
surrounding these technologies. This outcome would help any company to elaborate on their own strategy of acquiring Big Data platforms and may be of interest to their vendors and distributors.

How Do these Four Environments Compare

In these four different operational environments, which were evaluated through the framework in Table 1, we could see that they all have a different approach to the processing of Big Data. The steps within the processing may overlap but in general they differ in the way we configure the environment, collect and filter data and run SQL queries. There are quite sharp differences between the IBM Watson and Cloudera Apache Hadoop combination, and IBM Bluemix does not remedy that gap.

However, these big differences did not create an obstacle when defining the framework. On the contrary, we were able to compare the steps within each environment by finding out which one of them is not being used in which technology. This is an unexpected outcome of the research. “Empty” boxes within Table 1 clearly indicate what is NOT done within a particular environment.

This outcome has become particularly important when looking at open source software and comparing it to proprietary solutions. The process of using the technology in the former is more demanding. It may require the use of their commercial distributors, such as Cloudera for Hadoop, which is not free. However, if companies managed to deal with and invest into the combination of open source technology and its commercial distributors, then the processing of data collected in their repositories (such as HDFS in Hadoop) becomes very powerful. It is controlled by the user.

The proprietary solutions from IBM sheltered the user from most of the configuration problems and offered user friendly and powerful services within the environment, which could collect, filter and process Big Data. This does not come cheap, but it is a very elegant way of introducing Big Data technology within a company. The drawback is expected: users lose control of the way they wish to store and manipulate the data. This was a particular problem in WA and WA SM because we either were not allowed to perform SQL queries as we would prefer to or were not in a position to choose our own business question, which is expected to be answered by running SQL like queries.

Whichever outcome we would prefer, formulating the business question and answering it through Big Data processing is an important and delicate task. Its success may depend on the level of user intervention in the automated world of Big Data processing.

Table 1 also shows one interesting conclusion. There is space for a new generation of Big Data technologies, which would take the best possible practices from both “sides” Hadoop’s echo-system and IBM Big Data services. Currently, companies have to choose to go for either

- a solution where everything is prepared for them and is ready-to-use through “click and drop” and services tailored for various needs companies may have or
- a completely different solution where the environment chosen will give the flexibility of configuring, manipulating and processing of Big Data, but users will have to carry the burden of the complexity of such processing.

If we could allow for

- simpler configuration of Big Data technology than currently available in Hadoop, and exercising their mix and match to suit a company’s needs, and
- a higher level of user intervention than currently available in IBM and its Big Data services,

then we could probably have had a perfect technological solution for Big Data processing in the second decade of the 21st century.

Limitation of the Study and Future Work

The proposed framework in its current format could be of interest to companies which need to deploy Big Data technologies, and software vendors who are selling the technology on the market. In future, the framework could be tuned and show more details and criteria. However, in its current format the framework contains three basic aspects which may not change in future: configuration of the technology, collection and filtering of data and its processing.

However, the creation of the framework has been triggered by our own experience of using open source Big Data technologies, and our inability to manage it in the academic environment without substantial financial investment or assistance from the professional advice available in forums. This has had an impact on the methodology we used in the creation and the structure of the framework. More research has to be done in defining more generic evaluation criteria with substantial input from the practitioners and vendors from industry. The current framework is generated solely in an academic environment.

Furthermore, our numerous experiments, which were carried out over 5 months, were placed in the domain of Healthcare and Twitter social media. In future, we have to look at many other different situations in which companies acquire Big Data technologies but

- Big Data is generated from different sources and
- Big Data processing is not dependent on SQL like queries.

It would be interesting to see if our criteria available in Table 1 would hold.

The choice of using three IBM environments and Hadoop in this study was dictated by our previous interests in Hadoop echo-system and the availability of IBM services in academia. We have to expand our evaluation towards other software solutions which combine various Big data technology components, traditional algorithms from the artificial intelligence and software engineering solutions
which are currently interwoven is some the Big Data platforms.

REFERENCES
Bluemix available at https://www.ibm.com/cloud-computing/bluemix/
Fotso, J. Juric, R. The efficiency of Big Data Technology in Managing Information Important for public Health, under review for the International Journal on Health Systems
ABSTRACT

Convergence or collaboration across disciplinary boundaries is increasingly viewed as a vital element in forging solutions to complex technological, medical, and environmental challenges. Many articles outlining the potential benefits of convergence have been written, with some depicting this process as seminal to the future of medical research and new discoveries. However, even with the growing interest in this transdisciplinary collaboration, more research is required to translate the potential value derived from convergence into elements such as cost avoided, increased efficiency, or time saved. This paper, using a case study, provides an approach for quantifying the value of transdisciplinary problem-solving using convergence, engineering modeling, and communication theory. The technique outlined in this paper can be extrapolated for use in many domains that require transdisciplinary problem solving.

INTRODUCTION

Solutions for many of today’s engineering problems in healthcare, energy, and other critical engineering domains are not respecters of physical or virtual boundaries implemented to satisfy artificial control mechanisms (Fielder, Guldal, & Tanik, 2017). Integrated, convergent, and collaborative approaches are needed to bridge the disciplinary and communication chasms that impede effective and efficient solutions to these problems (Fielder, Tanik, Gattaz, & Sobrinho, 2016).

According to the National Academy of Sciences, convergence is an approach to problem solving that cuts across the disciplinary and communication boundaries by integrating tools and knowledge to solve societal challenges (National Research Council, 2014). Additionally, convergence is defined as the merging of distinct technologies, processing disciplines, or devices into a unified whole that creates a host of new pathways and opportunities (Massachusetts Institute of Technology, 2011). The ability of convergence-based approaches to integrate knowledge from disparate engineering disciplines and facilitate more timely problem resolution, provides an opportunity for generating solutions that are both reliable and “satisficed.” or “satisfactory and sufficient” (Simon H., 1956).

This research paper hypothesizes that convergence using engineering modeling and communication theory is a better approach to solving complex transdisciplinary problems compared to the traditional silo-based approach. Additionally, this paper hypothesizes that the value of convergence can be measured using engineering modeling techniques. As outlined in this paper, convergence principles will be vital in developing solutions to society’s most complex problems in industries as diverse as energy, healthcare, and other complex problem domains.

BACKGROUND

The traditional approach to solving complex problems generally involves leveraging silos of distinct disciplines, knowledge, technology, and tools. The silos are instantiated around a function or discipline such as mechanical engineering with its emphasis on moving or rotating parts, chemical engineering and its focus on the interaction of compounds, and software engineering with its focus on development of application code. Figure 1 illustrates the direct relationship between engineering silos and increased entropy or disorder in project communication. Disciplines have their own language, jargon, acronyms, tools, and process frameworks. Resultantly, these different languages, tools, and approaches to problem-solving create entropy in communication between the disciplines. As depicted, some amount of entropy is already resident in a single discipline environment. For instance, even within the electrical engineering discipline, there are different skillsets, different experience levels, different locations, as well as different tools that may be leveraged. These intra-discipline differences create entropy. However, this entropy increases markedly during inter-disciplinary communication. The large amount of entropy becomes an impediment to timely research and problem-solving since several iterations of communication are required to reach consensus on the basic foundational components that are necessary to unite the disparate disciplines into a cohesive research unit.
Transdisciplinary convergence enables common organization (goals, data, and jargon) of the research space. In turn, this common organization aids in entropy reduction. As illustrated in Fig. 2, convergence-based approaches facilitate a tighter integration amongst the disciplines by leveraging common jargon, common tools, processes, organizational alignment, and frameworks. The tighter integration can help reduce the amount of entropy on the research project.

The concept of disciplinarity represents the various ways in which disciplines (mechanical, electrical, software, material, project management, and biomedical) interact to solve problems. Table 1 provides a description of the various levels of interaction. These interactions range from disciplines working together without any formal integration (multidisciplinarity) all the way to full integration and dissolution of any boundaries that impede research (transdisciplinarity). Boundaries represent barriers to communication. The boundaries are often drawn along departmental, specialty, hierarchical, or geographical lines. In a non-transdisciplinary environment, boundaries may behave as physical constructs, limiting the free flow of information, increasing entropy, and impeding the research process. Transdisciplinarity is increasingly viewed as the bridge that overcomes disciplinary barriers when combined with convergence (Tanik & Atilla, 2008).

\[
N (N - 1) / 2, \text{ where } N \text{ is the number of resources working on the project}
\]

If there are \(25\) disciplinary resources on a research project, the number of possible communication channels is:

\[
\begin{align*}
\text{Channels} &= 25(25 - 1) / 2 \\
&= 25(24) / 2 \\
&= 600 / 2 \\
&= 300
\end{align*}
\]

Models (mathematical formulas, computer code, written language, simulations, virtual reality, or material representations such as statues) are invaluable elements of the problem solving process and key enablers of convergent approaches to engineering research (Fielder, Guldal, & Tanik, 2017). A model is a simpler realization or idealization of more complex real world entities (Ozaydin & Tanik, 2012). Modeling has the ability to depict both simplistic and complex engineering problems that transcend multiple boundaries (disciplinary, departmental, geographical, and technological). As illustrated in Fig. 3,
modeling is a formal process. It is not haphazard but rather methodical and controlled. The engineering model must be refuted and validated against reality, refined, and continually adapted since the real-world entity exists in an environment that is continually changing.

**Fig. 3 High-level modeling process flow**

Ozaydin & Tanik provide another depiction of the modeling process. As demonstrated in Fig. 4, the modeling process starts with first defining the real-world system that will be studied followed by the creation of a theoretical model. These steps are then followed by the generation of modeled results and interpretation of the modeled results to the real-world. This interpretation, in turn, may necessitate adjustments to the model (Ozaydin & Tanik, 2012). The arrows in the process signify the continual nature of modeling and resultant model. As the real-world environment changes, the model must also change.

**Fig. 4 The modeling process (Ozaydin & Tanik, 2012)**

Silos, discipline-specific jargon, geographical locations, departments, technological differences, multiple communication channels, and other barriers produce entropy on large research and other projects. Reducing this entropy requires an understanding of several foundational elements including *convergence* which has been researched by the National Research Council (NRC) and also Massachusetts Institute of Technology (MIT). *Disciplinarity*, as researched by Tanik & Atilla, is valuable in understanding how disciplines interact. Additionally, information on *communication theory* allows a clearer understanding of the complexity of communication that exists on projects. Finally, *modeling and tools* such as P3Tech, Mathematica, and others are key enablers of the transdisciplinary problem-solving process. Tools allow the problem domain to be modeled in a virtual environment thereby freeing the possible solution from the constraints imposed by the physical world.

Convergence improves the effectiveness of problem solving by increasing collaboration across the silos, sharing of critical data and technological infrastructure, and breaking down barriers that inhibit communication amongst disciplines. Furthermore, the traditional silo approach is no longer sufficient for solving society’s complex problems in a timely, quality, and cost effective manner. The traditional approach must be transformed from independent and autonomous agents to interdependent, collaborative, transdisciplinary, and convergent processes.

**METHODOLOGY**

The methodology utilized for this research paper and associated case study involved a phased approach and will continue over multiple years. The methodology consisted of the following phases of a planned multi-year effort:

1. **Analysis**
   a. During this step, *research* was conducted to identify a problem space that required collaboration across multiple engineering disciplines. The electrical substation design process was identified as a case study since it requires civil engineers, structural engineers, and electrical engineers to develop solutions to satisfy electricity needs. These disciplinary boundaries may impede timely problem-solving. The initial focus of the research was on the entire substation design process but was narrowed to more precisely focus on a critical subset of the overall process – the *bill of material (BOM)*. A bill of material is a key deliverable from the substation design process and represents a list of all of the individual parts (transformers, insulators, conductors, bushings, etc.) required to maintain an existing substation or construct an entirely new substation. It contains parts from multiple engineering disciplines. An incorrect bill of material can lead to costly construction delays. The capital budget for substation work at some energy companies can be hundreds of millions of dollars per year. During this analysis, it was determined that the bill of material process accounts for a big portion of the cost of
designing an electrical substation and consists of a number of manual steps that are both costly and prone to human error. Therefore, this problem domain provides an ideal space for researching and quantifying the value of transdisciplinary problem-solving in a convergent environment.

b. As part of this analysis, ethnographic studies of the problem space were performed. The ethnographic studies identified manual processing steps, automated processes, and opportunities for improvement.

c. Interviews of subject matter experts were conducted and documented. Some subject matter experts were engineers tasked with designing substations and others were engineers tasked with providing support for the bill of material solution. This diversity provided a cross-discipline view of the problem space.

d. An engineering model of the exiting substation design bill of material process was constructed, simulated, and reviewed with the subject matter experts. Metrics associated with time and costs were generated. This model served as a baseline point from which the value created by the new convergent solution can be measured.

2. Design/Construct/Implement

a. This particular phase required instantiation of a new solution to the manually intensive and potentially error prone bill of material problem into tangible software, hardware, and personnel solutions. Off-the-shelf software was purchased along with modifications to internally developed applications, databases, and graphical user interfaces. A key component of the solution was a centralized database where parts, from all disciplines, could be stored and accessible by all disciplines. Discipline-specific tools could also access the centralized data store.

b. Users were also trained to utilize the new solution. Training is vital to ensure the solution will be correctly utilized and the expected benefits are derived. This training is an ongoing process and continues into the future across an ever expanding base of users.

c. The new substation design bill of material process was modeled using P3Tech and automatically simulated to generate metrics associated with time and cost. These metrics provide a comparison point to the baseline metrics associated with the old process.

3. Validation

a. The simulated results from the new bill of material process model were compared to the simulated results from the old process model in terms of duration and cost. The difference between the two simulated models represents the quantified value demonstrated by convergence-based engineering modeling.

4. Synthesis

a. The research and findings from the electrical substation design bill of material case study were combined with the other research associated with convergence, communication theory, and transdisciplinary to form cohesive research paper.

b. Finally, the research paper was submitted to peer-reviewed conferences and journals for additional feedback from other researchers.

CASE STUDY

The case study in this research paper is associated with the bill of material (BOM) process utilized to design, maintain, or construct an electrical substation. Electrical substations are key components of the electrical grid infrastructure. Substations provide essential functions such as increasing or decreasing voltage levels along the grid which, in turn, allow the electrical energy to be consumed by a variety of customers – residential, commercial, or industrial. As noted earlier, a bill of material is a key document that lists all of the parts (transformer, conductor, etc.) required to maintain or construct the electrical substation.

An incorrect bill of material can lead to costly construction or maintenance delays, incorrect construction, or possible grid instability. Multiple discipline-specific (silo-based) engineering design drawings are input to the bill of material process. The engineering design drawings are created to document the type of work required on the substation such as excavation and drainage work (civil engineer), foundation and structural work (structural engineer), and electrical work (electrical engineer). Effective substation design requires transdisciplinary collaboration and convergence amongst multiple engineering disciplines, along with an integrated design approach.

The old bill of material process is manually intensive, has loose collaboration between the engineering silos (electrical, structural, and civil), and is more prone to human errors. The old process requires multiple manual review steps (up to three) to thoroughly review each of the drawings, assign part identification information to each part, and quantify the part information for creation of the bill of material. These multiple reviews are necessary to ensure the bill of material list is as thorough as possible. Construction maintenance delays due to an incorrect bill of material can be costly. Additionally, an incorrect bill of material can result in other costly issues including ongoing maintenance problems which can impact the reliability of the electrical grid.

The Fig. 5 symbolic model represents a subset (due column size limitations on the page) of the steps from Review Step #1 of the old process. As depicted, engineering design drawings are manually reviewed by engineers, the key mark (part identifiers) information for each part is manually
extracted from several software applications that contain lists of typical manufacturer or internally fabricated parts. Finally, though not depicted, the extracted key mark (part identifiers) is manually added to the design drawing and the design drawing is printed out and utilized as input into Review Step #2.

As depicted by Fig. 6, Review Step #2 of the old process basically performs the same steps as Review Step #1 with a goal of further ensuring the correct key marks (part identifiers) have been assigned to each part and that all parts have been correctly accounted. Review Step #2 is basically a manual quality check of the results of Review Step #1 using the same manually intensive steps.

Review Step #3 as depicted in Fig. 7, is the final pass utilized to validate the content of the drawings, ensure that all parts are accounted for, and all parts have the correct key marks (part identifiers) assigned. It uses the same manually intensive steps as Review Step #1 and Review Step #2. These multiple passes are indicative of the value placed on generating a quality bill of material and the potential cost that may result from an incorrect bill of material.

Figure 8 represents the final output from the multi-step process, the bill of material document. As depicted in Figures 5 through 8, the presence of so many manual steps in the old process creates inefficiencies and provides many opportunities to introduce errors into the process. Also, there is a lack of convergence or transdisciplinary integration amongst the engineering disciplines which creates additional inefficiencies. The introduction of convergence principles to the design of a new bill of material solution will allow the disciplines to work more collaboratively across disciplinary boundaries, reduce entropy and the overall cost, while improving the reliability of the process.

Automated simulation of the old model was conducted using the P3Tech value-based process modeling tool. Automated simulation in a virtual environment provides a significant advantage including the ability to trial various combinations of process changes without modifying the
physical environment. This is a less expensive approach compared to reconfiguring the physical environment and subsequently determining the process change does not work. Figure 9 provides the output from the P3 Tech simulated execution of the old bill of material process for a single design drawing. As depicted, the old process for a single bill of material takes approximately fifty (54) hours to complete at an estimated cost of approximately $7,502. If one hundred (100) of these bills of material documents are created by a company each year, the company will spend approximately $750,200 per year to generate its bill of material documents. Also, this dollar value does not consider the cost of errors resulting from an incorrect bill of material process.

| System: | PPO2 |
| Instance: | PPO2INS11 |
| Default Transition Duration: | 10.00 Hour(s) |
| Default Distribution: | Constant |
| Time Unit: | Hour(s) |
| Simulation Elapsed Time: | 54 Hour(s) |
| Total Transitions: | 16 |
| Activated Transitions: | 16 |
| Active Time Cost ($): | 5782.62 |
| Idle Time Cost ($): | 1718.94 |
| Fixed Cost ($): | 0.00 |
| Total Cost ($): | 7501.56 |

Fig. 9 Simulation results from P3Tech modeling tool (OLD)

Parameter of interest for the new process

The old bill of material process was analyzed, modeled, and simulated. Cost and duration metrics were generated for subsequent comparison to a future new process. For the new bill of material process, the parameter(s) of interest or goals and objectives were established. In this particular instance, the parameter of interest or goal of the new process is to “Improve the efficiency of the bill of material process by 25%.” This improvement will be accomplished through the reduction of manual activities and integration of discipline-specific knowledge and data into a convergent tool framework. A key element of achieving the goal is the implementation of software applications, tools, and a centralized data store to facilitate a more automated bill of material process. Establishing the parameter of interest upfront provides a holistic or overall view of the objectives of the project or research study. Holism (parameter of interest), when combined with reductionism (detailed modeling), allows a better understanding of the problem domain to be obtained.

Case study findings

A transdisciplinary core team of engineers from structural, electrical, civil, and software along with project management and information technology support was assembled to provide oversight and approval of the project deliverables. The solution was designed to be a “substation design solution” rather than a civil engineering solution, an electrical engineering solution, or a structural engineering solution. Transdisciplinary teams are a key component of problem solving and achieving the opportunities afforded by convergence principles. During the design/construct/implement phase of the new solution, discipline-specific software applications were still required but the disciplines now leverage an integrated database (library) of parts with automated connections to back-office systems. Some discipline-specific tools were still required since wiring diagrams utilized by electrical engineers are very different than site components such as hydrology, drainage, and retention elements required by civil engineers. However, all disciplines now take advantage of a catalog of available parts located in an integrated and centralized database. Furthermore, these parts should already have the necessary attributes (part identifier, model number, manufacturer, length, width, depth, etc.) assigned to them or the information can be more easily looked up using automated processes. This convergence of data and knowledge from all disciplines into a centralized data store reduced many of the manual steps in the old process that were required to manually locate, assign, and count the individual part quantities. Additionally, all of the bill of material reports for the disciplines can be generated from the same centralized data store. This improves consistency across all disciplines. As depicted in Figure 10, the new bill of material process is much more streamlined compared to the old process (figures 5 through 9) and has eliminated many of the manual steps such as manually searching for part identifiers and manually counting the number of each part type. In the new process, as the design is being created by the engineers, the attribute information (key mark or part identifier, manufacturer, etc.) is automatically assigned to each part. The information for all parts associated with the design is then stored in a centralized database. After the engineer does a final quality check of the design drawing, the key mark quantities are automatically calculated using the stored information and fed into an application utilized to format the bill of material document. The three rounds of manual reviews and manual counting of part quantities have been eliminated resulting in a streamlined and more automated process as outlined in Fig. 10.
An automated simulation of the new process using P3Tech was conducted. Per the simulation output depicted in Fig. 11, the new process for a single bill of material takes approximately one (1) hour to complete at an estimated cost of approximately $88. If one hundred (100) of these bill of material documents are created by a company each year, the company will spend approximately $8,800 per year to generate its bills of material documents. If one hundred (100) bill of material documents are created by a company each year, the company will spend approximately $8,800 per year to generate its bills of material documents. Also, this dollar value does not include the savings associated with a reduced likelihood of errors that result from a less manual process and a more accurate bill of material. The key point in the new bill of material process is that the engineering design drawings enter the process with all of the necessary part identifier (key mark) information already assigned by leveraging a centralized catalog of 3-dimensional modeled parts to build the design drawing. This centralized catalog of 3-dimensional modeled parts with attribute information is a key element in improving the efficiency and accuracy of the bill of material process and provides a point of convergence for multiple engineering disciplines and data values.

**Fig. 10 Final bill of material (NEW)**

**Results**

The old bill of material process was modeled and subsequently redesigned using convergence principles such as centralized data store and transdisciplinary collaboration. Likewise, the new bill of material process was modeled and simulated. The results from the comparison of the two automated simulations (Fig. 9 and Fig. 11) indicate a potential savings of $7,414 for one (1) bill of material document. If one hundred (100) bill of material documents are created, the company will save $741,400. These are tremendous savings and further underscore the value that can be derived from using convergence principles and modeling in the design of solutions to facilitate transdisciplinary collaboration across engineering disciplines.

The final results from the case study outlined in this paper must continue to be validated since the solution is being rolled out in a phased approach to the user base. Since the final rollout may take months to years, it is beyond the scope of this particular research article.

**CONCLUSION**

Engineering problems and associated solutions do not respect disciplinary walls or boundaries erected for human control. Therefore, problem-solving must evolve from silo-based approaches to transdisciplinary endeavors in order to solve these problems in a timely and cost-effective manner. The combination of convergence, disciplinarity, communication theory, and engineering modeling provides a good foundation that can be leveraged to solve complex engineering problems in an increasingly transdisciplinary environment.

While much progress has been made understanding and applying the foundational concepts (convergence, disciplinarity, modeling, etc.) individually, future research must lead to the development of a formal framework or methodology that can be leveraged to conduct transdisciplinary problem-solving if the true value of convergence is to be achieved.

**ACKNOWLEDGMENT**

This paper acknowledges information and support provided by Dr. Faud Sobrinho on the utilization of P3Tech for modeling and simulation.

**REFERENCES**


Massachusetts Institute of Technology. (2011). The convergence of life sciences, physical sciences, and engineering. MIT.


DESIGN AND EVALUATION OF STROKE REHABILITATION GAMES
BASED ON SERIOUS GAME AND CLINICAL KNOWLEDGE

Chien-Lung Hsu, Sheng-Hui Lu
Department of Information Management, Chang Gung University
Tzu-Wei Lin
Graduate Institute of Business and Management, Chang Gung University
{ chhsu@mail.cgu.edu.tw M04440004@stmail.cgu.edu.tw
 D0640001@cu.edu.tw }

ABSTRACT

Stroke is one the most common disease in Taiwan. The important part of curing stroke is rehabilitation because evidence shows that the state of stroke can be controlled effectively through timely treatment and rehabilitation. We designed a stroke rehabilitation game based on serious game and clinical knowledge by having discussions with experts about functions, interface, and other requirements of the stroke rehabilitation game. We also evaluated usability through interviewing users to collect feedbacks and satisfaction. Our main contributions are described as below. (a) We designed and developed stroke rehabilitation game by following Information System Research framework as a guideline. (b) Clinical knowledge is considered in movement design, game design, and interface design. The contributions make rehabilitation game in this study be suitable for stroke patients.

INTRODUCTION

World Health Organization had shown that stroke was the second among the top ten leading causes of death in 2015 which mean that there were over six million population passing away because of stroke (World Health Organization, 2017). World Stroke Organization says that rehabilitation is an important part of curing stroke. Through timely treatment and rehabilitation, the state of stroke can be controlled effectively (World Stroke Organization, 2017). In the past time, physical therapists help patients do rehabilitation with their bare hands or assistive devices. Most of the patients give up because they consider that rehabilitation is time-consuming and boring.

Serious game was proposed first by Clark and defined as a digital game applied in other professional fields except entertainment (Abt, 1970). Kinect is a device which can recognize movements of a user by tracing information of human skeleton (Microsoft, 2015). We utilized Kinect to develop a rehabilitation game based on serious game for stroke patients, and the stroke rehabilitation game can solve problems of traditional rehabilitation and make rehabilitation more fun. The properties of stroke rehabilitation game below may increase rehabbing willing of patients. (a) User is not restricted by time during rehabilitation. (b) We cooperated with experts in stroke to design rehabbing movements, game content, and interface presentation. (c) User can select difficulty level depending on physical situation. (4) Game information can be a reference for physical therapists.

RELATED WORKS

We discuss Information System Research framework, serious game and application of Kinect in stroke, usability evaluation, and user-centered design in this section.

Information System Research Framework

Information System Research (hereinafter referred to as “ISR”) proposed by Hevner et al. in 2004 includes behavioral-science and design-science. The behavioral-science aims at developing and verifying theories which explain or predict human or organizational behavior. The design-science paradigm aims at extending the boundaries of human and organizational capabilities by creating new and innovative artifacts. Both are foundational to information system discipline (Hevner, March, Park, & Ram, 2004).

Hevner proposed an enhanced ISR framework in 2007 to emphasize cyclical interaction including Relevance Cycle, Rigor Cycle, and Design Cycle. The Relevance Cycle inputs requirements from contextual environment into research, and then introduces the research artifacts into environmental field testing. The Rigor Cycle provides grounding theories and methods with domain experience and expertise from the foundations knowledge into the research, and then adds the new knowledge generated by the research to the growing knowledge. The Design Cycle supports a tighter loop of research activity for the construction and evaluation of design artifacts and processes. These three cycles in a research project clearly positions and differentiates design science from other research fields (Hevner, 2007). Schnall et al. had designed and developed a user-central mobile phone application with cyclical ISR framework (Schnall et al., 2016).

![Fig. 1 Design science research cycles (Hevner, 2007)](image)
We followed the cyclical ISR framework (Hevner, 2007) as a guideline and modified research process of Schnall et al. (Schnall et al., 2016) to design and develop the stroke rehabilitation game.

Serious Game and Application of Kinect in Stroke

Serious game was proposed first by Clark C. Abt and defined as a digital game applied in other professional fields except entertainment, such as education, medicine, military, etc. (Abt, 1970).

Webster et al. reviewed literature related to early treatment and stroke patients and found that Kinect has a high accuracy of capturing movement, so Kinect became a home rehabilitation tool (Borghese, Pirovano, Mainetti, & Lanzi, 2012; Webster & Celik, 2014). However, a rehabilitation tool should be based on functions of rehabilitation (Borghese et al., 2012). Galna et al. applied Kinect in Parkinson's disease research (Galna et al., 2014). Although they formed a workshop to recruit users and discuss system design together, the safety of rehabilitation movements still should be evaluated (Galna et al., 2014).

In this study, the stroke rehabilitation game is based on clinical knowledge of nursing staff and physical therapists. Rehabilitation movements were designed through cooperation between nursing staff and physical therapists’ experience and information technology staff, and we invited patients to join testing and discussion.

Usability Evaluation

Usability evaluation means that the products or services are evaluated by representative users, and the researcher can understand the thoughts and feelings of users when using products or services and find out potential problems. Nielsen et al. proposed that we could find most of the problems related to usability through 3 to 5 testers (Nielsen & Landauer, 1993). Lewis proposed post-study system usability questionnaire (hereinafter referred to as "PSSUQ") (Lewis, 1991, 1992). PSSUQ is a seven-scale questionnaire with 19 questions (Lewis, 1995). PSSUQ has four major criteria. Question 1 to 8 evaluate system usefulness, 9 to 15 evaluate information quality, 16 to 18 evaluate Interface quality, and 1 to 19 are overall evaluation (Lewis, 1995).

User-centered Design

User-centered design means that developers should develop system form user’s view (Abras, Maloney-Krichmar, & Preece, 2004). Developers should design and develop again and again to continually improve quality of system based on feedbacks. For example, Schnall et al. found system’s content and requirements of functions through a focus group, and they found most useful functions and user interface (Schnall et al., 2016).

We cooperated with nursing staff or physical therapists who has extensive experience to design the rehabilitation game which is suitable for stroke patients.

METHODS

We followed the cyclical ISR framework (Hevner, 2007) as a guideline and modified research process of Schnall et al. (Schnall et al., 2016) to design and develop the stroke rehabilitation game. The first part is design of research process, and the second part is system structure of stroke rehabilitation game.

Design of Research Process

We consulted professional nursing staff and physical therapists at Relevance Cycle of ISR framework. The nursing staff should possess at least 10 years’ experience of care treatment, and physical therapists should possess over 10 years’ experience of rehabilitation. We discussed with the nursing staff and physical therapists about requirements of rehabilitation. At Rigor Cycle, we searched for literature related to rehabilitation and clinical knowledge. We developed and continually improved stroke rehabilitation game through interview experts. Finally, we invited users to test the stroke rehabilitation game, and we interviewed users after testing and collected feedbacks for improvements. Figure 2 illustrates research process of the study.

![Fig. 2 Research process of the study](image-url)
System Structure of Stroke Rehabilitation Game

Figure 3 illustrates the system structure of stroke rehabilitation game. The stroke rehabilitation game includes upper limb, lower limb, and weight shifting drill. There is a database for storing information of user in the game. Rehabilitation movements for each drill were designed through discussions with nursing staff and physical therapists. The information of invited nursing staff and physical therapists are shown at Table 1.

SYSTEM IMPLEMENTATION

We introduce development environment and tools, implementation and improvement, and testing in this section.

Development Environment and Tools

We used Windows 8.1 as our development environment. We utilized Kinect, Vizard 5.0, Faast, and Kinect SDK 2.0 as development tools. Vizard is a virtual reality development software which can create 3D content and support Python. Faast is an open source middleware released by MxR Lab, University of Southern California, and Faast can detect information of human skeleton between game and virtual reality application. Kinect SDK 2.0 is a software for developer to develop non-commercial software through Kinect. We used Python as program language and MySQL to build database.

<table>
<thead>
<tr>
<th>Expert</th>
<th>Title</th>
<th>Year of working experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Neurosurgeon Nurse head of surgical ward</td>
<td>12.5-year MICU of neurology 0.5-year SIICU of neurosurgeon 5.5-year neurosurgeon surgical ward</td>
</tr>
<tr>
<td>B</td>
<td>Neurology Nurse head of medical ward</td>
<td>9-year neurosurgeon surgical ward</td>
</tr>
<tr>
<td>C</td>
<td>Neurosurgeon Nurse head of surgical ward</td>
<td>12-year neurosurgeon surgical ward 8-month neurology medical ward</td>
</tr>
<tr>
<td>D</td>
<td>Registered nurse</td>
<td>4-year general medicine 5-year ICU 1-year transplantation surgery 1.5-year neurology</td>
</tr>
<tr>
<td>E</td>
<td>Physical therapist (team leader)</td>
<td>25-year physical therapist</td>
</tr>
<tr>
<td>F</td>
<td>Physical therapist</td>
<td>6-year physical therapist</td>
</tr>
</tbody>
</table>

Table 1 Experts’ Experience

Implementation and Improvement

We implemented two upper limb drills named “Pick Fruits” and “Shoot Balloons”, a weight shifting drill named “Pick Eggs”, and two lower limb drills named “Step on Cockroaches” and “Dance Machine”. Figure 4 to 8 show game screens of each game. We took about 2 weeks to develop prototype for each game. The most important thing was satisfying requirements of experts through interview. The experts should be at least one experience nursing staff and one physical therapist, and they provided feedbacks and suggestions. Implementation and improvement would end when no more feedbacks and suggestions existed.
Testing

“Pick Fruits” was tested by a tall and strong stroke patient. The physical therapist taught the patient rehabilitation movements in advance. There were other nursing staff and physical therapists for preventing dangerous situations during playing, and we observed and recorded patient’s playing situation. After game ended, we interviewed patient for feedback about game and advise. We also discussed with physical therapist for improvement and interface design.

“Shoot Balloons”, “Pick Eggs”, and “Dance Machine” were tested by nursing staff. They evaluated games based on their clinical experience. We discussed with them several times for improvement and interface design. We also adjusted parameters of difficulty level under suggestions form nursing staff.

“Step on Cockroaches” was tested by two stroke patients. One is the same as “Pick Fruits”, and the other one is thin and full of confidence. The physical therapist taught them rehabilitation movements in advance. There were other nursing staff and physical therapists for preventing dangerous situations during playing, and we observed and recorded patients’ playing situation. After game ended, we interviewed patient for feedback about game and advise. We also discussed with physical therapist for improvement and interface design.

CONCLUSIONS

Rehabilitation is the important part of curing stroke because the state of stroke can be controlled effectively through timely treatment and rehabilitation (World Stroke Organization, 2017). We designed a stroke rehabilitation game based on serious game and clinical knowledge by discussing with experts. We designed and improved game through discussing with nursing staff and physical therapists. The stroke rehabilitation game was designed by iterative and incremental development. We discussed with experts, tested, and improved functions continually for achieving rehabilitation effect for stroke patients. The medical staff can know the rehabilitation situation through database of stroke rehabilitation game which records user’s information, and they can adjust difficulty level. We also evaluated usability through interviewing users to collect feedbacks and satisfaction. The contributions are described as below. (a) We designed and developed stroke rehabilitation game by following ISR framework as a guideline. (b) Clinical knowledge is considered in movements design, game design, and interface design. These contributions make rehabilitation game in this study be suitable for stroke patients.

DISCUSSIONS

Limitations

We only implemented the stroke rehabilitation game and related database at the present stage. If other related game is developed, we should extend our database for other developer to reduce development time. Although we got suggestions from professional nursing staff and physical therapists, we still want to collect more suggestions from experts of other related professions. We need more stroke patients to join stroke rehabilitation game testing to collect more feedbacks since we had only 2 stroke patients in this study. The interface and content of stroke rehabilitation game should be updated continually and able to be chosen by users to maintain using wills of users. There is still some difficulty about synchronisation of user’s movements. We could collect user information in stroke rehabilitation game through Kinect as reference for nursing staff or physical therapists.

Future Works

We will do quantitative research using PSSUQ to evaluate usability for stroke rehabilitation game. We will evaluate acceptance and continually use with flow theory (Csikszentmihalyi, 1975), and we can use Flow State Scale (Jackson & Eklund, 2002) to know the flow condition of tester when playing. We will also interview testers to collect feedbacks for improvement.

REFERENCES


A GENERALIZED APPROACH FOR PMT PERFORMANCE PREDICTION

Fan Xiong, Ed Marquette
Life Science Group, Bio-Rad Laboratories
{fan_xiong@bio-rad.com edward_marquette@bio-rad.com}

ABSTRACT
This paper proposes a simple mathematical model to predict PMT voltage based on a PMT datasheet. The model can also be applied to calculate the optical signal level of a particular channel in a flow cytometer given the PMT voltage and the channel number of the acquired signal.

INTRODUCTION
Flow cytometry is a technology to characterize cell populations while the sample passes a laser beam in a single file stream (Xiong & Murat, 2015) (Xiong & Marquette & Murat, 2015). Photomultiplier Tubes (PMT) are widely used in flow cytometry instruments as light detectors.

A typical PMT structure is shown in Fig. 1 which includes a photocathode, several dynodes and an anode. Photocathode is the surface that the photon hits and that hopefully emits a photoelectron. Dynodes are surfaces that either a photoelectron or electrons produced by photoelectrons hit and emit multiple secondary electrons. Each of the several dynodes in a PMT has an electrical potential which are arranged to create a force on the photoelectron or electron which will accelerate it into the next dynode with sufficient energy to produce multiple secondary electrons. Finally, the anode is the output pin that collects the output of the last dynode.

Fig. 1 PMT structure

A photon hits the photocathode and will excite an electron in the photocathode. If that electron has more energy than the work function of the photocathode, it will escape. The electron, once it escapes, can be accelerated at fairly high energy into a dynode with a low work function. There can be enough energy to produce multiple electrons. The process repeats and what started as one photon becomes on the order of a million electrons. The probability of getting a photoelectron in response to a photon is the quantum efficiency and is typically ~15% at 488 nm wavelength.

PMTs are made of low work function materials like alkali compounds. Different PMTs use different photocathode materials to trade off optical bandwidth, cost, noise, temperature sensitivity, etc. PMT manufacturers generally supply data as shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1 Example PMT information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cathode</td>
</tr>
<tr>
<td>Luminous</td>
</tr>
<tr>
<td>Sens.</td>
</tr>
<tr>
<td>PMTs</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Cathode Sensitivity can be either radiant (mA/W) or luminous (µA/Im). Radiant sensitivity is fairly simple. One 500 nm photon per second is 4.0*10⁻¹⁹ W. one mA is 6.2*10¹⁵ electrons per second. If the radiant sensitivity is 50 mA/W, one photon per second gives 50*6.2*4.0*10⁴ = 0.124 electrons per photon on average which is the quantum efficiency. On the other hand, Luminous sensitivity is complicated since it factors in the response of the human eye.

Anode Sensitivity is similar to Cathode Sensitivity but after the electron multiplication process, so the units are larger which are A/W or A/Im. Unlike Cathode Sensitivity, this is a function of the voltage applied to the PMT. Anode Sensitivity / Cathode Sensitivity is the electron multiplication gain, typically around 10² at 1000 V, 10⁶ at 700 V for the Hamamatsu R928 PMT which is studied here.

Anode Dark Current is a measure of how much signal is produced when there are no incident photons, expressed in nA. The statistical fluctuation of this current is the dominant source of baseline noise in a PMT.

R/W (Red/White) Ratio is indication of sensitivity at higher wavelengths, measured by applying white light with and without a red filter cutting off at certain wavelength, about 650 nm for most PMTs used in flow cytometry. The typical value is about 0.3, and a higher value is better, especially for higher wavelength channels.

Blue Index measures sensitivity in the 420 nm region which is of interest in scintillation counting. This parameter usually is not considered in flow cytometry.

PMT VOLTAGE BACKGROUND
When the PMT voltage changes, the electric field between dynodes changes, accelerating electrons into the next dynode with more or less energy and ejecting more or fewer electrons. Functionally, the signal goes up exponentially with voltage. For example, in the Hamamatsu R928 PMT an 11% increase in PMT voltage causes the output signal to roughly double.

PMT voltage optimization is recommended for flow cytometry instrument operation, and the relationship of PMT voltage, CV (Coefficient of Variance), signal to background ratio has been discussed thoroughly in (Perfetto 2013) (BD 2008). Poor optical alignment will result in higher PMT voltages. Poor dynode gain also results in higher PMT voltages. Both scenarios exhibit high PMT voltages, but poor optical alignment will show up as having high CVs, causing imprecision in the data, while poor dynode gain has a minimal effect on CVs and is not very problematic. If a high PMT voltage is set to run experiments on flow cytometers, some concerns may arise. This paper proposes an approach to predict PMT voltage based on the PMT datasheet data.

Applying the values to the dynode gain model \( G_d = K_1 \cdot V^{K_2} \), we have:

\[
10^5 = K_1 \cdot V^{K_2} \\
10^7 = K_1 \cdot (2V)^{K_2}
\]

Solve the equations and we have

\[
K_2 = 6.643856189775 \\
K_1 = 1.1706617558 \times 10^{13}.
\]

The specific datasheet supplied for each PMT only gives dynode gain at 1000 V – this means we can’t recalculate this equation for each PMT without assuming either \( K_1 \) or \( K_2 \) being fixed. There are two ways to model PMTs with different gains, one varying \( K_1 \) and one varying \( K_2 \). \( K_2 \) seems like the most realistic way to model, but may make it harder to come up with a correction. Using \( K_1 \) to model it has the disadvantage that the intercept is different, which doesn’t seem realistic. To ensure that we can still get a reasonable value for gain as a function of voltage by fixing either \( K_1 \) or \( K_2 \), we compared the results by fixing \( K_1 \) or \( K_2 \) as shown in Fig. 4, and it doesn’t matter much which parameter we fix.

Fig. 2 PMT and light signal detection

When laser light hits a cell, the light is scattered and fluorescent molecules are excited. The resulting signals are collected by PMTs and converted to electrical signals as shown in Fig. 2. The fluorescence and scatter signals displayed in flow cytometers are determined by

\[
S = K \cdot NP \cdot GP \cdot GD,
\]

where

- \( S \) is the signal level in the plot
- \( NP \) is the number of photons reaching the PMT
- \( GP \) is the quantum efficiency of the PMT (electrons/photon)
- \( GD \) is the dynode gain
- \( K \) is everything else (charge of an electron, etc.)

CVs are almost entirely determined by \((NP \cdot GP)^{1/2}\). Dynode gain is a very small factor in CVs, but Dynode gain does have big impact on signal level. Dynode gain can be expressed as \( GD = K_1 \cdot V^{K_2} \) and so dynode gain is a steep function of voltage.

QUANTIFYING DYNODE GAIN VS VOLTAGE

The R928 PMT data sheet specifies a dynode gain of \( 10^5 \) at 500 V and a gain of \( 10^7 \) at 1000 V as shown in Fig. 3.
**PMT Voltage Prediction Model**

We proposed a mathematical model to predict PMT voltage based on PMT testing data. The simple model is

\[ S = K \times G_c \times G_d , \]

where

- \( S \) is the signal strength (\([0,255]\) in linear scale, \([0,10000]\) in log scale)
- \( K \) is the optical signal with units of counts*lumens/\( \mu \)A
- \( G_c \) is the photocathode sensitivity with units of \( \mu \)A/lumen
- \( G_d \) is the dimensionless dynode gain

In order to verify the model, we have run a series of experiments. The experiment procedure is:

- Run Bio-Rad ProLine™ rainbow 8 peak beads and make sure that the performance is good in all channels, one example is shown in Fig. 5.

- Using Bio-Rad ProLine™ calibration beads, record PMT voltage needed to get signal to a level of 5, 10, and 20 * cathode sensitivity.

- Repeat for a wide distribution of PMTs.

- For each PMT and each of the three signal strengths, calculate a correction voltage and apply to the actual voltage needed.

- Compare the corrected and uncorrected voltages and see if the spread is sufficiently improved to warrant using the correction in the software.

**RESULTS**

We have experimented with 13 PMTs and rotate the PMTS to the FL1, FL3, FL4 channels of the S3 cell sorter. FL2 channel was used as a reference. Based on

\[ K = \frac{S}{G_c \times G_d} \]

\[ G_d = 1.1706617 \times 10^{-13} \cdot V^{k_2} \]

We have

\[ G_d_{1000} = \frac{\text{Anode sensitivity}}{\text{Cathode sensitivity}} \times 10^6 = 1.1706617 \times 10^{-13} \cdot 1000^{k_2} \]

We can solve \( K_2 \) first, and then solve \( K \).

\[ S = K \times G_c \times G_d \]

\[ S = K \times G_c \times 1.1706617 \times 10^{-13} \cdot V^{k_2} \]

Thus we have

\[ V = 10^{\frac{S}{K \times G_c \times 1.1706617 \times 10^{-13} / k_2}} \]

The actual voltage and predicted voltage comparison was shown in Fig. 6.

---

Fig. 4 PMT gains and voltage by fixing K1 or K2

Fig. 5 An example 8 peaks beads plot
We also compared the voltage at signal intensity of 5000 as shown in Fig. 7, which is half of the dynamic range of S3 cell sorter. Based on the comparison, we can tell that the prediction voltage is closely matched to the actual voltage with an average difference of 24v.
To further verify the proposed model, we repeated the same experiments on another type of PMT and the result is consistent.

CONCLUSION

This paper proposes a simple mathematical model to predict PMT voltage based on PMT testing data. The individual PMT dynode gain has virtually nothing to do with separation, sensitivity, linearity, and noise. It does affect the amplification of very weak signals. Based on the experiment data, predicting PMT voltage based on the PMT datasheet for the target signal level is feasible. Furthermore, this approach can be applied to other type PMTs for voltage prediction and PMT screening.

REFERENCES

Hamamatsu (2006), R928 PMT datasheet
ABSTRACT

Modern education has recognised the importance of serious games in achieving learning goals and games’ effectiveness in addressing learning through fun, enjoyment and motivation. The education material is often disseminated through interactive environments, which set up learning goals, evaluates student progress, allows unlimited repetition of game playing and personalizes learning. In this study, we illustrate the design and implementation of a novel serious game named Duke’s Disclosure, which addresses learning needs of optometry students. A particular development model was used to create the game architecture and define its rules. The architecture shows scalability: It could accommodate as many learning goals as required. The model of the game is reusable and could work across STEM education. The game has been developed with the Unity engine and tested in real life in a higher education environment.

INTRODUCTION

This research illustrates the design and implementation of a serious game, named Duke’s Disclosure. It can be used by optometry students to achieve learning goals related to clinical assessment, clinical reasoning and evidence-based practice in vision care. The purpose of the paper is to show important aspects in the design and implementation of the game:

a) An architectural overview of the game, which fits special requirements of the environments where serious games are designed and developed,

b) The relationships between the learning outcomes and the model of the game, applicable in optometry learning and

c) The level of reusability of the game in terms of its constant adaptation to learning outcomes and student assessment.

This is an experimental study, which proved the concepts proposed in the game design. It would need further investment if the game were to be commercialized in future.

The outcome of this study is interesting: We should be able to apply the same game design principles across various subjects in optometry and possibly in Science, technology, engineering and mathematics (STEM) education. The prerequisite is a clearly defined process, which would take the player towards achieving learning goals. Mapping the process to the overall architecture and model of the game has secured the achievement of learning goals. However, all other aspects of serious games are in place: players experience fun and learning through motivation.

The paper is organized as follows. In the next section we briefly overview related work. We could not find any published papers, which emphasized an overall architecture of serious games applicable in optometry learning and focus on a-c) above. Therefore, in this section we summarize the literature which underpin our study. In the Modelling section, we describe the scenario in which we place the development of the game, we define the SOAP structure (Cameron & Turtle-Song, 2002) in optometry learning and clinical optometry, outline the planning of the game and illustrate our design principles through the Use Case model and overall Architectural View of the game. The section finishes by outlining game rules and illustrating the relationship between the data input into the game and resources essential for playing the game. In the section named Creating the Game with Unity, we elaborate on user interfaces of the game, levels and scoring system and the challenges in the game. In conclusions we look at the results of the study and outline future work.

RELATED WORK

The growing interest in exploiting the power of serious games in education is still drawing attention of educators all over the world. Serious games research community have created numerous examples which illustrate how serious games positively affect the delivery of a modern curriculum to modern generation of our students. There are numerous
peer reviewed publications which highlight the power of serious games and which, at the same time, look at the way we can tune the games to fit educational needs. Computer game based working environment has a huge impact on learning environments and affects learning outcomes (Chatterje et al., 2011), which was recognized more than two decade ago (Randel let al., 1992), (Reiber et al., 1998), (Williamson, 2009), (Robertson and Good, 2005). Serious games and gamifications are sine qua non in many fields of training and education (Gentry et al., 2016), (McGrath and Bayerlein, 2013), (DeGloria et al., 2009), (Loh et i, 2009), (Kapp, 2012), (Connoly et al., 2012). We have no doubts that they have positive impact on achieving learning outcomes at various levels of education of an individual (Casey et al., 2014), (Fisch, 2005), (Kintsakis and Rangoussi, 2017). Computer games and serious games in particular exploit “fun and learning though motivation” as one of the main trigger to attract interest of modern learns ((Habgood and Aworth, 2011), (Malone, 1981), (Jovanovic et al., 2008), (Denis and (Jouvelot, 2005), (Mayo, 2007), (Huang et al., 2012), (Malone et Lepper, 1987). They are also coming close to instructional goals in learning (Zapusek et al., 2011), (Garrist et l., 2002) in which instructional content and game characteristics are supposed to be interwoven into the game model. These are every important factors when building a game engine (Serrano-Laguna et al., 2015), (Freirre and Fernandez-Manjon, 2016). It is also important to mention that we do have a valid evaluation mechanism within games, which would look at learning outcomes (Gagne et al., 1992), tools and methods for their efficient design (Marfissi-Schottman et al., 2010) and frameworks which we can use in serious games validations (Yousoff, 2010), (Mayer, 2012).

Serious games in medical education have also come of age (Aki et al., 2010) (Bochennek et al., 2007), (Blakely et al., 9) but there are no available publications which focuses on the issues relevant for optometry students. We have found only one publication relevant to our work (Trevino et al., 2016) in which the authors compared the effectiveness of educational board game with interactive didactic instructions when teaching optometry students basic and applied science. Their results show that educational game and interactive didactic instruction can be equally effective, and that both modes have the potential to be equally engaging and enjoyable.

In this study we taken this a step further to serious computer games, conducting the design and implementation of the proposed Duke’s Disclosure according to the information and research results we collected by reading the literature listed in this section.

MODELLING

The Scenario

Our game is an action adventure video game for smart devices which is meant to be fun to play and should motivate players to sharpen their knowledge in biomedical fields. The game’s goals are to raise the study quality through technology, improve student progression, increase grades, raise the teacher’s competence in e-learning, and give them more opportunities to follow up on the students’ progress.

The player of the game is an optometry student who has to apply theoretical knowledge in anatomy, physiology, neurology, pharmacology and pathology, practice and develop knowledge about clinical assessment of patients with vision and eye related problems, define the problem(s) and prepare a plan on how to remedy the problem(s). The player should follow the SOAP structure (Cameron & Turtle-Song, 2002) for problem-oriented medical records as point of departure for playing the game The game should give opportunities to the player to use multiple choice questions in order to experiment with possible solutions to the problem and collect results of all decisions made by the player. These decisions could not solely be connected to the rules of the game. The scoring system must also be able to monitor player’s decisions when examining patients. The learning outcome would be achieved if the player ends up the game with an adequate score.

Duke’s Disclosure draws its name from the world of the famous count from Transylvania.

The SOAP structure

SOAP structure is an organized approach for clinical assessment and medical record keeping (SOAP notes). SOAP is an acronym for subjective (S), objective (O), assessment (A), and plan (P), which each represent sections of the patient record (Cameron & Turtle-Song, 2002), (Weed, 1964). The SOAP structure follow the natural flow of the clinical encounter in optometry; that is patient history (S), clinical assessment (O), clinical reasoning (A), and management (P) following evidence-based practice. When studying optometry, students should gain knowledge, skills and competency in the process of providing vision and eye care. The process contains activities which have been grouped into Subjective, Objective, Assessment, and Planning. The student will get a subjective explanation of the symptoms from the patient (S), then create a more objective description of the condition by choosing methods of clinical examination (O). After that, the student will have to assess the test results (A) and make clinical decisions about diagnosis, etiology, status of vision and health, and prognosis and choose how to deal with the problem (P).

Planning the Game

Planning of serious games should conform to nine general steps of instructions essential in learning (Gagne and Briggs, 1992). In our preliminary research, we were able to articulate the importance of developing serious games in optometry learning because of the lack of available games that could support learners in this field of science. Therefore, we planned to place learning goals in optometry within a framework, which helps in learning and remembering and follows SOAP structure. In the current version of the game, we avoided memory overload and provided guidelines for learning through additional boxes and animation. We were
also able to secure that the learner (player) practices skills and applies knowledge.

When developing the game, we have also conformed to the advise from (Loh, 2009) and (Marfisi-Schottman et al., 2010) and included an analysis of choices of game models, a detailed scenario and narratives for the game, and pedagogical quality control. We constructed the game from reusable software components in order to address the scalability and reusability and focused on usability testing before the game was released.

**Use Case Model of the Game**

Figure 1 shows the main functionality of the game, including both: the roles and interests of the Game and Player as main actors. It is important to note that UML use cases are expected to show how these two actors would use “computations” which exist within the game. In other words the UML semantic requires that we interpret the functionality collected in use cases of the perspective of actors, which use that functionality. The relationship between actors and use cases must be defined as “use”. Game actor uses “manage inventory” and “manage cases” within the game, but also “updates the score” of the player, when required. Therefore we use a stereotyped dependency <<extends>>. Player actor uses use cases which follow the functionality of the SOAP structure and therefore secures that the player experienced the procedure (i.e. following the “steps” within the game) which would lead to either the assessment of player’s performance or achieving learning outcomes specified in the inventory of the game.

The reader should pay the attention that in some of the use cases, identified in the SOAP structure, the player either performs examination or conducts a dialog with the patient. At this abstract level of modelling, we cannot reveal which.

The reader should pay the attention that in some of the use cases, identified in the SOAP structure, the player either performs examination or conducts a dialog with the patient. At this abstract level of modelling, we cannot define exactly level of computations or “coding” within the game would be required for performing examination of the patient or having a dialog with him/her, because it is difficult to predict which choices the player will make. Data collected through the dialog, and during the examination of the patient, would be relevant for continuing playing the game and consequently performing the rest of SOAP activities. In other words, the data collected at this level would determine the execution of functionalities found in our use case model.

However, the <<extend>> use case “Use Dilation Remedy”, for the base use case “Perform Objective examination of the patient”, remains a very important “computational” part of the game, particularly in the process of achieving learning outcome.

**Fig. 1 Use cases for Modelling the Game**

**Overview of Game Architecture**

Figure 2 shows an architectural overview of the game. The top layer in the architecture illustrates the types of interfaces created in the game, which are self-explanatory considering that we are developing serious games. Interfaces specific to the SOAP structure are taken from the use case model in Fig. 1 and grouped into an eye examination.

The choice of resources are very relevant to the game because they secure both: playing the game and achieving learning goals. We overview them in paragraphs below.

Dialog Content is a resource which holds information on the dialog between the player and non-playable characters (NPC). It stores information on what NPC says to the player. We store ready-made answers on the dialog in XML files. The response of player may trigger response from NPCs. Every time the player talks to NPC, the dialog is initiated. Under the eye examination, from the SOAP structure, the dialog is crucial for preparing the SOAP notes for the patient record.

Patient Information is a resource, which stores all relevant information on eye examination. It stores images for the problem patient may have, descriptions of the problem and solutions to the problem which appear as multiple-choice questions in the game. In examinations, we have to start with the dialog first (subjective, as required by the SOAP structure!) and therefore this resource must know which dialog content is relevant to the examination.

Level Score is a resource, which keeps track of player’s performance in eye examination.

The inventory is a resource that keeps track of all items the player has collected in the game world that may be used later in eye-examinations. The inventory is a part of the main character object and is made to store multiple objects of a certain type. The objects that are stored by the inventory found in the levels of the game and can have any sprite, name, and description.
The co-relation between user interfaces and resources in Fig. 2 is color coordinated. This means that information available through a particular interface reaches various resources. Therefore Patient Information is fed by information from all interfaces from Eye Examination and Subjective interface will be the only one having impact on Dialogue content. Game World interface can influence Level Score and Start/Pause menu is important for the game engine, as expected. The semantic behind this color coordination is a part of our design decision on how to create the game. It should be revisited in future for two reasons. The first one is the choice of resources, which we may change in future because we may want to have better matching between interfaces and resources. The second reason is the enhancement of the assessment of the student progress, in which cases Level Score resource might not be sufficient.

Any of these changes would not affect the architecture in Fig. 2. We can add more resources and interfaces, as long as we can define the co-relation between them. The layered architectural style from Fig. 2 guarantees the flexibility and will not limit options in the number and type of interfaces and resources we may have. As in any other software architectural style, the relationships between layers of software components may be interpreted as constraints and therefore affect the game, and thus the flow of information between interfaces and resources should be revisited whenever we change the learning goals and assessment.

In Fig. 3 we illustrate the pathway from the Objective interface in Eye examination interface group towards Inventory, Patient Information and Level Score resources which is illustrated as a RED line in Fig. 2.

**Game Rules**

The game is played on a smart device (iPhone, iPad, android device) individually and is not designed to be used by groups. The only requirement to play the game is to have the game installed. Prior knowledge in relevant medical fields is not required but recommended to get the most out of the game. The player will control a character in a virtual world to help patients, overcome obstacles and explore mysteries of the game-world.

The game is designed to be played in short burst mode on a daily basis. The game is not depended on sounds, it can be played without internet connection, and can be paused anytime so it can be used in most situations where the player has some spare time.

**CREATING THE GAME WITH UNITY**

**Creating User Interfaces for the Game**

Figure 2 has introduced a variety of user interfaces, which convey information to the player in the game, but also let the player interact with the game by controlling the main character and using menus. The user interfaces that are used, while controlling the main character in the game, consists of the actions the main character can perform, the player’s inventory, the player’s life bar and a pause menu. The other important interfaces are Start and Pause menu, and various dialogues and interactions with characters in the game, as collected in the Eye Examination from Fig. 2. All user interface elements in the game follow the same theme to fit the Transylvania inspired style of the game: Many buttons and interfaces look like they are made of wood.
Figure 4 shows how the buttons which control the characters are designed. They can be easily recognized and placed for easy access for thumbs: arrows left and right to move, a hand to interact and use certain objects in the game world, a gun reticle to shoot, and an arrow facing upward for jumping. The player also has some information about their status in the game. At the top left corner, the player’s health status (life bar) is located. The health display will automatically extend if the player’s health changes.

In the game, we also have interfaces for giving the player plain text, having a dialogue with another character, and for examining a patient. The interface for displaying plain text to the player is used when the player finds and reads notes in the game or if the main character says something to him/herself. The dialogue interface is used to talk with characters met by the player. It lets the player choose from different responses which can influence what the other character might say. As you can see from “Fig. 5”, the player is presented by a question, may have multiple answers to choose from. The player may also leave the conversation at anytime by pressing the “Exit” button.

An instance of interfaces related to the patient examination is shown in “Fig. 3” and “Fig. 6”. The interface contains information about the problem the player must resolve and up to four different solutions to the problem (resembling MCQ). In some cases, an image that can be viewed may add further information to the player for solving the problem. In “Fig. 6”, the player is inspecting an image of the patient’s retina and has to find out which dilation remedy to use to provide a better view. The player has found that two of the four are possible answers. If the player wanted to choose one of the grayed-out remedies, the player would need to find them in the game world first. This creates a link between “what the player does in the game world” and the “eye examinations”. Consequently, the player “feels” that performing the examinations is quite realistic when playing the game.

Challenges in the game

One of the most important aspects of having fun when playing games is facing interesting challenges. In Duke’s Disclosure, they come in many forms. In addition to challenging the player faces when manipulating optometry knowledge, the game also has other challenges such as puzzles to resolve, enemies to defeat, and traps to overcome.

The puzzles in the game can be created by certain items that must be destroyed, with buttons that must be pressed, labyrinths to be concurred or something else, where the player progresses by making choices.

The enemies make the game livelier. They move around in the game world with their own rules on how to behave when they spot the player. The player is tasked with learning how the enemies behave and finding the best way to defeat them.
In “Fig. 7”, we see the bat enemy and its flying trajectory. The bat enemy often waits for the player “under the roof” and when the player approaches the bat, the bat flies towards the player. However, the bat is approaching the player in a diagonal line, which creates an interesting scenario for the player, because the player can only shoot in a horizontal line. Therefore, the player must either jump and then shoot or wait for the bat to fly lower and then shoot. Jumping up may lower the distance between the player and the bat, which could be a dangerous move. However, it may be a necessary action if there is not enough room behind the player. Moving away and letting the bat get lower and lower, before attacking it may be safer, but it comes with the risk of being cornered.

The traps in the game often challenge the player’s skills in the game. It is a reaction test in form of a rolling boulder that needs to be jumped over, a jumping challenge in forms of a pit that must be crossed, and an observation test in form of stalactite that might fall when the player walks under. These traps are also used in puzzles, where the player might have to let a stalactite fall from the ceiling in a cave, in order to use them as a higher platform for a jump.

Levels of the game

The levels are always designed around the eye examinations. The objective of each level is to complete the eye examinations and go through a final door. The eye examinations are mostly spiced up with various gameplay objectives the player must complete to find either the patient or the tools needed in the examination.

We have designed the levels to be around four to eight minutes in length and with two to three examinations in each level. With this amount of time, the players can face moderate challenge each day if they play the game.

The levels in the game have the potential to vary in difficulties in terms of both: Gameplay difficulty and the eye examination difficulty. Enemies can be adjusted to be weaker or stronger, traps can be adjusted to be easier to overcome, and the level of the professional content can be varied across any subject in the optometric/medical field. This secures that we can tailor the game levels for a specific purpose.

Scoring System

The scoring system is important for several reasons. It is meant to motivate the player to put more effort in the optometry challenges within the game and let players know how well they are doing in a particular subject. The score is accumulated from the players performance under the patient examinations in the game.

After completing levels in the game, the player will receive a star rating from zero to three. The score is displayed over the entrance to each level. If the score is low the player knows that he/she needs more training in the particular medical subject the level contains. The scoring system can also motivate players to challenge each other to see who can achieve the most of three-star ratings.

CONCLUSIONS

Evaluation

One of the most important outcomes of this study is the use case and architectural models of the game, available in Fig. 1 and Fig. 2, which adhere to the SOAP structure and allows amendments in resources available for the game. Therefore, in situations where we have to change either the process of achieving learning outcome or the SOAP structure, we could still use the same model of the game and pay attention to the co-relation between user interfaces and resources, enabled through the game engine. The resources and their content available in Fig. 2 may change in future. However, as long as we could map these relationships with the process of learning, such as the SOAP structure, we can achieve learning outcome. We will be able to amend dialogs, inventory and scoring system and influence decisions on what should “patient information resource” contain.

Consequently, the architectural overview of the game in Figure 2 shows explicitly its reusability. Changes in any of the game resources will not affect the game model and the changes in the way learning outcomes are defined, would require updating of these resources. Also, the adherence to the Model-View-Control pattern (Gamma et al., 1994) in Figure 2 additionally secures the reusability of the model.

This paper deliberately focuses on the modeling aspects of the games and goals specified in a)–c) from the introduction. This means that all other components of the game design, which make it a serious game, are still present, but the subject of a different publication. It is important to note that the game has been planned according to practices essential in serious games developments and the technologies available and therefore players experienced fun and are motivated to play the game. Our short illustration of interfaces, levels and challenges of the game give a glimpse of learning through fun and motivation.

There are a few issues, which should be debated before we undertake future work. Our choice of resources and their co-relation with user interfaces through the game engine, available in Fig. 2, is a result of experimenting in the modelling of our game for optometry learning. Information in these resources may overlap, they may be accessed through various user interfaces from the top layer of the architectural model and therefore they do not store the semantic of significantly different resources. Therefore, more work has to be done in the analysis of the role of these resources for the reusability of the game. Currently, Fig. 2 works perfectly well in the setting where SOAP is followed, and scoring system works as defined by optometry professionals. If we wished to expand the game towards different fields in optometry learning then the choices of resources and their contents should be revisited.
Future Work

One of the first steps in our future studies is to address three future important aspects of the roles of serious games in optometry learning.

1) We should revisit our procedure of planning, designing the game according to widely recognized practices and include informative feedback, assess performance tests and enhance retention and transfer as suggested in (Gagne et al., 1992).

2) We should look at our overall architectural model and assess if it fits any other known generic conceptual framework, such as the one proposed in (Zolotaryova and Plokha, 2016). In our model, the relationship between game resources and the interfaces which facilitate the data entry to these resources are done through the engine, which can be seen as a limitation of our architectural model. However, if we manage to create resources in Fig. 2, which can address assessment measurements and their criteria, as advised in (Zolotaryova and Plokha, 2016), this would automatically make our solution more reusable. This means that we have to focus more on the assessment (summative and formative) and investigate if our game would not only fit proposed conceptual frameworks for serious games, but also show a clear correlation between learning outcomes and the format of the assessment utilized by the game. This is one of the most important goals of modern education.

3) We have to look at interactive learning instruction designs in serious games and compare it with the functionality of our designs form Fig. 1 and Fig. 2.

4) The effect on learning outcome should be assessed in a randomized controlled trial.

Acknowledgement

We would like to thank our fellow students Konrad Gislerud Andresen, David Viking Danielsen Lund, Henrik Bekkhus Solberg, John Morgan Haugen, Steffen Amandus Nielsen, Ivan Hansen, and Kim André Myre from Department of Science and Industry systems, University College of Southeast-Norway, Norway, and Theodora Anitei and Vanessa Giovannini from the Department of Computing and Information Systems, University of Greenwich, for their contribution in creating and testing the game. This work was funded by Norgesuniversitetet (P19/2015) and Faculty of Health and Social Sciences, University College of Southeast-Norway.

REFERENCE LIST


Gama, E., Helm, R., Johnson, R., & Vilssides, J. ( ). Design Patterns: Elements of Reusable Object-Oriented Software. Published by Addison-Wesley Professional.


Huang, C., Liu, E., Liu, L., & Lin, Chun. (2012). A development and Evaluation of Educational Board Game Design Course: An Example of Pre-service
COMPARING USABILITY OF BUSINESS RULE VALIDATION PROCESSES IN HEALTH INFORMATION SYSTEMS

Ramaraju Rudraraju, Nick Timkovich, Ryan Cantor
Kirklin Institute for Research in Surgical Outcomes, University of Alabama at Birmingham
{ramaraju@uab.edu ntimko@uab.edu rycantor@uabmc.edu}

ABSTRACT

There was an exponential growth in the usage of Health Information Systems (HIS) in the last three decades. In addition to delivering patient care, health care professionals are often burdened with complicated regulatory and documentation requirements. Existing literature shows that introduction of a poorly designed HIS into a healthcare setting causes additional burden to the health care professionals. Building well designed HIS based on sound principles of usability engineering is of prime importance. In this paper, we compare two HIS developed by our team with contrasting validation processes namely soft validation and hard validation, based on the usability best practices suggested in the literature. We begin with a hypothesis that the system developed using soft validation is more usable than the one developed using hard validation. We compare the timestamps from our ticketing system and data change logs in order to determine the time taken to fix user mistakes in both the systems, in order to evaluate their usability. Our analysis of the data affirmatively confirms our hypothesis that a system implemented using soft validation is more usable than the system implemented with hard validation.

INTRODUCTION

Intermacs (Interagency Registry for Mechanically Assisted Circulatory Support)

Established in 2005, Intermacs is a North American registry for patients who are receiving mechanical circulatory support device therapy (implants) to treat advanced heart failure. Clinical data related to mechanical circulatory support devices (MCSDs) from initial implant time through follow-up evaluations are collected in Intermacs. Follow-up data after the first implant are collected at 1 week, 1 month, 3 months, 6 months and every 6 months afterwards. Major events after implant, e.g., death, explant, re-hospitalization and adverse events, are entered within 30 days of occurrence and also as part of the defined follow-up scheduled intervals. Currently 170 hospitals from all over the U.S participate in Intermacs enrolling 22,257 patients. Data is entered by sites using a Web Based Data Entry (WBDE) system. The collected data is analyzed to study outcomes such as survival rate, patient’s level of function and quality of life (“Intermacs,” 2017).

Fig. 1 Example of Hard Validation
PHTS (Pediatric Heart Transplant Study)

PHTS aims to advance the science and treatment of children both during the listing period and after heart transplantation. The study begins tracking the patients while they are listed and continues to track them periodically with annual followups after transplantation. PHTS began data collection in 1993. Currently 53 hospitals from all over the U.S. participate in PHTS enrolling 8,497 patients, 8,246 listings and 6,250 transplants (“PHTS,” 2017).

Both Intermacs and PHTS collect patient data and have validation processes to maintain data integrity. While developing Intermacs and PHTS we adopted contrasting approaches to validation. Following is the description about the validation processes.

Hard validation

We term the validation process in place for Intermacs as hard validation. In this type of validation, users are blocked from entering data when they make a mistake that affects integrity of the data. Also, in conjunction with hard validation, some of the critical data fields are “locked”, or made “read-only” once the user submits the form (Fig. 1). Assistance from an administrator is required in order to allow the user to make a correction and continue with data entry. We rewrote Intermacs WBDE from scratch in 2014 in order to modernize the user interface and improve user experience. During the rewrite we implemented hard validation for the following reasons: 1) matching the processes followed in the old WBDE; 2) following the preference of administrators to maintain data quality by minimizing errors; and 3) based on assumption that users will make very few mistakes while entering important data for Intermacs. The assumption about user mistake proved to be wrong since we processed close to 1200 tickets to resolve erroneous data entry in Intermacs.

Soft validation

Learning from the experience while developing and maintaining Intermacs, we decided to improvise validation process to be more user friendly. We term this type of validation as soft validation (Fig. 2). In this approach, we removed the locks placed on critical fields and therefore allow users to correct their own mistakes. This approach was successful in reducing the administrative burden involved in making corrections. Due to soft validation, we process a minimal number of data tickets since launch in 2015.

In this paper we hypothesize that soft validation is more effective than hard validation both in terms of improving the user experience and reducing the administrative burden. In our opinion, the drawbacks of hard validation in terms of usability are: 1) not allowing the users to undo incorrect actions and fix their own mistakes, even though they are capable of doing so; 2) requiring time-consuming administrator intervention to rectify the mistakes; and 3) users not being able to continue data entry until the mistakes are resolved.

Fig. 2 Example of Soft Validation
RELATED WORK

In this section, we present a review of the literature related to the topics covered in this paper. We begin by defining HIS and usability from the literature and summarize the findings of research studies under the topics of: 1) errors in HIS; and 2) measuring usability in HIS. Although the literature reviewed in this section is related to Electronic Health Records (EHR) systems, they still apply to patient registry software since both systems collect information about the patient from health care organizations.

Health Information Systems (HIS)

Health information system is defined as an “inclusion of computer-based components which are used by healthcare professionals or the patients themselves in the context of inpatient or outpatient patient care to process patient-related data, information, or knowledge” (E. Ammenwerth, Keizer, & others, 2005). In this paper, we use HIS as a general term that can include a medical registry, electronic health record (EHR) systems, and clinical information systems.

Usability

As defined by Nielsen “Usability answers the question on how well users can use an information system’s functionality, concerning learnability, efficiency of use, memorability, errors and satisfaction.” (Nielsen, 1994). While Nielsen argues for usability in all systems, his conclusions are particularly relevant to HIS due to the need for high-quality data (A. W. Kushniruk, Triola, Borycki, Stein, & Kannry, 2005).

Errors in HIS

In a literature review (Ash, 2003) authors analyzed errors in HIS from qualitative research studies conducted in the United States, the Netherlands, and Australia. These studies discuss several scenarios where health information systems fostered errors instead of reducing them. The review focused on silent errors that arise from a mismatch between functionality of HIS and real-life demands of healthcare work. These errors are not easily found by technical analysis of HIS and can only emerge when the technical system is integrated into a working organization. Further, they can vary from one organization to the other.

The review interprets the nature of silent errors from various social science perspectives such as information science, sociology, and cognitive science. The study categorizes errors into two categories: 1) errors while entering and retrieving information; and 2) errors in communication and coordination process that HIS are utilized to support. The authors suggest that silent errors can be addressed by improvements in education, design, implementation, and research.

In another study (A. W. Kushniruk & Patel, 2004) authors conducted usability testing to evaluate a hand-held prescription writing application. In this study, authors rename silent errors to “technology induced” errors. Ten physicians were asked to perform a series of tasks to enter prescriptions from medication lists into the application. Participant’s interactions with the application were audio and video recorded. These recordings were analyzed in two phases: 1) By usability experts to identify usability problems and 2) By medical experts to identify user errors while entering prescriptions. Results indicated several types of usability problems, most of them related to interface design issues. The analysis also found that some of the usability problems were closely associated with occurrence of specific types of user errors in prescribing medications. The authors conclude by discussing implications of their findings to identify and predict technology-induced errors in order to improve the safety of HIS.

Another research study (Elske Ammenwerth, Gräber, Herrmann, Bürkle, & König, 2003) discussed the problems during evaluation of HIS (called in this study as IT in healthcare), and propose possible solutions to these problems. These problem areas, along with consequences and possible solutions, are presented based on the experience of authors and literature review. The study concludes by providing general recommendations including a broadly accepted framework for evaluation of HIS to address these problems.

A report published in 2015 by Patient Safety Organization (PSO) of ECRI (Emergency Care Research Institute) identified that data integrity failures from incorrect or missing data in records stored in HIS as one of the top patient safety concerns for healthcare organizations (“Wrong-Record, Wrong-Data Errors with Health IT Systems,” 2015).

Measuring Usability in HIS

Nielsen developed a methodology for heuristic evaluation (Nielsen, 1994) (Nielsen, 1995) based on general design principles such as visibility of system status, user control and freedom, and consistency. The following heuristics (good design practices) were used to conduct heuristic evaluation of a wide range of information systems.

Heuristic 1: Visibility of system status.
Heuristic 2: Match the system to the real world.
Heuristic 3: User control and freedom
Heuristic 4: Consistency and standards
Heuristic 5: Error prevention
Heuristic 6: Support recognition rather than recall
Heuristic 7: Flexibility and efficiency of use
Heuristic 8: Aesthetic and minimalist design
Heuristic 9: Help users recognize, diagnose, and recover from errors
Heuristic 10: Help and documentation

This heuristic evaluation is more suitable for HIS (A. W. Kushniruk & Patel, 2004) (Kushniruk, Patel, & Cimino, 1997). Heuristic 3 and Heuristic 9 are applicable to the topic of discussion in this paper. These heuristics are discussed below.
Heuristic 3: User control and freedom

This heuristic emphasizes that user should feel that they are in control of the system and not the reverse. Main guidelines are to: 1) providing clear way to exit from user actions; 2) support undo and redo; and 3) prevent users from performing irreversible actions.

Intermacs clearly violates this heuristic since users do not feel in control of the system especially when they make an error (Fig. 1). Undo and redo functionality is not supported and users have to go through a manual process to resolve the errors and undo their mistakes. PHTS on the other hand implements this heuristic well (Fig. 2) by supporting undo and empowering the users to recover from mistakes by making changes themselves.

Heuristic 9: Help users recognize, diagnose, and recover from errors.

This heuristic suggests giving clear and easy to understand error messages including the details about how to recover from the error. Intermacs violates this heuristic as well by displaying cryptic error messages such as “Invalid Error” (Fig. 1) that include no details about how to recover from the error. PHTS implements this heuristic well by giving gentle warning messages that do not block the users and provide constructive information about recovering from the errors (Fig. 2).

Kushniruk and Patel (A. W. Kushniruk & Patel, 2004) describe the various cognitive usability engineering methods used for the evaluation of HIS. The study explains the drawbacks of traditional summative evaluation methods such as questionnaires, retro-interviews, and focus groups that are generally used to evaluate completed systems. The study also explains the advantages of modern formative evaluation methods such as usability tests that are based on cognition and rapid prototyping, used to evaluate systems in all phases from project planning to maintenance. The study then describes usability evaluation methods and usability inspection methods. Usability evaluation methods are conducted with a sample of the user population while performing a sample of tasks in the system. Different phases used in performing usability evaluation are described along with a case study for illustration. Usability inspection methods on the other hand are based on evaluation of information systems conducted by a trained usability expert. The authors pick two types of usability inspection methods viz. heuristic evaluation and cognitive walkthroughs as more suitable for evaluation of HIS. These two inspection methods are described in detail along with examples.

In spite of strong evidence from literature suggesting that acceptance and adoption of health care system relies on the degree of system usability (Lærum, Ellingsen, & Faxvaag, 2001) (A. Kushniruk, 2002), majority of HIS architects and programmers tend to consider usability as an afterthought. This is demonstrated in a literature review (Peute, Spithoven, Bakker, & Jaspers, 2008) of usability studies on HIS that found that 73% of the studies performed usability evaluations on working systems and only the other 23% of them conducted usability studies as part of system development cycle.

DATA ANALYSIS

Comparison of Timelines of Data Correction

In any instance of incorrect data entry, a timeline of events occurs in a defined sequence. First, the user enters a value, and the value is believed to be correct. In fact, it is wrong, but the user does not realize it yet. Then, at some indeterminate time after initial entry, the mistake is recognized, and user wants to correct the mistake. At this point, the user can take one of two divergent paths, based on how the system implements locks and validations on data entry fields. In the case of soft validation in which no fields are locked by the system, the user simply opens the relevant form, changes the data to the correct value, and continues working. In the case of hard validation, which prevents erroneous entry and locks the entry points, the user cannot change the data autonomously, but must instead contact support staff and request the change. The change request triggers the administrative process of reviewing the mistake, evaluating the need for the change, documenting and communicating the change, then of course making the change through alternate channels (rather than the nominal case of changing it directly in the system). Fig. 3 illustrates a representation of the timeline of these paths.
It is important to note that in the case of hard validation with locking, the system locks only the most important fields, such as Date of Device Implantation in Intermacs. These fields carry the most dependencies, meaning that a significant number of subsequent data points (and their own validation logic) depends on the value of the source field. This is important because if the source field is incorrect, the user’s work cannot continue.

In our analysis, we compared change requests in Intermacs that originated from users and were related to the Date of Device Implantation, on which multiple forms depend for scheduling and validation, to changes in PHTS to the Transplant Date, which is also a key field on which multiple other data elements depend. In both cases, if the date saved for these fields is incorrect, work cannot continue by the user.

The ideal measurement of effectiveness of each approach to validation would be to compare the length of time from the moment in which the user realizes the error to the moment in which the correction is complete and work can continue. Unfortunately, we cannot document the point in time that the user realizes the error. That moment in time is unknowable and fleeting. The time points we can measure accurately are as follows:

- The time of first data entry
- The time of correction of the data
- The time a ticket was created to make a correction
- The time the ticket was completed and closed

In the case of PHTS, we measured the time to make a correction as the difference from the point of initial, validated data entry to the time of most recent correction. For example, if the first entry time for a Transplant Date was March 3, 2017 at 9:38am, then a correction was made on March 4, 2017 at 8:13am, then another correction was made on March 5, 2017 at 9:48am, then the time to make the correction was recorded as 2 days, 10 minutes – the time from the initial entry to the time of the latest correction. Using the first entry time as the starting point of the timeline will underestimate the time to make a correction.

In the Intermacs system, we measured the time to make a correction as the difference from the creation time of the change request ticket to the time of completion of the same ticket. For instance, if the change request ticket to modify an Implant Date was created on March 3, 2017 9:38am, and the ticket was marked as complete on March 17, 2017 11:38am, then the time to make the correction was recorded as 14 days, 20 minutes. Important considerations about this method are as follows:

- The time of ticket creation forms the beginning of the timeline because this is closest to the unknowable moment that the user realizes an error occurred. Using the ticket creation date will underestimate the time to make a correction.
- The time of completion of the ticket forms the endpoint of the timeline, and not the time of completion of the change itself. The ticket completion time is the closest recorded time to when the user can continue working, since the end user relies on notification from administrative staff. Even if the data is in fact already correct before this moment, the end user cannot continue working because he/she is not aware that the change has been made.

The geometric mean for time to completion in PHTS was 26.675 days, or 640 hours and 19 minutes, with a sample size of 41. The median time to completion in PHTS is 20 hours and 45 minutes. If the sample size of corrections is reduced to only include changes that happen with 24 hours (i.e. on the same day as the initial point of entry), the average drops to 79 minutes and 32 seconds across 20 corrections. This reduction is sensible considering that changes that take more than 24 hours to complete are likely driven by external factors, such as reports or peer review, rather than direct recognition or awareness of the error.

The geometric mean for time to completion in Intermacs was 60.81 days with a sample size of 58. The median time to completion is 32.09 days. The difference between the mean and the median shows a much slimmer margin of variation in completion times. There is no logical method to segment the data further.

Time to completion had a highly-skewed distribution (Fig. 5). Therefore, the Inter Quartile Range (IQR) representing 25th to 75th percentile was used to compare the aggregate data and the non-parametric Mann-Whitney U test was used to test the statistical significance of the completion times between the registries. The median (IQR) for Intermacs tickets is 32 days (20 days and 103 days). The median (IQR) for PHTS tickets was 0.9 days (4 minutes and 34.4 days). This difference is statistically significant (p < 0.0005, Mann-Whitney U test). This can be visualized by using the cumulative distribution function (Fig. 4) of ticket completion time within each registry. The cumulative distribution function shows that close to 60% of changes in PHTS were completed in less than a day, where as it took close to 30 days to complete 60% of tickets in Intermacs. The box plot (Fig. 5) shows that the number of days needed to complete tickets in Intermacs are clustered from 25 to 100 where as the number of days needed to make changes in PHTS are clustered from 0 to 25. We can observe that both Intermacs and PHTS have a few outliers that do not have a major effect on the overall analysis.

CONCLUSION AND FUTURE WORK

We started this paper by introducing two medical registries and their contrasting validation processes. We compared the time to ticket completion for correcting user errors in Intermacs to time taken for users to fix their own errors in PHTS. By analyzing the data and comparing the aggregates we deduced based on the principles of usability engineering that soft validation used in PHTS is more usable than hard validation used in Intermacs. Currently, we plan to still continue to use hard validation in Intermacs for two main reasons: 1) maintain the data quality that is important for the research and regulatory compliance; and 2) we do not have
definitive data from the users to support that they prefer soft validation over hard validation. But, the analysis conducted in this paper provides valuable evidence in favor of soft validation. We hope that this analysis will be useful for other groups considering different validation processes.

In future, we plan to conduct a survey to record the user experience on both Intermacs and PHTS and analyze the responses in a follow-up study. Since we have hospitals that use both Intermacs and PHTS the opinion of users from those sites will be highly beneficial in doing the direct comparison and get focused responses. We also plan to do a study comparing data quality in both Intermacs and PHTS to compare the effects of validation processes on data quality. We plan to use the results of both of these studies to make iterative changes to Intermacs to improve usability and guide the development of new registry systems in future.

ACKNOWLEDGEMENT

Suggestions from our colleague John Pennington saved us a lot of time while formatting the paper. We appreciate his help and guidance.

Fig. 4 Cumulative distribution function of time (days)
REFERENCES


Fig. 5 Box plot
A MULTI-TIERED DISTRIBUTION STRATEGY FOR IMMERSIVE VIRTUAL REALITY CONTENT IN BLENDED DELIVERY INTERPROFESSIONAL HEALTH CARE EDUCATION

David E. Robbins, John A. Galdo, and Sarah Waller
College of Health Sciences, Samford University
{drobbins@samford.edu, jgaldo@samford.edu, swaller@samford.edu}

ABSTRACT

Although virtual reality (VR) promises to enhance health care education, delivering VR content to students presents significant challenges in a blended online/on-site delivery program. Responsive to this challenge, we present a multi-tiered approach for delivering VR content across a spectrum of immersion and accessibility. Students can self-select within this immersion spectrum based on their immediate objectives and the availability of VR hardware.

INTRODUCTION

Using virtual reality (VR), students of health care professions can experience the world through the senses of their patients, their colleagues, and their interprofessional team members. We previously presented a roadmap for integrating these experiences into interprofessional health care education (Robbins, Neumeier, & Galdo, 2016). During the implementation of this roadmap, however, we discovered significant challenges to the dissemination of VR educational content integrated with classroom activities.

The most immersive virtual reality technologies require significant investments in electronic hardware, such as a paired VR headset and a high-powered computer. This equipment can pose a significant cost burden to students, with a full system costing around $2000.00 USD in 2017. Although we anticipate costs will continue to decrease as new technology becomes available, we simultaneously anticipate more immersive technologies to become available similar or higher price points.

Purchasing multiple sets of highly immersive VR systems is not feasible in an on-site education context, where they can be shared as in a lab setting. However, as more and more students take courses online, their use can present an obstacle to online learners who may be incapable of traveling to the equipment.

Since March of 2015, the online video platform YouTube has supported the distribution of 360-degree video over the web. Playback of 360-degree video in a web-browser requires the use of buttons on the video player to “look around” within the video – hardly immersive. However, when played back on a smartphone with a gyroscopic sensor, looking around is achieved by moving the smartphone. This more natural interface enhances the immersion the VR viewing significantly. Further enhancement can be achieved by pairing the smartphone with low-cost virtual reality accessories.

To provide the most immersive VR educational experiences possible, we developed a multi-tiered distribution strategy for VR content. This multi-tiered approach provides a spectrum of immersion that students can move within according to their immediate objectives and their modality of participation in the course.

METHODS

Our multi-tiered approach to producing and delivering virtual reality experiences (VRE) involves recording health care scenarios using high-resolution, 360-degree cameras; editing these recordings using commercial-off-the-shelf software; and finally distributing the recordings through multiple channels with varying levels of fidelity and immersion.

In our pilot we recorded 7 VRE scenarios using a Nikon Keymission 360 camera. These recordings were made from the patient perspective. That is, the camera was placed approximately where the patient’s head would be, such that a viewer of the scenario will see and hear things as though they were in the patients position.

The VRE scenarios were edited for length and transitions using Adobe Premiere Pro and Apple iMovie. Following editing, scenarios were injected with spatial metadata using Google Spatial Media Metadata Injector (https://github.com/google/spatial-media). This injection is required to enable playback of the scenarios as 360-degree media, where the viewer can alter their viewpoint or “look around.”

VRE scenarios were delivered to students either through upload and playback through YouTube or playback on an Alienware 17 high-performance laptop equipped with an Oculus Rift VR headset. When delivered through YouTube, students viewed the experiences in one or more formats depending on preferences and hardware availability. When viewed through the YouTube website, students manipulated the viewport using integrated graphical user interface controls. When viewed on a gyroscopically enabled smartphone, students “looked around” by moving the smartphone like a small window into the world of the simulation. Finally, by using a smartphone in conjunction with a VR accessory such as Google Cardboard or Samsung Gear VR, students experienced the simulation with nearly the same fidelity as a high-powered VR headset.
All VR scenarios were posted on the Health Care 360: Seeing is Believing YouTube channel, available freely online at https://www.youtube.com/channel/UCPU0wll4IuknAn7J4oDeTMQ.

RESULTS

Using a multi-tiered distribution approach has successfully enabled us to deliver immersive VR education content to both on-ground and online students in Samford University’s College of Health Sciences. The 7 developed virtual reality scenarios have together been viewed 424 times through August 2017. Additionally, the virtual reality scenarios have been delivered to multiple on-ground students using Oculus Rift virtual reality headsets.

DISCUSSION

We believe our multi-tiered distribution strategy has proven to be a successful technique for delivering immersive educational content to students in both on-site, online, and blended learning settings. As with any success, our experiences have exposed multiple tradeoffs and challenges associated with implementing virtual reality experiences as a component of health care education.

In most health care settings, written material forms a key element of the visual environment. From medication labels to instructions on a discharge sheet, patients are expected to be able to see and read written information. Unfortunately, the Nikon Keymission 360 cameras proved incapable to reliably capturing the environment with sufficient resolution to allow reading of written materials. Although their resolution of 3840x2160 pixels (4K UHD) is more than adequate for this purpose in a standard rectangular viewing format, it is simply not enough pixels when viewed in a 360 degree format. To correct this, we’ve ordered a 7680x3840 pixels (8K) capable camera – the Insta360 Pro. Acquiring high-resolution cameras at reasonable cost remains a significant challenge.

Compared to the preparation of other educational video content, such as lecture videos and narrated slide decks, 360 degree video simulations requires significant post-production effort. We expect this will improve as the software tooling available to produce this type of content increases in capability and automation.

With currently available VR hardware and software, educators and students face tradeoffs in the degree of immersion available, cost of required equipment, and the reach and availability of VRE. The most immersive equipment available, high-powered VR headsets paired with powerful computers, generally cost $2,000 to $3,000 – a significant cost burden for a student. This type of equipment can be provided in limited numbers in a classroom or laboratory setting, but each set of equipment will only be usable by one student at a time. In contrast, making use of student smartphones paired with inexpensive VR accessories provides a slightly less immersive experience, but can be provided to students in larger numbers or purchased by students for personal use. Paired with distribution via YouTube, this technique allows students to view VRE provided as a component of online or blended-online classes.

The immersion spectrum also provides students with a wide range of options for engaging in VRE. During an initial experience of a VRE, a deeply immersive setting is preferable to maximize impact. After this experience, the availability of

![Figure 1](https://www.youtube.com/channel/UCPU0wll4IuknAn7J4oDeTMQ)

Figure 1. A multi-tiered distribution spectrum for virtual reality content in blended delivery interprofessional health care education. In this figure a selection of delivery technologies are arranged, from left to right, according to their degree of immersion, cost, and accessibility. Also shown is the learning context (on-site or online), the playback device, and whether the VR file is played directly or viewed using YouTube’s 360-degree video player. Notably, personal computers provide both the most immersive and least-immersive experiences.
less immersive formats enables review and reflection with minimal additional effort.

As a relatively new technology in interprofessional health education, the use of VR in online and blended learning settings needs further assessment of its efficacy in achieving student outcomes.

CONCLUSION

By delivering VR content using a multi-tiered dissemination strategy, health care educators can provide maximally immersive experiences to their students. Constraints on immersion arise from the cost of VR gear and the students learning modality – online or on-site. With VR content provided in tiers, students can select the degree of immersion based upon their immediate objectives (e.g. review) and the availability of VR technology.

ACKNOWLEDGMENT

This work was supported by an Innovative Technology Grant from Samford University.

REFERENCES

 USING VIRTUAL REALITY TO TEACH INTERPROFESSIONAL EDUCATION

John A. Galdo, Ashley Parish, Rebecca Warr, David Robbins
College of Health Science, Samford University

(jgaldo@samford.edu, aparish@samford.edu, rjwarr@samford.edu, drobbins@samford.edu)

ABSTRACT

Interprofessional education is now a required component of health professional education. However, various barriers, like physical availability on non-medical higher education campuses of different healthcare workers creates sometimes insurmountable obstacles for education. Virtual reality is a low-cost alternative teaching method, similar to simulation, that allows various healthcare providers to see the eyes of other professions.

I. INTRODUCTION

Interprofessional education (IPE) is the practice by which various disciplines of healthcare providers, pharmacists, nurses, physical therapists, dentists, physicians, learn and train together to provide team-based, patient care. This entry level training helps prepare each member of the team to understand the roles and responsibilities of each provider. The most common definition of IPE adapted from the Centre for the Advancement of Interprofessional Education in the United Kingdom and the World Health Organization.

Additionally, the WHO Study Group on Interprofessional Education and Collaborative Practice studies the impact of IPE on educational and health policy. Ultimately, IPE can improve healthcare delivery. However, IPE is not fully implemented due to the barriers, like physical availability, associated with teaching IPE.

Recently, some medical professions, like pharmacy, have implemented IPE requirements into accreditation for US-based schools. Yet, not all accredited pharmacy schools are integrated within larger medical centers, which sustains barriers to IPE. Additionally, many health professionals still practice in silos – namely, dentistry, community pharmacy, physical therapy, and others. The practice-based silos prevent IPE during externship opportunities. Therefore, schools are required to find innovative ways to bring different disciplines into the classroom to foster IPE.

One such innovation is the use of virtual reality (VR). Previous work has shown a roadmap for increasing provider empathy through virtual reality. Part of the framework describe the use of virtual reality to help teach IPE; this article takes a deeper look into three different healthcare providers and the use of VR in IPE education.

The Healthcare Provider Perspective

Role of Physical Therapy in Healthcare

The traditional role of physical therapy in healthcare is to examine, diagnose, and treat movement dysfunction secondary to medical conditions. Movement dysfunction in turn, limits an individual’s functional ability. Physical therapists work with individuals throughout the life span including newborn to geriatric populations. Physical therapy has a growing part in prevention of disability through health promotion and wellness in addition to the traditional role. Due to the diversity of practice type, physical therapists work in a variety of settings including hospitals, sports facilities, work settings, nursing homes, home health and outpatient settings.

Impact of Physical Therapy

Because physical therapists provide treatment to improve functional activity and not necessarily treat the primary medical conditions, it is easy to fall into practice silos. The silo model of healthcare does not support communication or collaborative practice, which contributes to breakdown between disciplines as well as optimal health care for the patient. Though physical therapists are rehabilitation experts with a specialized vocabulary, it is imperative to work effectively with other professions to provide coordination of care. In regards to physical therapists, they frequently collaborative with other healthcare professions, referral sources, durable medical equipment suppliers, and well as community health and wellness programs.

Physical Therapist’s Role for Other Healthcare Providers

Critically ill patients are living longer, increasing hospital lengths of stay and periods of immobility with the advancement of medical technology and care. Since immobility is directly linked with decline in functional ability, physical therapy has gained a vital role in early mobility in the intensive care unit setting in recent years. Though not standardized practice, physical therapists are now ambulating patients who are mechanically intubated and on extracorporeal membrane oxygenation (ECMO). Research has demonstrated that early mobility is associated with decreased mortality, length of stay, and hospital costs. It is a safe and effective intervention that can have significant impact on functional
outcomes. Early mobility involves a team approach in which all professions involved must work together to communicate patient status, monitor vital signs, and manage equipment.

Individuals must understand the role of each provider in order to communicate effectively about the care of patients. Virtual reality will allow other professions to observe the role of physical therapy. In healthcare, there is overlap in skill sets. Despite this knowledge, time and location can provide challenges with interdisciplinary training. Virtual reality allows cross over in skill sets. For example, virtual reality could be used to teach both nursing and physical therapy students how to transfer a patient from the bed to chair in the hospital setting. Virtual reality would allow for the student learners to be in the same location without the time constraints of simulation. Use of virtual reality has been proven to raise interest and motivation to effectively support skill acquisition and transfer. Use of virtual reality to demonstrate the role of other healthcare professions will improve communication and collaboration, which favorably impact patient outcomes.

**Role of the Nurse in Healthcare**

“Once viewed as subservient and subordinate, nurses are now serving as full and essential partners on interdisciplinary health care teams.” The role of the nurse in health care includes technical skills such as starting intravenous fluids, giving medications, performing sterile procedures, and assessing the client. In addition, nurses must use critical thinking skills to determine the best interventions to use with the patient; they are present to care for the patient twenty-four hours a day, seven days a week while the patient is hospitalized. Nurses serve a variety of roles in healthcare: identify needs and goals, intervene to help the patient meet those goals, administer medications, identify side effects, prevent medication errors, provide for prevention of injury, such as from falls or skin breakdown, and coordinate the care of the patient with other health care professionals. The document from The Campaign for the Future of Nursing states, the nurse is “both the front lines of health care, as well as the backbone of patient treatment. We see nurses as innovators in health care—like pain detection in newborns. We see how their observational skills, advanced knowledge, interventions and compassionate care help patients manage their medical needs.”

**Impact of Nursing**

When the nurse concentrates on managing the patients assigned and completing the tasks and technical skills needed, he or she is working in a silo. The patients may be receiving good technical care but not the quality that could prevent complications and repeated hospitalizations.

The Institute on Medicine, in a 2010 report, stated that nurses improve access to quality care and to lower costs. They also play vital roles in achieving patient-centered care, strengthening primary care service, delivering more care in the community, and providing seamless, coordinated care.

**Nurse’s Role for Other Healthcare Providers**

Nurses not only implement the medical treatment plan, including administering medications as well as other treatments, they evaluate the safety for the patient before implementing the plan. Nurses are constantly assessing and using critical thinking skills to provide safe care for the patient, deciding what interventions are needed and when. The nurse is the person that identifies when the patient’s condition requires immediate attention, calling a health care provider for emergency medication to prevent a further decline; or calling for a rapid response team when resuscitation might be immanently needed.

Nurses can shortend the length of time a patient is in the hospital; because the nurse knows the patient’s goals, the nurse is able to help ensure that treatments by other health professionals are received as needed.

Nurses work with all members of the health care team to help provide quality patient care. More than just performing technical skills to a person, the nurse cares for the patient’s physical, mental, emotional, social, and spiritual needs. The nurse works with other members of the health care team to accomplish an individualized plan to help the patient meet their goals.

**Using Virtual Reality to Explain Nursing**

Virtual Reality can give a glimpse of the nurse as the client is assessed for their problem, intervention provided, medications administered and evaluated for side effects and therapeutic outcomes, and performing sterile procedures in a compassionate manner. The nurse can also be seen identifying the patient’s goal and ensuring that the patient receives treatment from other health care team members to meet those goals.

**Role of the Pharmacist in Healthcare**

Often considered the medication expert, pharmacists practice in all areas of healthcare. Various responsibilities include delivery of direct patient care, self-care in the community setting, optimization of medications, and population health through managed care. The most common practice for pharmacists is in the community setting, usually at a retail location, such as CVS Health or Walgreens. Stereotypically, the pharmacist is viewed as an assembly worker—packaging and preparing medications; however, part of the preparation includes a drug utilization review, which examines the clinical correctness and safety of care.

**Impact of Pharmacy**

One of the main modalities for the treatment of medical conditions is medications. For example, a 2010 CDC
report stated at 37 percent of older Americans use five or more prescription medications per month. This trend is only continuing; the report also stated that medication use has doubled from 1999 to 2010.18 Succinctly, the role of the pharmacy can have a major impact on day to day management of patients.

**Pharmacist’s Role for Other Healthcare Providers**

Pharmacists are an integral aspect of the healthcare system, yet are at a financial disadvantage. Pharmacists are the only healthcare provider reimbursed for a product, rather than cognition. This misalignment of skill and reimbursement model has limited the impact of a pharmacist in the management of patients. Additionally, community pharmacies, where majority of pharmacists practice, are a silo in healthcare. The community pharmacist has no access to a patient’s medical record and often guesses at the indication for drug therapy. Alternatively, the pharmacist is also the most accessible provider; patients can walk into any pharmacy, sometimes 24/7, to ascertain answers to clinical questions.

**Using Virtual Reality to Explain Pharmacy**

Virtual reality can best help explain the thought process of a pharmacist. More often than not, little consideration is given to a pharmacist’s professional judgement. For instance, a recent Chicago Tribune article described pharmacists missing more than half drug-drug interactions.19 Yet, the study did not take into account all the characteristics of filling the medication, including the aforementioned drug utilization review. The use of virtual reality can help other providers understand the day to day operations of preparing medications and the clinical review of therapy.

**Conclusion**

The use of virtual reality technology is not just limited to the three disciplines mentioned above. Many healthcare professional schools struggle to meet Interprofessional education accreditation requirements, coordinate schedules between institutions, and overcome historical training biases. The expanding role of many healthcare providers, complexity of patient care, and the need to truly understand patient-centered care have necessitated innovative methods to education. Virtual reality is the first in many steps to produce well-rounded practitioners.

**ACKNOWLEDGMENT**

This work was supported by an Innovative Technology Grant from Samford University.

**REFERENCES**


16. UT-InterProfessional Education, The University of Texas at Austin The Nurse’s Role 
http://www.healthipe.org/healthcare-roles/nursing

17. Role of Pharmacist in Healthcare. 


ADDING ADVANCED FEATURES TO A DATABASE COURSEWARE

Mario Guimaraes, Professor & Chair of CS
Saint Martin's University
mguimaraes@stmartin.edu

ABSTRACT

This project describes the redesign and advanced features that are being developed based on an existing Database Courseware (adbc.kennesaw.edu). It describes the major functions of the existing software, redesign requirements and new modules under development. Unique features described include Data Warehouse, Data Mining, MongoDB and a Database Video Game.

INTRODUCTION

The database courseware being redesigned is based on a Database Project funded by two NSF grants (0089412 and 0717707). This software contains four modules (Database Design, SQL, Transactions, Database Security). This software allows code execution and visualization in a step by step sequence at the student’s individual pace. Figure 1 shows a screen shot of this software. In the example, the sub-module Database Design – Scenarios to ER is highlighted, where a student is given a scenario obtained from an interview and selects the corresponding E-R diagram.

This software has gone through many different type of evaluations. Among the strong points frequently identified in evaluations are the opportunity for users to associate existing knowledge with new knowledge, through multiple windows displayed in parallel, a self-pace instruction and the high degree of interactivity. Justification for this type of software can be found at (Shackelford, 2006), (Guimaraes, 2006, 2009) and (Dietrich, 2016)

The Database Design Module was developed in Flash and it is still accessible by most web-browsers. The other three modules were developed as java applets and most web browsers have restricted access to this part for security reasons. Therefore, there was a need to re-design the entire software as an application instead of applets. For standardization, all four modules are being redesigned as an application.

DATA WAREHOUSE

This module was chosen due to the high demand of data warehouse jobs and the shortage of training resources. We chose a stock market domain for example because it has several interesting characteristics for teaching. Stock market data can be found and downloaded for free from many different sites such as http://finance.yahoo.com/q/hp?s=YHOO and http://www.quantshare.com/sa-43-10-ways-to-download-historical-stock-quotes-data-for-free.

Furthermore, in these applications there is a need to perform a daily Extract-Transform-Load (ETL), as well as an intra-day analysis. This led us to create two star models. One for intra-day analysis with a Date and a Time Dimension and one to analyze stocks between days, requiring just a Date Dimension. Figure 2 shows the star schema with only analysis between days. We are simulating the import of multiple stocks from multiple stock markets. This is an excellent way to demonstrate a need to automate the ETL process through tools such as Microsoft’s SQL Server Integration Services (SSIS).
Two other features that are being illustrated by the software are ETL of a dimension table (Figure 3) and ETL of a fact table (Figure 4). Note that before the ETL of the dimension table, records are removed for a Full Extract. Meanwhile, the ETL of the fact table is done incrementally. These animations will allow the user to visualize each step of the ETL flowchart and explanation windows as the flowchart is executed.

Figure 2 – Star Schema for a Stock Market Data Warehouse

Figure 3 - ETL of a Dimension Table

Figure 4 - ETL of a Fact Table

DATA MINING

In the data mining model, we are developing animations to introduce data mining concepts using SQL Server’s data mining tools. The initial design, includes features illustrating Data Mining with Decision Trees and Time Series Models. Decision Trees was chosen due to the relative ease that one can learn and use them. Figure 5 shows the Decision Tree option chosen from among 9 Data Mining Models. Figure 6 shows the Input Attributes and the Attribute to be predicted. Figure 7 displays a decision tree being generated. The darker colors shows the most likely paths. The software will allow the user to move one step forward or backwards at all times. It focuses not only on how to use the Data Mining tool, but also the advantage and drawback of each Data Mining Model. In future versions, the courseware will include the same example displayed through two different modules for side by side in equivalent windows.
The Time Series Module was chosen for forecasting. The high volume of applications that require time series information (financial institutions, weather forecast, e.g.) and the public availability of this data justified this decision.

**NOSQL AND MONGO DB**

This first sub-module presents basic MongoDB (CRUD) commands, allow users to execute them and compare to the equivalent SQL commands (Figure 8). In the example, one MongoDB command is highlighted and executed as well as the corresponding Relational command (in another window). The MongoDB design sub-module introduces mongo objects comparing them to the equivalent Relational objects (Collection to Table, Document to Row, e.g.). This design sub-module also compares linking to embedding documents, with examples where it makes more sense to link and examples that make more sense to embed. The indexing module shows visualizations with examples of the balanced tree (default index), the text index (or inverted list index) and spatial indexes. The last sub-module illustrates how sharding (horizontal partitioning) in MongoDB.

```sql
SELECT cust_id, 
FROM orders 
GROUP BY cust_id 
HAVING count(*) > 1 
```
VIDEO GAME MODULE

The videogame is a reinforcement presenting topics and challenges of the other modules in a project base style. At level 1, the player makes decisions while moving through Obtaining Requirements, Constructing E-R Diagrams, Creating Logical Schema, Creating Tables, and Inserting, Updating, and Querying Data.

At level 2, the player will tables, alter columns of existing tables and create integrity constraints.

At level 3, the database is under attack and the player applies Database Security techniques.

At level 4, the player creates a stock market Data Warehouse by extracting data from yahoo finance.

Video Games provide active learning. Players benefit from multiple well known pedagogical features: learning within a familiar context, well ordered exercises from easy to complex, consequence of failure lowered by ability to re-play, performance viewed at all times, decision making with immediate consequences, and an assessment component (score). Designing the interface was one of the major challenges. Decisions can’t be simple multiple choice solution. Meanwhile, it can’t be expected that the user will input solutions through the keyboard. The solution was a library style interface containing all possible answers. The library is arranged by topic. The user will navigate through the library and retrieve the answer. It is like a multiple the multiple choice, but instead of four options, there will be more than a hundred arranged in units of approximately seven options per room.

ACKNOWLEDGMENT

We thank the National Science Foundation for funding the development of the original project.

REFERENCES


INTERNATIONAL CONVERGENCE ON THE
TRANSDISCIPLINARY RESOURCE CORRELATION PLATFORM

U. John Tanik, Sang C.Suh, Abdullah Arslan, Kaoning Hu, Hemanth Bandi
Department of Computer Science and Information systems
Texas A&M University Commerce
John.tanik@tamu.edu, Sang.Suh@tamu.edu, Abdullah.Arslan@leomail.edu,
Kaoning.Hu@tamu.edu, hbandi@leomail.tamuc.edu

ABSTRACT:

A novel approach is introduced to efficiently bridge the current disconnect between graduate students and highly successful inter/trans-disciplinary researchers and employers in universities, government, and industry. A new graduate education model can be piloted at Texas A&M – University Commerce (TAMUC) with the new Computational Science program that now includes tracks in Computational Biology, Linguistics, and Business Analytics. To stimulate growth, this semester TAMUC CSIS department has already improved the degree plan and introduced more well-defined tracks and relevant courses to encourage its growth with more options forthcoming. Innovation can be accomplished by developing a virtual platform that functions as a transdisciplinary resource correlation service that can provide all types of synergistic thesis support for students, faculty, and industry. This model will be based on a framework that provides a virtual bridge to a global network of transdisciplinary researchers, starting with members in the Society for Design and Process Science (SDPS), a transdisciplinary organization with members including Nobel Laureates, reputable international researchers, and industrial sponsors who can assist with job placement and internships. Since a common framework was needed for all SDPS student chapters accessing a resource portal, the concept for a Transdisciplinary Resource Correlation Platform (TRCP) has materialized.

Keywords: Convergence, Transdisciplinary, Correlation, Platform, Computational.

INTRODUCTION

A novel approach is introduced to efficiently bridge the current disconnect between graduate students and highly successful inter/trans-disciplinary researchers and employers in universities, government, and industry. A new graduate education model can be piloted at Texas A&M – University Commerce (TAMUC) with the new Computational Science program that now includes tracks in Computational Biology, Linguistics, and Business Analytics. To stimulate growth, this semester TAMUC CSIS department has already improved the degree plan and introduced more well-defined tracks and relevant courses to encourage its growth with more options forthcoming. Innovation can be accomplished by developing a virtual platform that functions as a transdisciplinary resource correlation service that can provide all types of synergistic thesis support for students, faculty, and industry. This model will be based on a framework that provides a virtual bridge to a global network of transdisciplinary researchers, starting with members in the Society for Design and Process Science (SDPS), a transdisciplinary organization with members including Nobel Laureates, reputable international researchers, and industrial sponsors who can assist with job placement and internships. Since a common framework was needed for all SDPS student chapters accessing a resource portal, the concept for a Transdisciplinary Resource Correlation Platform (TRCP) has materialized.

In a span of three years, the TRCP pilot is expected to boost enrollment, research opportunity, and job placement in Computational Science. The first pilot of the proposed platform framework (TRCP) advancing the TAMUC Computational Science program will be capable of inter/trans-disciplinary research and development initiatives with industry. The gap between senior and graduate researchers will be resolved by accurately orchestrating a multitude of services provided through a robust online portal and community. SDPS celebrated its 20th anniversary in 2015 and TAMUC CSCI Department established the university sponsored Design and Process Laboratory (DPSL). Plans were made in Dallas, Texas during the SDPS 2015 conference to develop the DPSL as a center to manage SDPS student chapters at various universities that serve to develop transdisciplinary talent. In Orlando, Florida during the SDPS 2016 conference, discussion was focused on how UAB and TAMUC student chapters could collaborate through innovative means to transfer knowledge from advanced members to next generation. And now, in Birmingham, Alabama the next SDPS 2017 conference is scheduled in November, which will include discussion of a framework that can manifest as a virtual platform that can distribute research services seamlessly to TAMUC, UAB, and UT-Arlington, and other universities which can benefit from the network. The TRCP development includes concentrated effort by many SDPS members to develop a scalable, modular, reconfigurable platform that facilitates thesis research for graduate students based on more than twenty years of experience in transdisciplinary research. The TRCP development enables graduate students from a broad array of universities to utilize the services available on the virtual platform for thesis work and other research support as needed.
OVERVIEW OF TRANSDISCIPLINARY RESEARCH CORRELATION PLATFORM FOR GRADUATE RESEARCH

One of the emerging challenges of the 21st century is teaching research and development techniques to the next generation of graduate research students who have unprecedented computational power available at a moment’s notice anywhere in the world, but do not know how to properly leverage it. Solving this problem is especially important at Texas A&M University-Commerce (TAMUC) for advanced students pursuing thesis option in the Computer Science and Information Systems (CSCI) Department. In fact, there is a widening gap and disconnect between those practicing research in industry and students learning research skills in the university as disciplinary boundaries blur. Students are even more confused with the information explosion generating overlapping disciplines that make the research and development (R&D) job market a fast-moving target upon graduation. Students need more direction on relevant projects that will garner them job prospects where they can apply their computation skills immediately to an employer of choice in any research domain. The question of how to bridge the gap between industry and university interests so students succeed in R&D environments can be answered by the results of this initiative.

The CSCI Department has tackled this issue in 2017 by revamping the Computational Science degree plan as a starting point. However, after the faculty meetings, the director and Co-director of the Computational Science Program recognized the importance to carry the momentum of change and met with the Chair of the Department to develop this initiative with the Society for Design and Process Science (SDPS). At least 21 consultants have agreed to facilitate the development of a research platform called the Transdisciplinary Research Correlation Platform (TRCP) to educate and train the next generation of research students with industry-driven projects for thesis that directly utilize computational science techniques and technology for any domain. This platform has three essential layers, namely presentation, logical, and data layer to manage the I/O, services, and storage functions for research and development with industry (Fig 1). The test pilot can be accomplished over three years with the new course CSCI 575 entitled “Cyber-physical Systems and the Industrial IOT”. As the list of companies grow in the Industrial Internet Consortium, more opportunities abound as R&D projects can be supplied by such networked companies to our platform for student selection.

COMPUTATIONAL SCIENCE UTILIZING TRANSDISCIPLINARY RESEARCH CORRELATION PLATFORM

The Computational Science program degree plan has added three tracks, namely Computational Linguistics, Computational Business Analytics, and Computational Biology. Among the new courses being offered by the growing faculty is CSCI 575 Cyber-physical Systems and Industrial IOT. The pilot for this platform can be evaluated by the industrial advisory board (IAB) that can assist validation of the testbed with guidance provided as needed to the pilot course CSCI 575 (Table 1).

In addition to feedback from the IAB, student teams in courses can develop modules appropriate to the learning objectives for each course. At least 80 students per semester can assist platform module requirements, design, modeling, and even implementation (Table 2). The platform can form a convergence area for SDPS student chapters networked together and managed from the TAMUC Design and Process Science laboratory (DPSL). The student chapters developed their constitutions after the 20th anniversary of SDPS celebrated in 2015 with next generation leadership taking the helm to pursue expansion in the research sphere advancing academic, industrial, and government interests in the United States and abroad. The senior membership began developing general plans to roll out a general transdisciplinary platform that networks all resources together for effective research collaboration, education, and training (Table 3).
Table 2. Courses taught that can assist with platform development

<table>
<thead>
<tr>
<th>Course</th>
<th>Hours</th>
<th>Topic</th>
<th>Students</th>
<th>Teams</th>
<th>Activity as part of class</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSCI 595</td>
<td>3</td>
<td>Research Literature and Techniques</td>
<td>20</td>
<td>4</td>
<td>Development of platform modules guiding students on research project techniques</td>
</tr>
<tr>
<td>CSCI 524</td>
<td>3</td>
<td>Analysis and Design</td>
<td>30</td>
<td>6</td>
<td>Development of platform module requirement specifications and modeling</td>
</tr>
<tr>
<td>CSCI 530</td>
<td>3</td>
<td>Operating Systems</td>
<td>30</td>
<td>6</td>
<td>Development of platform module processes that function as CPU analogy</td>
</tr>
<tr>
<td>CSCI 575</td>
<td>3</td>
<td>Cyber-physical Systems and Industrial IOT</td>
<td>TBD</td>
<td>TBD</td>
<td>Development of platform modules that guide students on IOT product development with Wolfram testbed using industry projects</td>
</tr>
</tbody>
</table>

Table 3. SDPS Student Chapters networked together contributing to the platform

<table>
<thead>
<tr>
<th>SDPS student chapter</th>
<th>Development goals for platform starting with first year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texas A&amp;M-Commerce (TAMUC)</td>
<td>Officers and members are in charge of STEM projects for involving supplies acquisition from vendors for IOT product development with computational software such a Wolfram.</td>
</tr>
<tr>
<td>UAB</td>
<td>Officers and members are in charge of the training MS and PhD students for R&amp;D education and training supported by university and industry.</td>
</tr>
<tr>
<td>UT-Arlington</td>
<td>Officers and members will formulate constitution and hold their first meeting to coordinate R&amp;D education and training with UAB and TAMUC from common platform.</td>
</tr>
</tbody>
</table>

CSCI 575 CPS and Industrial IOT is an ideal vehicle to test the pilot that utilizes the TRCP services with plenty of opportunity for students to showcase their computational skills when conducting R&D thesis work for the course that is directly relevant to industry that provides research projects ideas.

Consultants from SDPS were invited to assist in platform development and 21 confirmed their interest (Table 4). These consultants will form teams (Table 5) to develop modules that facilitate the education and training of R&D skills relevant to employers with industry-driven projects supplied to the platform directly. These research projects posted by member companies will train students on relevant computational software, if selected by the student and approved by faculty.

The proposed platform will follow a general software design lifecycle defined by the Rational Unified Process (RUP) having four distinct phases (Fig 2) for platform development, namely inception, elaboration, construction, and transition (with currently unknown activity levels expected per layer). In year 1, planning and prototype development is initialized to ascertain specific requirement specifications to develop the prototype. In year 2, testing and verification of aspects of platform operations is accomplished. In year 3, piloting and validation of CSCI 575 utilizing all the functionality of the platform (TRCP) module services is accomplished. Formative and summative reporting can be conducted as needed (Section 5).

Fig2. Transdisciplinary Resource Correlation Platform Activity Diagram based on RUP

Processes for modular platform activity with stakeholders will follow standards such as IEEE 830, 1016, 1058 for software requirement specifications (SRS), software design description (SDD), and project management plan (PMP), respectively, for coordinated work by all stakeholders, including administration, faculty, students, consultants, industry, and investigators (Table 6).
Table 4. Consultants mapped to platform modules for computational science R&D education

<table>
<thead>
<tr>
<th>Consultant</th>
<th>Platform module</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wolfram Mobile Application Development Environment (A)</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>Wolfram Mathematica Functions with Wolfram Alpha (B)</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>Wolfram Cloud Programming (C)</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
</tr>
<tr>
<td>13</td>
<td>Wolfram Discovery Platform (D)</td>
</tr>
<tr>
<td>14</td>
<td>D</td>
</tr>
<tr>
<td>15</td>
<td>Wolfram Data Science Platform (E)</td>
</tr>
<tr>
<td>16</td>
<td>E</td>
</tr>
<tr>
<td>17</td>
<td>Wolfram System Modeler (F)</td>
</tr>
<tr>
<td>18</td>
<td>F</td>
</tr>
<tr>
<td>19</td>
<td>Wolfram Connected Device Framework (G)</td>
</tr>
<tr>
<td>20</td>
<td>G</td>
</tr>
<tr>
<td>21</td>
<td>G</td>
</tr>
</tbody>
</table>

Table 5. Transdisciplinary Computational Science platform modules developed for Research: Industry-driven projects posted by member companies will train students on relevant software

<table>
<thead>
<tr>
<th>Consultant Teams</th>
<th>Assessments</th>
<th>Computational Science topic descriptions with select Wolfram tool boosting R&amp;D productivity with industry-driven projects (See section 5d for learning outcome detail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team 1:</td>
<td>A1-4</td>
<td>Wolfram platform API allows development of multitude of mobile applications with instant API functionality saving time in class. Appery is an example mobile development service that can be connected for rapid prototyping of research projects that have a mobile element.</td>
</tr>
<tr>
<td>Wolfram Platform API</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team 2:</td>
<td>B1-6</td>
<td>Wolfram Mathematica is a high-level language that makes it possible for students to learn the basics of mathematical concepts. It has a technology called Computable Document Format (CDF) that enables interactive documents to be created. There are nearly 5000 built-in functions that save time for students to get started. Over 150,000 examples are available, so students do not need to start from scratch.</td>
</tr>
<tr>
<td>Wolfram Mathematica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team 3:</td>
<td>C1-6</td>
<td>Wolfram Cloud provides high efficiency computation that allows students to devise high-computation solutions for complicated problems on inexpensive computer platforms. Programming Cloud can be done directly on the Web, making it very portable and cross platform. Cloud programming allows students to call API from other languages, enabling written code to be reused. The capability to embed in webpages and static document enables teaching from fully interactive websites and developing thesis documents with dynamic content for presentations.</td>
</tr>
<tr>
<td>Wolfram Cloud</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team 4:</td>
<td>D1-6</td>
<td>Wolfram Discovery Platform rapid prototyping ability enables students to sharpen their creativity. Seamless workgroup sharing makes it easier for students to form collaborative teams. Students can create modular symbolic components enabling large projects to be done in small parts. Students can create their own private cloud, which is a good place for students to work collectively on homework.</td>
</tr>
<tr>
<td>Wolfram Discovery Platform</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team 5:</td>
<td>E1-6</td>
<td>The Wolfram Data Science Platform allows importing data semantically so there is no need to learn new commands. Students can connect to external databases, which allow students to use specified private databases. The results in the Data Science Platform are not just strings and numbers, so analyzing data is easier with built-in images and graphs. The platform can read hundreds of formats, so many database structures can be considered for</td>
</tr>
<tr>
<td>Wolfram Data Science Platform</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
calculation. Machine learning is supported, so developing engines for specific purposes would be easier.

Team 6: Wolfram System Modeler

Team 7: Wolfram Connected Devices Framework

Table 6. Phased development of platform (TRCP) over three years

<table>
<thead>
<tr>
<th>Phased Development</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rational Unified Process (RUP)</td>
<td>Inception/ Elaboration</td>
<td>Elaboration/ Construction</td>
<td>Construction/ Transition</td>
</tr>
<tr>
<td>Platform Activity with stakeholders</td>
<td>Planning and prototype Development</td>
<td>Testing and Verification</td>
<td>Piloting and Validation</td>
</tr>
<tr>
<td>Platform module development in layers (L1-6+) with focus on industry driven projects supplied by companies and IOT training using Wolfram computation software</td>
<td>Modular Project Supplies Acquisition and IOT innovation in DPSL Consultants develop platform modules with CSCI 524/530/595 and SDPS student chapters Modular design and modeling of university-industry project framework according to requirements standard SRS-IEEE 830, design standard SDD-1016, and project standard PMP-IEEE 1058 Modular Project Supplies Acquisition</td>
<td>Modular Project Supplies Acquisition and IOT innovation in DPSL Consultants test platform modules with CSCI 524/530/595 and SDPS student chapters Modular development and testing of platform services with DPSL Industry supplies projects to students via platform</td>
<td>Modular Project Supplies Acquisition and IOT innovation in DPSL Consultants pilot CSCI 575 based on platform modules developed by CSCI 524/530/595 and SDPS student chapters Modular platform service utilization with project supplies primarily by CSCI 575 course pilot CSCI 575 course is piloted with fully operational platform (TRCP)</td>
</tr>
</tbody>
</table>

TRANSDISCIPLINARY RESEARCH
INTEGRATION GOALS FOR INDUSTRY SYMBIOSIS

SDPS has many activities and proceedings since its founding in 1995 to address the needs of transdisciplinary thinking in a world of overlapping disciplines requiring symbiotic relationships for productive discourse. These silos of disciplinarily can be defined in various ways (Table 7), leading to a definition of transdisciplinary that has eluded so many. Computational Science draws heavily from the lessons learned from such theory and puts it to practical use as evidenced by Society for Industrial and Applied Mathematics (SIAM) and other organizations like SDPS. SDPS conferences since 1995 have cultivated a unique group of leaders who can contribute to the platform by facilitating industry-driven project work. The SDPS Next Generation Leadership in the November conference is poised to assist in platform develop that will make transdisciplinary research correlation services a reality, not just for TAMUC CSIS, but for any university interested in joining the growing network of student chapters managed by the DPSL. In 2015, one of the objectives was to establish the DPSL at TAMUC to start the SDPS student chapter network. Five publications at the Dallas, TX conference established the goals of DPSL with ancillary benefits (Table 8).

The SDPS conference in Orlando, Florida further solidified plans to develop ties between the student chapters at TAMUC and UAB, as constitutions were formed while undergraduate accreditation was taking place successfully. As SDPS expanded its leadership capacity to manage new student chapters, TAMUC CSIS began its emphasis on improving the graduate program by revamping the
Computational Science Program to attract more students and make it more competitive. We hope that development of the platform with SDPS will create new opportunity by bridging the gap between senior members in R&D and our thesis students.

The conference program highlights the authors (Dr. John Tanik, Dr. Sang Suh) will co-chair this workshop in November 2017 “…The Internet of Things (IoT) is an integrated set of tools, technologies, and processes used to monitor and control the functioning and operations of a wide variety of environments such as smart meters, real-time health devices, manufacturing field devices, and process sensors in plants. According to forecasts by Gartner, 8.4 billion connected things will be in use worldwide in 2017 with a spending on endpoints and services reaching almost $2 trillion. It is expected that the number of connected things will reach 20.4 billion by 2020. In this workshop, we will discuss examples of the platform and tools that can be used for the design and implementation of IoT solutions…” The STEM workshops related to our work include topic in transformative education, STEM education & training, computational thinking, Entrepreneurship and innovation, and Advances in STEM education for academicians. We plan to invite the participants of this conference to contribute to our platform. Traditionally, at the end of every conference future themes are discussed. This year we plan to develop a theme advancing our transdisciplinary research correlation services for thesis students who can benefit from industry-driven projects that need computational R&D work. Interdisciplinary research and engineering has been in demand for several decades. Computational Science and Engineering plays an important role in interdisciplinary work. To improve the efficiency of education and collaboration of CSE in interdisciplinary work, people have already been creating platforms to allow students to remotely join research and engineering projects. Such platform will prepare the students to face the future work in research and engineering, as well as filling the gaps between science and industry. In “The engaged university; providing a platform for research that transforms society” (Whitmer, Ali, et al. & quot; 2010), discussed creative ideas and tested models that provide opportunities for conversation and thoughtful consideration about how such institutions can facilitate the dialogue between scientists and society. To efficiently provide interdisciplinary education to the students, interdisciplinary training has already been considered as a separate course other than traditional courses. In “Interdisciplinary teamwork in HPC education: Challenges, concepts, and outcomes” (Skulmoski, Gregory, Francis Hartman, and Jennifer Krahn, 2007) and “Interdisciplinary graduate training in teaching labs” (Lundstrom, Kristi, and Wendy Baker, 2009), it is shown that when interdisciplinary team research is offered as a separate course, it improves the quality of education and helps students to succeed in interdisciplinary projects (Planas Lladó, Anna, et al, 2014), (Planas Lladó, Anna, et al, 2014). (Neumann, Philipp, et al, 2017). (Vale, Ronald D., et al, 2012). (Reynolds, Julie A., and Robert J. Thompson, 2011). (Yang, Hen-L., et al, 2011). (Wang, Liya,2012). Many platforms and related topics were reviewed as preliminary research to get ideas on features to add beyond industry-driven project development with computational science training (Chi-Sheng Shih, Chia-Lin Yang, Mong-Kai Ku, 2005). (Chengling Zhao , Lin Wang , Desheng Wu, 2009). (Di Yang, Pedro Martins, Vaibhav Saini, 2017). (Bernadette M. Randles, Milena S. Golshan, Irene V, 2017). (M. Meyer, E. Autio,2004). (Zhenfan Ding, Rong Li,,2010). (Alexandra Broillet, Constance Kampf, Sabine Emad, 2014). (Mark Reynolds, 2011), (Li Chen-Guang, 2015). As platforms evolve, more advances will be reviewed over the next three years and included in the platform module specifications as needed by the university, industry, and students.

PERFORMANCE ASSESSMENT AND PROJECT EVALUATION

A Logic Model shows expected input, activity, and output for a given proposal, which is provided (Table 9). Platform development (TCRP) can progress for new Comp Sci program with guidance from IAB and 21 DPS consultants, including input from 80+ grad students in PI courses and SDPS student chapter participation. The DPSL will serve as an intellectual control over supplies acquisition for IOT product innovation activity. Curated technologies that interoperate with Raspberry Pi and Wolfram Framework can be connected together to produce products with emergent properties. Variable range goals for platform development that facilitates this emergent development, is provided in terms of short-term, medium-term, and long-term goals (Table 10). Formative and summative reporting includes statistical information on demographics and related information provided by department (Table 11), as well as periodic and final reporting descriptions with collection dates (Table 12). Detailed assessments can be compiled after collecting from various stakeholders the results of the industry-driven projects testing key topics in computational science applied in R&D by students in CSCI 575 and thesis students (Table 13).
Disciplinarity is the study of phenomena, assumptions, epistemology, concepts, theories, and methods-distinguish it from other knowledge formations.

Unidisciplinarity is the process in which researchers from a single discipline, field, or area of established research and education practice work singly or collaboratively to study an object or to address a common question, problem, topic, or theme.

Multidisciplinarity juxtaposes two or more disciplines focused on a question, problem, topic, or theme. Juxtaposition fosters wider information, knowledge, and methods, but disciplines remain separate and the existing structure of knowledge is not questioned.

Interdisciplinarity integrates information, data, methods, tools, concepts, and/or theories from two or more disciplines focused on a complex question, problem, topic, or theme. The scope and goals of research programs range from incorporating borrowed tools and methods and integrating them into the practice of another discipline to generating a new conceptual framework or theoretical explanation and large-scale initiatives. The key defining concept of interdisciplinarity is integration, a blending of diverse inputs that differs from and is more than the simple sum of the parts.

Transdisciplinarity transcends disciplinary approaches through more comprehensive frameworks, including the synthetic paradigms of general systems theory and sustainability. In the late 20th century, it also became aligned with problem-oriented research that crosses the boundaries of both academic and public and private spheres.

### Table 7. Disciplinary Definitions

<table>
<thead>
<tr>
<th>Topic</th>
<th>National Academies Press Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disciplinarity</td>
<td>Particular branch of learning or body of knowledge whose defining elements—such as objects and subjects of study, phenomena, assumptions, epistemology, concepts, theories, and methods—distinguish it from other knowledge formations.</td>
</tr>
<tr>
<td>Unidisciplinarity</td>
<td>Process in which researchers from a single discipline, field, or area of established research and education practice work singly or collaboratively to study an object or to address a common question, problem, topic, or theme.</td>
</tr>
<tr>
<td>Multidisciplinarity</td>
<td>Juxtaposes two or more disciplines focused on a question, problem, topic, or theme. Juxtaposition fosters wider information, knowledge, and methods, but disciplines remain separate and the existing structure of knowledge is not questioned.</td>
</tr>
<tr>
<td>Interdisciplinarity</td>
<td>Integrates information, data, methods, tools, concepts, and/or theories from two or more disciplines focused on a complex question, problem, topic, or theme. The scope and goals of research programs range from incorporating borrowed tools and methods and integrating them into the practice of another discipline to generating a new conceptual framework or theoretical explanation and large-scale initiatives. The key defining concept of interdisciplinarity is integration, a blending of diverse inputs that differs from and is more than the simple sum of the parts.</td>
</tr>
<tr>
<td>Transdisciplinarity</td>
<td>Transcends disciplinary approaches through more comprehensive frameworks, including the synthetic paradigms of general systems theory and sustainability. In the late 20th century, it also became aligned with problem-oriented research that crosses the boundaries of both academic and public and private spheres.</td>
</tr>
</tbody>
</table>

### Table 8. Prior work preparing for platform development

<table>
<thead>
<tr>
<th>Publications related to proposal</th>
<th>Focus areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Design and Process Science Lab: STEM Virtual Platform for Transdisciplinary Device Innovation Promoting Venture Development</td>
<td>Focus on establishing a common platform for IOT device innovation</td>
</tr>
<tr>
<td>1.2 Transdisciplinary Convergence on the DPSL Platform: STEM Development with IOT Mobile Applications in Wolfram Framework</td>
<td>Focus on STEM training with Wolfram and Mobile computing</td>
</tr>
<tr>
<td>1.3 DPSL Virtual Platform Module: SDPS Student Membership Web Database Development with Python, MySQL, and Apache server</td>
<td>Focus on STEM management using database and Cloud</td>
</tr>
<tr>
<td>1.4 SDPS Mobile Application Module for DPSL Virtual Platform: Enabling Next Generation Collaboration and Convergence for SDPS Student Chapter Network</td>
<td>Focus on Mobile application development for virtual platform</td>
</tr>
<tr>
<td>1.5 DPSL Platform Advancing Rapid-Prototyping Capability in IOT Product Design: Value Driven Modeling and Simulation Approach with the P3Tech Advantage</td>
<td>Focus on rapid prototyping capability</td>
</tr>
</tbody>
</table>

### Table 9. Logic model depicting general input, activity, and output over three years

<table>
<thead>
<tr>
<th>Input</th>
<th>Activity</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Director/Co-Director of Comp Sci Program/Chair of CSCI Dept</td>
<td>Director/Co-Director work together to manage platform development (TCRP) for new Comp Sci program with guidance from IAB and 8+ SDPS consultants, including input from 80+ grad students in courses</td>
<td>TRWF launch complete (pilot with CSCI 575 that utilizes platform resources)</td>
</tr>
<tr>
<td>8+ SDPS consultants</td>
<td>Admin (dean et al) evaluates Comp Sci program according to metrics</td>
<td>CSCI Computational Science program recruitment and quality increases with platform global resource support for research and development.</td>
</tr>
<tr>
<td>CSCI 524/530/595+ (30 +30+20 grad students)</td>
<td>SDPS conferences annually will develop workshops on IOT platform development until TRCP launch</td>
<td>More students will be prepared for MS/PhD thesis research using computational software applied to actual projects from industry.</td>
</tr>
<tr>
<td>IAB (companies)</td>
<td></td>
<td>Clear pathways to internships and job placement based on projects from industry.</td>
</tr>
<tr>
<td>SDPS conference (annual), 2017 workshop on IOT platform and convergence</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 10. Short-term, medium term, long-term goals for TRCP development

<table>
<thead>
<tr>
<th>Short term goals</th>
<th>Medium term goals</th>
<th>Long-term goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop TRCP as a platform framework to support new transdisciplinary TAMUC course CSCI 575 as pilot</td>
<td>Develop TRCP as a platform framework to support other transdisciplinary course courses in TAMUC computational science program tracks</td>
<td>Develop TRCP as a platform framework to support transdisciplinary resource correlation with other universities</td>
</tr>
<tr>
<td>Demonstrate CSC 575 as a way to improve research quality and marketability with advanced computational software (Wolfram) applied to actual industry projects</td>
<td>Demonstrate other courses in Comp Science program as a way to improve research quality with advanced computational software (Wolfram) applied to actual industry projects</td>
<td>Demonstrate other courses from various universities that would benefit from improved research quality with advanced computational software (Wolfram) applied to actual industry projects</td>
</tr>
<tr>
<td>Develop SDPS student chapter at TAMUC managed by DPSL utilizing the TRCP as a platform to attract global resources to advance student research interests and preparation</td>
<td>Expand SDPS student chapters at TAMUC, UAB, and UT-Arlington managed by DPSL utilizing the TRCP as a platform to attract global resources to advance student research interests and preparation</td>
<td>Rollout SDPS student chapters with networked universities managed by DPSL utilizing the TRCP as a platform to attract global resources to advance student research interests and preparation</td>
</tr>
<tr>
<td>SDPS workshop on IOT platform in 2017</td>
<td>SDPS workshops annually (with meetings in between) on IOT platform development until launch</td>
<td>SDPS workshops on platform module expansion beyond computational science research education</td>
</tr>
<tr>
<td>Increase enrollment in Computational Program and thesis research</td>
<td>Inspire MS students to continue to PhD with partner institutions that will help develop PhD program at TAMUC</td>
<td>Develop PhD program at TAMUC/Invite companies to supply thesis projects with networked opportunities</td>
</tr>
<tr>
<td>Invite companies to supply thesis projects</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11. Sample statistics on students in CSCI Computational Science research program

<table>
<thead>
<tr>
<th>MS thesis students with software knowledge</th>
<th>Internship experience</th>
<th>Financial aid received</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate degree graduation date</td>
<td>PhD/job bound</td>
<td>Disability status</td>
</tr>
<tr>
<td>Degree plan/track selected</td>
<td>MS student total</td>
<td>Professional society</td>
</tr>
<tr>
<td>Job interest</td>
<td>MS thesis student total</td>
<td>Undergraduate degree/GPA/GRE</td>
</tr>
<tr>
<td>Gender/Demographic area/Citizenship</td>
<td>MS thesis completion</td>
<td>Under-represented Minority (URM)</td>
</tr>
</tbody>
</table>

Table 12. Formative and summative reporting description and collection dates

<table>
<thead>
<tr>
<th>Report type</th>
<th>Description</th>
<th>Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formative reporting by administration</td>
<td>Feedback metrics will be provided on statistics related to the computational science program/attributes/demographics</td>
<td>Annual</td>
</tr>
<tr>
<td>Formative reporting by 8+ consultants</td>
<td>Consultants will provide feedback on how to improve the platform framework, research quality, and job placement according to industry projects</td>
<td>Semester</td>
</tr>
<tr>
<td>Formative reporting</td>
<td>Director/Co-Director will provide surveys from students in computational science program including three courses taught that help improve graduate student preparation in research methods (CSCI 530/524/595)</td>
<td>Semester</td>
</tr>
<tr>
<td>Formative reporting by IAB</td>
<td>Local company feedback from industrial advisory board (IAB)</td>
<td>Annual</td>
</tr>
<tr>
<td>Summative reporting by administration (by dean and/or chair)</td>
<td>A final report will be prepared on the results of the program by administration</td>
<td>Annual</td>
</tr>
<tr>
<td>Summative reporting by investigators (PI and/or Co-PIs)</td>
<td>A final report will be prepared on the results of the program by PI with support from Co-PIs</td>
<td>Annual</td>
</tr>
<tr>
<td>Summative reporting by Principal investigator (PI)</td>
<td>A final report will be prepared on the results of the program by PI</td>
<td>3rd year</td>
</tr>
</tbody>
</table>
Table 13. Assessing industry-driven research projects using computational science software

<table>
<thead>
<tr>
<th>Mobile Application Development</th>
<th>Consultant Team 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Features</strong></td>
<td><strong>Benefits</strong></td>
</tr>
<tr>
<td>Cross-Platform Development</td>
<td>Create mobile apps for device platforms</td>
</tr>
<tr>
<td>Visual Development</td>
<td>Visual Editor ergonomics</td>
</tr>
<tr>
<td>Integrated mobile backend services</td>
<td>Focus on apps, not infrastructure</td>
</tr>
<tr>
<td>API Express Plug-ins</td>
<td>Enables enterprises to easily and securely integrate apps with any back-end system.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mathematica</th>
<th>Consultant Team 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Features</strong></td>
<td><strong>Benefits</strong></td>
</tr>
<tr>
<td>High level language</td>
<td>Could be learned quickly by users</td>
</tr>
<tr>
<td>CDF</td>
<td>Interactive documents could be created</td>
</tr>
<tr>
<td>~5000 built-in function</td>
<td>Saves time</td>
</tr>
<tr>
<td>Real-time database</td>
<td>could be done analyze dynamic big data</td>
</tr>
<tr>
<td>parallel computing</td>
<td>Fast setup</td>
</tr>
<tr>
<td>150,000+ examples available</td>
<td>No need for starting from scratch</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Programming Cloud</th>
<th>Consultant Team 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Features</strong></td>
<td><strong>Benefits</strong></td>
</tr>
<tr>
<td>High efficiency computation</td>
<td>Big problems solved with cheap PCs</td>
</tr>
<tr>
<td>On the web</td>
<td>Portable and cross platform</td>
</tr>
<tr>
<td>Create an instant API</td>
<td>Saved APIs saves class time</td>
</tr>
<tr>
<td>Call from other languages</td>
<td>Written codes reused</td>
</tr>
<tr>
<td>Embed in webpages</td>
<td>Fully interactive websites</td>
</tr>
<tr>
<td>Symbolic deployment</td>
<td>Codes similar to math equations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discovery Platform</th>
<th>Consultant Team 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Features</strong></td>
<td><strong>Benefits</strong></td>
</tr>
<tr>
<td>Rapid prototyping</td>
<td>Creativity of the students sharpened</td>
</tr>
<tr>
<td>Seamless workgroup sharing</td>
<td>Collaboration vital for any work</td>
</tr>
<tr>
<td>Data from anywhere – Including devices</td>
<td>Obtained experimental data analyzed</td>
</tr>
<tr>
<td>Create modular symbolic components</td>
<td>Big projects done by smaller pieces</td>
</tr>
<tr>
<td>Your own private cloud</td>
<td>Private clouds good place for homework</td>
</tr>
<tr>
<td>Interactive documents</td>
<td>CDF precise dynamic results</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Science Platform</th>
<th>Consultant Team 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Features</strong></td>
<td><strong>Benefits</strong></td>
</tr>
<tr>
<td>Import data semantically</td>
<td>No need to learn new commands</td>
</tr>
<tr>
<td>Connect to external databases</td>
<td>Specified private databases.</td>
</tr>
<tr>
<td>Not just string &amp; numbers</td>
<td>Analyzing data is easier with built-in images,</td>
</tr>
<tr>
<td>Read hundreds of formats</td>
<td>Feature covers more database structure</td>
</tr>
</tbody>
</table>
PRIOR WORK TO DEVELOP COMPUTATIONAL BIOLOGY MODULE FOR PLATFORM

Prior NSF (National Science Foundation) work provides opportunities to utilize the platform in future computational biology course, considering the achievements of the recent grant awarded by NSF. A new representation for RNA secondary structure was introduced. A new algorithm for efficient searching of RNA substructures was presented. New algorithms for finding largest common substructures in multiple RNAs were developed. These algorithms exploit the properties of the new structure representation and propose efficient algorithms for computationally difficult problems. New structure drawing tools were also developed. By using a drawing tool, the user can draw a substructure and generate a substructure representation for searching. More publications (Arslan, A. N., He, D., He, Y., and Wu, X. 2016), (Arslan, A. N., George, B., and Stor, K. 2015), (Wu, X., Zhu, X., He, Y., Arslan, A. N. 2013) were produced in past five years based on an NSF award received more than five years ago that could benefit from the platform. The project formulated a computational problem motivated from finding in biological sequence repeating patterns with wildcards and length constraints and with also the condition that found letters cannot be shared by different occurrences.

PROPOSAL CONCLUSION TO TRANSDISCIPLINARY RESEARCH CORRELATION PLATFORM

This proposal concludes with the prospect of developing an effective platform for transdisciplinary resource correlation services that can be developed with the assistance of 21 consultants (or more) from the Society for Design and Process Science. A maximum of 8 letters of collaboration have been furnished with support from university administration to achieve this goal over three years. The DPSL will manage a growing SDPS student chapter network while 80+ students in courses can be given modules to develop in teams as part of their course each semester. The platform will be called Transdisciplinary Resource Correlation Platform and run a pilot of CSCI 575 Cyber-physical systems and Industrial IOT. An essential component of the course is to develop student projects that utilize computational software with hardware for industry-driven projects.

REFERENCES


DPSL Virtual Platform Module: SDPS Student Membership Web Database Development with Python, MySQL, and Apache server, SDPS Proceedings, Dallas, TX, 2015.


M. Meyer, E. Autio.,(2004) “Academic inventiveness and entrepreneurship: is there a relationship between science and technology fields and the utilization of academic


AN EMPIRICAL STUDY ON THE MEASURABILITY OF MEDITATION

Hong Lin¹, Steven Lawrence Fernandes², V. Rajinikanth³

¹University of Houston-Downtown, Houston, Texas, USA
²University of Alabama at Birmingham, USA
³St. Joseph’s College of Engineering, Chennai, India

{linh@uhd.edu stevenlf@uab.edu rajinikanthv@stjosephs.ac.in}

ABSTRACT

Physiological signal measurement is widely adopted in health centers to assess the activity of internal human organs. Electroencephalogram (EEG) is one of the important physiological signal gives the information regarding the Electrical Activity (EA) of brain. In the proposed work, an experimental study is implemented to evaluate the EA of brain during various conditions, such as idle, talking and meditation. During this study, a multi-channel EEG signal patterns are recorded using a wireless sensor device, Emotiv EPOC. The signals picked by the electrodes (14 numbers) are then recorded using a personal computer equipped with the necessary pre-processing and classifier algorithms. The main task considered in this work is, differentiating the EEG signature of meditation state with other states considered in this study. Finally, five well known classifier units, such as nearest neighbor, 1R algorithm, Naive Bayes classifier, decision trees, and random forest are considered to classify the EEG patterns and found that, random forest outperforms the alternatives by providing the classification accuracy of greater than 99.92% for all the cases.

1. INTRODUCTION

Human physiological signals play a significant role in the examination of vital internal organs. The modern clinical and therapeutic approaches always require appropriate tools to record and examine the electrical activity of the internal organs during the normal, abnormal and relaxed conditions.

Due to its clinical significance, in recent years, a number of automated and semi-automated approaches are widely proposed by the researchers to examine physiological signals, such as electromyography, electrocardiography, electrooculography, and electroencephalogram. These signals are normally examined to ensure the activity and the healthiness of the organs. Among them, electroencephalogram (EEG) is considered as the chief physiological signal, usually record to verify the electrical activity of the brain (Dissanayaka et al., 2015). The EEG signals generated due to the activities of neurons are normally recorded using a single or multichannel electrodes placed on the scalp region. During the clinical diagnosis, the required EEG signals are recorded by placing the electrodes at some predefined locations. Recently, the EEG signals are widely considered to examine syndromes, like dementia, Alzheimer’s disease, sleep disorder, and epilepsy.

The work by Al-Qazzaz et al. (2014) reports that, the change in EEG frequency pattern can be considered as the chief information to identify neurodegenerative disorder known as the dementia. Colloby et al. (2016) proposed a hybrid approach based on EEG-MRI to examine the Alzheimer’s Disease (AD) and dementia by means of Lewy bodies. Hata et al. (2016) presented a study by relating the resting stage EEG with the Alzheimer’s disease and confirmed that artifact-free EEG data provides better result in the detection of AD. Lee, Brekelmans & Roks (2015) confirmed that, the grand total EEG score will help to distinguish dementia with Lewy bodies from AD. Kulkarni & Bairagi (2017) proposed a novel approach to extract and evaluate the AD based on EEG signal and Support Vector Machine (SVM).

Abad and Guilleminault (2003) presented a detailed examination on the diagnosis of sleep disorders (SD) based on the EEG signals and possible treatment procedures. Aboalayon et al. (2016) provided a detailed assessment for the sleep stage classification based on the EEG signals. This work consist of the procedures, such as the pre-processing using digital band-pass filter, feature extraction based on maximum-minimum distance of the EEG signal peaks, and classification of these signals using well known classifier algorithms. This study confirms that, assessment of EEG signal is enough to identify the cause for the sleep disorders. The work of Siddiqui, Srivastava, & Saeed (2016) reports short time frequency analysis of Power Spectrum Density (PSD) of EEG signals to examine SD. Tan et al. (2012) discussed about the Empirical Mode Decomposition (EMD) approach to identify the SD from the EEG patterns.

An EEG based analyzing tool was recently discussed for evidence-based diagnosis of Epilepsy (Lagunju et al., 2015). Giovannini et al. (2013) discussed an EEG based approach to examine the Epilepsy in ring 14 syndrome. This study is clinically tested and validated using 22 patient’s EEG signals recorded at various intervals. Pedreira et al. (2014) presented an integrated approach between the EEG and fMRI to examine the partial epilepsy. The research work by Zwoliński et al. (2010) presents a combined examination of EEG and MRI database collected from Epilepsy patients. This work confirms that, the mapping of the EEG with the MRI can be used to detect the region and rate of the Epilepsy. This work also provides a detailed assessment of on the multichannel EEG signal to examine the disease.
The recent work by Thanaraj & Parvathavarthini (2017) confirms that, the Epilepsy can be detected and classified using the time–frequency entropy measure of the Multichannel EEG signal. In this work, the abnormality is sensed by examining the interictal spike activity of the EEG.

The EEG signals are also considered to examine the electrical activity of the brain during the meditation process. At the time of the meditation, the brain muscles will be in relaxed state, which alters the electro-chemical signals generated at various parts of the brain. In order to examine the brain activity during the meditation process, these signals are recorded using the external brain electrodes at a controlled environment. After recording the neuro-physiological signatures (EEG), the recommended pre-processing, post-processing and classification approaches are implemented to examine the conditional changes in the brain. Due to its importance, in recent years, a considerable research work is proposed to examine the EEG during the meditation process. Banquet (1973) proposed a spectral analysis on the real time EEG signals recorded during the meditation and confirmed that, meditation level can be measured by analyzing the frequency and amplitude patterns of the EEG signals. Takahashi et al. (2005) conducted an experimental investigation on Zen meditation with twenty volunteers and examined their physical and mental state. This study confirms that, Zen meditation will influence in fast theta power and slow alpha power on EEG signature.

Surangsrirat & Intarapanich (2015) analyzed the change in rhythmic activity of EEG during meditation. Gupta, Ramdinmawii & Mittal (2016) presented a detailed investigation on changes in alpha brainwaves during the meditation/yoga. The work by Thomas &Cohen (2014) presents a detailed review on the meditation research. In this work, various meditation practices, such as Zen, Qigong, Tibetan Buddhist, Yoga, Ananda marga, Kundalini, etc. were examined based on the EEG signals and the psychophysiology of the volunteers were examined. This work also confirms that, balanced meditation will improve the physical and mental ability of the users. Recently, Lin & Li (2017) proposed an experimental investigation to measure the meditation process based on the recorded multichannel EEG signals in the controlled environment. In their investigation, brain EEG samples with pseudo names, like meditation 1, meditation 2, idle, talking and inexperienced meditation were considered. Entropy assisted approach is considered to process the obtained dataset and finally the recorded EEG signals are then classified using the classifiers, like tree bagging, Gaussian mixture model and support vector machine. This work confirms that, measurement of the meditation is possible with multichannel EEG data.

In this research work, the dataset of Lin & Li (2017) is considered to investigate the structural change in EEG signature pattern during the normal (idle), abnormal (talking, and inexperienced meditation) and relaxed (meditation 1, and meditation 2) conditions. In this work, a feature extraction and classification is proposed for the recorded EEG pattern in order to measure the meditation level with a comparative study using the EEG patterns of idle and talking conditions.

2. METHODOLOGY

In this paper, detailed investigations on the change in the electrical activity of the brain during the normal and meditation conditions are investigated using the dataset of the recent research work reported by Lin & Li (2017). The necessary signal pattern is recorded using non-invasive approach with the help of multi-channel wearable EEG sensor equipped with 14 electrodes. The recorded EEG dataset is initially pre-processed in order to identify and remove the low Contact Quality (CQ) signals and then classified using the well known classifiers existing in the literature.

2.1 EEG Signal Acquisition

The EEG database for the normal, abnormal and meditation conditions are primarily recorded using the EPOC wearable sensor headset, a low cost individual brain-computer interface unit developed by Emotiv Corporation (EMOTIV). This sensor is designed to provide a 14 channel EEG signal in order to support the brain-computer interface. This headset is prepared with 14 high quality primary electrodes and two secondary (reference) electrodes. The primary electrodes of EPOC are stickered as AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4,F8, and AF4 and the secondary electrodes are marked as P3 and P4. This EPOC tool is also supplied with some fundamental software tools for signal pre-processing, training, testing and classification of EEG signals (Benitez, Toscano & Silva, 2016).

In the proposed work, the electrodes are placed as depicted in Fig. 1. Each electrode is responsible to record the certain electrical activity of the brain region. During the real time experimental work, the major problem is low signal strength due to the electrode contact quality. In the proposed research work, the signals are obtained using Emotiv Education Edition to record EEG signals from 14 channels Emotiv EPOC headset. The sampling rate of EPOC headset is 128. EEG data are transferred to CSV format and saved in computer by the Emotiv software.

![Fig. 1 Placement of electrodes to record the EEG pattern](image-url)
The main task considered in this paper is measuring the brain activity during the meditation process. In order to obtain the EEG dataset, an experienced meditator is chosen as the volunteer. During this experimental work, the EEG pattern is recorded during the deep meditation state, at rest and during a conversation. All three activities are separately recorded using a personal computer and the signal patterns are then examined after implementing the possible pre-processing. This examination is revealed that, the EEG recorded during the meditation is very smooth compared with the EEGs recorded during the conversation and idle condition. From this data, it is clear that, if we can measure the amplitude and the frequency level of these EEGs, it is possible to build soft-computing models which can classify the EEG dataset into idle, talking and meditation.

![Flow chart depicting the stages adopted in this research work](image)

Fig. 2 Flow chart depicting the stages adopted in this research work

Fig 2 shows the work flows considered in the proposed research to record and classify the EEG signals. The accuracy of the EEG signal depends mainly on the contact quality and the placement of the electrodes. Hence, initial care is taken during the placement of the record to the volunteer. Further, the proposed experimental work is executed over subjects during different brain activities, viz., meditation, talk, and idle. The duration of each recording is 3-5 minutes.

### 2.2 Data Preprocessing

The raw data coming out from the sensor always requires some initial processing steps in order to smooth the measured variables by removing unwanted signals. In the proposed work, pre-processing involves in identification and removal of outliers, missing information and EEG with low CQ value. Firstly, it was guessed that values with low CQ would result in poor data quality which would need to be removed prior to model evaluation. Based on the quality, the range for the CQ is fixed from 0 to 4 and the electrodes signal with low CQ (ie. CQ < 2) is removed from the dataset. This procedure is implemented in all the three (normal, abnormal and meditation) cases.

This work is implemented based on the assumption that, lower CQ values would represent more spread in the data and outliers than the data with higher CQ. Based on the graphical evaluation there appears to be no such correlation and have determined that although there are three outlier points which need to be removed are spread across all CQ values (including CQ of 4 which is supposed to represent the best quality data). This also confirms that, CQ values did not relate to data quality, it is closely related with the outliers, which is to be identified and removed. The pre-processed data is then considered to build the classifier unit.

### 2.3 Feature extraction

Feature extraction plays essential role in the examination of EEG signals. In the literature, a substantial amount of approaches are available to extract the feature of EEG signals (Al-Fahoum & Al-Fraihat, 2014). Normally, EEG signals are complex and random in nature. Hence, in the proposed approach, well known procedures, such as entropy based extraction and principle component analysis are adopted to extract the features.

#### 2.3.1 Entropy Based Approaches

Usually, entropy is widely considered to examine the uncertainty existing in the test signals (Al-Fahoum & Al-Fraihat, 2014). In the meditation analysis, entropy can be used to compute the level of chaos (Lin & Li, 2017) in brain-computer interface systems, entropy can be used to measure the level of chaos of the system. It is a non-linear measure quantifying the degree of complexity in a time series. Let us consider, Abe a set of finite discrete random variables \( A = \{ a_1, a_2, \ldots, a_n \} \), aic \( R^2 \), then the Shannon entropy, \( H(X) \), can be expressed as follows:

\[
H(A) = -c \sum_{i=0}^{c} p(a_i) \ln p(a_i)
\]
where, c is a positive constant, and \( p(a_j) \) is the probability function with the following equation:

\[
\sum_{a=0}^{c} p(a_j) = 1
\]  

More details regarding the entropy approach is available in (Phung et al., 2014; Lin & Li, 2017, Thanaraj & Parvathavarthini, 2017).

### 2.3.2 Principal Component Analysis

Principal Component Analysis (PCA) is widely adopted in data processing fields in order to find the likeness and dissimilarity in data patterns (Lindsay, 2002). The merit of the PCA compared with the other approach is, after finding the required pattern; the dataset can be compressed without losing the information. Due to its importance, the PCA based approach is widely adopted by the researchers to examine the EEG patterns (Polat & Gunes, 2008; Lekshmi, Selvam & Rajasekaran, 2014).

### 2.4 Classification

Classifiers are widely used to support the automated examination of the signal/data. Initially, the classifier unit is to be trained based on the requirement and later, the performance of the classifier is to be tested to confirm its efficiency. In the proposed work, classifier system is used to distinguish the meditation EEG dataset.

#### 2.4.1 Nearest Neighbor (KNN)

In pattern recognition domain, K- nearest neighbor (KNN) approach is widely considered to solve the classification and regression tasks (Yazdani, Ebrahimi & Hoffmann, 2008). The KNN classifier calculates the dataset’s class based on the K training trials with respect to nearest neighbors to the test sample, and relates it to a group which has the principal category probability (Suguna & Thanushkodi, 2010). It is a non-parametric approach and guided with the following mathematical expression:

Let us consider the sample dataset is \( X \). If there are \( j \) training groups, like \( C_1, C_2, ..., C_j \), and the total after the feature reduction is \( N \), then it can be represented by a m-dimensional feature vector. If the training samples are represented as \( X_1, X_2, ..., X_m \), then the similarity (S) between samples can be computed as:

\[
S(X,d_j) = \frac{\sum_{j=1}^{m} X_j \cdot d_j}{\left(\sum_{j=1}^{m} X_j\right)^{\frac{1}{2}} \left(\sum_{j=1}^{m} d_j\right)^{\frac{1}{2}}}
\]  

where \( i \) is the sample value with \( i=1,2,\ldots,N \).

The probability of \( X \) can be computed as:

\[
P(X,C_j) = \sum_{d} S(X,d_j) \cdot y(d,C_j)
\]  

where, \( y(d,C_j) \) is an attribute function with the following value:

\[
y(d,C_j) = \begin{cases} 
1, & d_j \in C_j \\
0, & d_j \not\in C_j 
\end{cases}
\]  

Here, the KNN considers the \( X \) which as the leading \( P(X,C_j) \).

#### 2.4.2 1R Algorithm

The 1R algorithm is an easy rule-based classification procedure for distinct features. The aim of this approach is to assume a rule that forecast the class given the values of the features. The 1R algorithm prefers the most informative single feature and considers the rule on this feature alone. More information regarding this algorithm can be accessed from (Holte, 1993; Nevill-Manning, Holmes & Witten, 1995; 1R)

#### 2.4.3 Naive Bayes Classifier

Naive Bayes (NB) classifier systems will constantly aims to achieve the finest premise \( U \) through a given training dataset. The Bayes theorem permit one to compute the posteriori probability based on the a priori probability (Machado & Balbinot, 2014).

The mathematical expression is:

\[
E(V_j|B) = \frac{E(B|V_j) \times E(V_j)}{E(B)}
\]  

where \( V_j \) is the hypothesis \( j \) in the set of hypotheses \( V \), and \( B \) is the set of attributes \( b_1,b_2,...b_m \) which relating the data. When \( B \) has more than one attribute, it is then necessary to estimate \( E(V_j|b_1,b_2,...b_n) \) in order to calculate \( E(b_1,b_2,...b_n|V_j) \).

The joint probability for the above condition can be expressed as:

\[
E(b_1,b_2,...b_n|V_j) = \Pi_i E(b_i|V_j)
\]  

Finally, the classifier output is given as follows;

\[
V_{MAP} = \arg \max_{V_j \in V} E(V_j) \times \Pi_i E(b_i|V_j)
\]  

Where, \( V_{MAP} \) is the maximum a posteriori probability considered inside the region of hypotheses \( V \).

#### 2.4.4 Decision Trees

Decision Tree (DT) approach considers a tree like structure with a series of test questions. DT approach considers the attribute test conditions as the root and internal nodes and the class label (Yes/No) form the terminal node. Once the DT structure has been constructed, classification is achieved easily based on the decision taken in the each branch of the tree. More details regarding the EEG classification based on the DT can be found in the following articles (Aydemir & Kayikcioglu, 2014; Aboalayon et al., 2016; Bastos, Adamatti & Billa, 2016).
2.4.4 Random Forest

In recent years, Random Forest (RF) technique is widely considered to classify the complex datasets. RF was initially proposed by Breiman (2001) and its detailed justification can be found in the paper by Chen et al. (2014).

Let us consider, \( T \) is the training set with \( (x_1, y_1), (x_2, y_2), \ldots (x_n, y_n) \), \( N_{\text{tree}} \) is the number of tree to be built, \( M_{\text{try}} \) is the number of variables chosen for splitting at each node, then the classification is performed based on majority vote among the \( N_{\text{tree}} \). Detailed theory on random forest can be found in (Chen et al., 2014; Nguyen, Huang & Nguyen, 2015).

3. RESULTS AND DISCUSSIONS

This section presents the results obtain with the experiment and its evaluation. This section presents the details regarding the EEG signal acquisition, pre-processing and the classification procedures.

Initially, the essential EEG data is recorded using the 14 channel EPOC electrode as discussed in section 2 with the help of experienced meditator. Fig. 3, 4 and 5 depicts the EEG data recorded during the actions, like meditation, idle and conversation respectively. From these figures, it can be observed that, during the meditation process, the activity of the brain is slow hence, a very smooth and linear EEG patterns are extracted by the electrodes compared with the other actions.

From Fig.5, it can be observed that, random and non-linear spikes are extracted by the electrodes, which confirm the rapid change in brain’s electrical potential. These datasets are further evaluated to identify the change in EEG pattern during the meditation process.

Fig 6 depicts the raw data obtained from the electrodes AF3, F7, F3, and FC5 for the three conditions like meditation, idle and talking. This figure confirms that, the band occupied by the meditation dataset is comparatively small than the idle and talking actions. Similar result is obtained for the remaining 10 electrodes existing in the EPOC. The raw data is then pre-processed with eliminating the outliers and low CQ data before implementing the feature extraction and classification procedures.

After the possible pre-processing and the classifier training approach, a sum of 26700 data points from each test
case is considered to test the classifier accuracy. The robustness of the classifier unit is also tested using the raw data with 26700 data points and the corresponding results are depicted in Table 1 and Fig 7 and 8. Table 7 shows the data points of the meditation EEG patterns and similar results are obtained with the EEG points of idle and talking. From this results, it can be noted that, RF approach out performs the other approaches by offering the better values of % right and % error. This result also confirms the efficiency of the DT and KNN compared with the 1R and NB. This table also confirms that, the clean data points will help to achieve better classification result compared with the raw data points.

Table1. Results of various classifier units for the raw and clean data (meditation dataset)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Right points</th>
<th>Error points</th>
<th>% Right</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>25499</td>
<td>1201</td>
<td>95.50</td>
<td>4.50</td>
</tr>
<tr>
<td>1R</td>
<td>20073</td>
<td>6627</td>
<td>75.18</td>
<td>24.82</td>
</tr>
<tr>
<td>NB</td>
<td>19664</td>
<td>7036</td>
<td>73.65</td>
<td>26.35</td>
</tr>
<tr>
<td>DT</td>
<td>26649</td>
<td>51</td>
<td>99.81</td>
<td>0.19</td>
</tr>
<tr>
<td>RF</td>
<td>26680</td>
<td>20</td>
<td>99.93</td>
<td>0.07</td>
</tr>
<tr>
<td>KNN</td>
<td>25537</td>
<td>1163</td>
<td>95.64</td>
<td>4.36</td>
</tr>
<tr>
<td>1R</td>
<td>20140</td>
<td>6560</td>
<td>75.43</td>
<td>24.57</td>
</tr>
<tr>
<td>NB</td>
<td>21.608</td>
<td>5092</td>
<td>80.93</td>
<td>19.07</td>
</tr>
<tr>
<td>DT</td>
<td>26662</td>
<td>38</td>
<td>99.86</td>
<td>0.14</td>
</tr>
<tr>
<td>RF</td>
<td>26686</td>
<td>14</td>
<td>99.95</td>
<td>0.05</td>
</tr>
</tbody>
</table>

The experimental result confirmed that, the RF approach offered better classification result compared with the alternatives. In this work, RF algorithm is implemented using 500 trees, which will increase the computation time of CPU. Hence, the RF approach is further examined in order to improve its evaluating capacity. Fig 8 shows the mean decrease in Gini of every electrode. Fig 9 and 10 presents the convergence of RF algorithm during the classification task. The results of Fig 9 and 10 confirm that, around 200 trees are sufficient to attain the required result. If the initial tree value is minimal, then without degrading the performance of the classifier, better results can be achieved.

4. CONCLUSION

Proposed work implements an experimental investigation to identify the meditation process by analyzing the EEG signature. In this work, a multi-channel EEG is recorded using the personal EEG sensor headset, known as EPOC. This sensor is designed to extract the EEG using 14 dedicated electrodes marked with specific labels. The main task is to measure the meditation level with the EEG, hence an experienced meditator is chosen as the volunteer to obtain the EEG database. In order to examine the other activities of the volunteer, along with the meditation signal, other signals like idle and talking signals are also recorded. These signals are pre-processed to attain a clean signal by correcting the outliers and low CQ values. Finally, the recorded signals are classified using most successful classifier algorithms, such as...
KNN, IR, NB, DT, and RF. The experimental investigation confirms that, RF case-d classifier offers better result compared with the other approach considered in this study. From this work, it is confirmed that, by analyzing the amplitude, frequency and the pattern of the EEG data, it can be possible to predict the meditation level.

ACKNOWLEDGEMENT

We thank Dr. K. Suresh Manic of Caledonian College of Engineering, Sultanate of Oman, Muscat, for his help in our review process of the paper.

REFERENCES

1R[http://www.cs.nyu.edu/faculty/davise/ai/1R.html]
EMOTIV [https://www.emotiv.com/epoc/]


Phung, D., Tran, D., Ma, W., Nguyen, P. & Pham, T. (2014), Using Shannon entropy as EEG signal feature for fast person identification, In ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), 23-25 April 2014, PP:413-418.


A CLOSER LOOK AT VIDEO BASED SIDE CHANNEL ATTACKS ON THE SMARTPHONE USERS’ PIN

Diksha Shukla, and Vir V. Phoha
Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY-13210
{dshukla@syr.edu, vyphoha@syr.edu}

ABSTRACT
We take a closer look at video-based side-channel attacks on the smartphone users’ sensitive information specifically for the eavesdropping over the users’ Pins. The paper utilizes the short video clips of smartphone users’ hand’s movements while they type on their smartphone screen. We assume that the attacker does not have access to the smartphone screen display and can only utilize the video recordings of the users’ hands movements information.

We discuss existing attack methods and investigate various parameter settings to analyze the effect of these settings on the users’ Pin information. Our work considers various realistic scenarios such as different typing style, the angle of recording, recording frame rate etc. and suggests the combination of parameters which results in high success rate.

Our results suggest that although the performance of the attack degrades under realistic scenarios, the amount of information leak significantly reduces the exhaustive search complexity.

Keywords — Security; Smartphone; Pin-lock; Passwords; Video Analysis; Side Channel Attack;

INTRODUCTION
Modern wearable devices, such as, smart watches, smart glasses, smartphones, and tablets are equipped with cameras which are not only capable of capturing pictures but also can unobtrusively record a very good quality video. While the camera-equipped hand-held devices enable us to capture moments, record and store important events, they also pose and amplify the security risks.

Recent researcher has shown that video recordings may be used as side channels to steal the smartphone users’ sensitive information such as pin, password, and text typed (Backes, Durmuth, & Unruh, 2008; Balzarotti, Cova, & Vigna, 2008; Maggi, Gasparini, & Boracchi, 2011; Miluzzo, Varshavsky, Balakrishnan, & Choudhury, 2012; Raguram, White, Goswami, Monrose, & Frahm, 2011; Shukla, Kumar, Serwadda, & Phoha, 2014; Y. Xu, Heinly, White, Monrose, & Frahm, 2013; Z. Xu, Bai, & Zhu, 2012; Yue, Ling, Fu, Liu, Ren, et al., 2014). Although these attacks have shown from moderate to high level of security risks, the practicality of these methods is not very clear. For example, how the attack performs if the video is recorded from a remote distance? Does the attack work with low-resolution video recording? How much the attack performance depends on the smartphone screen dimensions, recording frame rate, the angle of recording, etc.? Does this attack depend on the users’ typing behavior? To address these questions, we designed our experiments on a dataset of 240 short video clips of pin entry process from 30 volunteer participants in five different settings using three different video recorders. We design our attack model like the attack design by Shukla et al. (Shukla et al., 2014) to analyze the effect of various parameters settings.

Our results show that the attack is effective however the performance is affected by the users’ typing behavior, recording frame rate, video resolution, recording distance, and the angle of recording. On the Pin entry process on an HTC One phone keypad, the attack could correctly infer 86% of the Pins using video recording captured from 3 meters from the user. We evaluated the effectiveness of all the six parameters and found that video recordings captured from the front, less than 10 meters distance with ×10 zoom, and lighting better than 600 lumens, results in good performance of the attack (see Table 1).

Figure 1 shows the scenario which we analyzed in our prototype experiment. A user is typing on her mobile phone, and an attacker is using a high-resolution camera with optical zoom to capture the videos of the user’s hands movements from various recording distances, angles, under different light conditions, and at different frame recording rate.

The following are the contributions of the paper:

1. We evaluated various parameters settings and analyzed the practicality of the video based attack on the smartphone user’s Pin. We provide the detailed analysis on the effectiveness of the attack with different video recorder settings and the users’ typing behavior. Since the prototype could work in realistic scenarios however the prediction accuracy was affected, our work shows a serious threat with the new range of sophisticated cameras now introduced in the market (such as smart glasses and smart watches).

2. Using a dataset of 240 video recordings in different settings from 30 volunteer participants, we classify the recorder parameter settings in three different categories: (1) recommended if the attack performed higher than 65% Pin prediction accuracy, (2) acceptable if the attack performed with more than 50%, and (3) not recommended if the attack accuracy is less than 50% (see Table 1).
The rest of the paper is organized as follows. We first discuss the related work, datasets, and our prototype experiment design and analysis. At the end, we present the results and draw our conclusions.

Fig. 1. A user is typing her Pin on the mobile phone screen and an adversary is recording the video of her hands’ movements using a video recorder with optical zoom from different recording distances, various angles, frame rate, and resolution.

RELATED WORK

Previous research has raised awareness on various attack designs which show text reconstruction using the video clips of the user’s typing (Balzarotti et al., 2008; Maggi, Gasparini, et al., 2011; Shukla et al., 2014; Y. Xu et al., 2013; Z. Xu et al., 2012; Yue, Ling, Fu, Liu, Ren, et al., 2014; Yue, Ling, Fu, Liu, Yu, et al., 2014; Yue, Ling, Liu, Fu, & Zhao, 2014). Some of these methods use direct observation techniques i.e. the attack model assumes that the adversary has the access to the user’s smartphone screen display.

Attack design by Xu et al. (Y. Xu et al., 2013) shows the text reconstruction using a low-resolution video recording of the fingertip and the screen on which typing is being done. They show the video recording captured from a long distance such that it was impossible for the user to see the attacker. They show the attack to work even with the recording of the reflection of the typing finger and phone screen.

The attack design by Xu et al. was based on the classifier trained on the appearance of the screen while a key is being typed. For users who can hide their screen while they type, the attack by Xu et al. will fail.

The work by Raguram et al. (Raguram et al., 2011) and Backes et al. (Backes et al., 2008) also show the text reconstruction using reflection. In their work, they use state of the art image processing techniques to reconstruct the text.

The attack design by Balzarotti et al. (Balzarotti et al., 2008) uses a video recording of user’s typing on the desktop keyboard. The video recording was done in such way that camera directly points to the keyboard. Balzarotti et al. used a series of computer vision analysis followed by language modeling techniques to infer the text typed on the desktop keyboard. They use the information about which keys are not visible in the video while a key is being pressed and puts them as candidate keys to be pressed.

The attack by Maggi et al. (Maggi, Gasparini, et al., 2011; Maggi, Volpatto, Gasparini, Boracchi, & Zanero, 2011) uses the video recording while the user is typing on the smartphone screen. Their attack takes advantage of display feedback mechanism (the enlarged key display while it is being typed). In their attack set up, the recording was done while the camera was directly pointing to the smartphone screen. They use a classifier trained on the appearance of the enlarged display of characters to determine the characters typed in a stream of video recording. A similar kind of attack recently been proposed by Yue et al. (Yue, Ling, Fu, Liu, Ren, et al., 2014; Yue, Ling, Fu, Liu, Yu, et al., 2014; Yue, Ling, Liu, et al., 2014) on password and text entry process. In their work, they rely on the video recording of the smartphone screen and fingertip from a distance using advanced camera-enabled devices such as Google Glass. They employ advanced image processing techniques to estimate the touched locations on the screen which in turn maps to the actual key presses.

In all three works by Balzarotti et al. and Maggi et al. and that of Yue et al., the cameras directly pointed at the keyboard to capture the video of user’s typing, i.e. direct observation of the appearance of the keyboard or the text typed. There is a series of work which explored the vulnerabilities posed by sensor data on the mobile phone. Owusu et al. (Owusu, Han, Das, Perrig, & Zhang, 2012) used the accelerometer sensor data to infer the password typed on the smartphone screen.

A closely related work to this paper is by Shukla et al. (Shukla et al., 2014). In their paper, Shukla et al. designed an attack on the pin entry process on the mobile phones. The attack was shown to work with the video recording of a user’s hand movement while not compromising any information from the mobile phone screen display. Authors’ show the user’s pin prediction accuracy close to 85%.

Although the attack was shown to predict the smartphone users’ pin with high accuracy, the practicality of the attack is unclear. In this paper, we design a prototype attack similar to the attack by Shukla et al. (Shukla et al., 2014) and evaluate the model under various realistic scenarios.
Fig. 2. Plots showing the effect of various parameters settings on the performance of the attack. Y-axis shows the percent of average accuracy of the Pin prediction and X-axis shows the parameter settings.

EXPERIMENT DESIGN AND EVALUATION

Data Collection

Following approval from our university’s Institutional Review Board (IRB), we collected a dataset of 240 short video clips of Pin entry process from 30 volunteers. 8 pin entry videos were collected from each volunteer participant with different recording settings. Three different video recorders were used: (1) Sony Camcorder with optical zoom, (2) iPhone 6 Plus camera, and (3) HTC One mobile phone camera. Each participant typed Pins on an HTC One phone.

Sony camcorder was used at 6X optical zoom and the mobile phone cameras recorded the videos without any optical zoom. Two different frame recording rates were used 30 fps and 60 fps using Sony camcorder, and iPhone 6 Plus camera. Sony camera was used at a resolution of 1920x1080p, and from 2-10 meters from the volunteers while smartphone cameras used lower recording resolution and recording was done from 2-10 meters from the users.

To record the video of user’s pin entry process, we asked each volunteer participant to select and set the pin on the HTC One phone. Video recorded phone unlocking session was preceded by a practice session in which participants were asked to practice unlocking the phone with the pin for 10 times. We included the practice session before the actual pin entry to get the participants familiarize with the set pin and the phone used in our experiment. All participants were students, faculty or staff at our university.

The following section discusses each parameter and its influence on the pin inference results.

Distance of Recording

We compared the performance of our prototype attack for pin inference on several different values of the distance of recording varying from 2 meters to 10 meters. The optical zoom in the camera was not used (or not varied in case of Sony camcorder) while recording the video to analyze the effect of distance of recording on the performance.

We observed that the performance of designed prototype decreases while increasing the recording distance (see Fig. 2 (a)-(c)). We believe that the quality of video decreases with the recording distance and hence the performance. Since the prototype method uses video object tracking step which is highly dependent on the quality of the video and dissimilarity of the object of interest from the surrounding objects, the
tracking step does not perform well in case of low-quality videos and hence results in lower pin prediction accuracy.

In our experiments, the good performance (i.e. the prediction accuracy higher than 65%) was achieved with the videos recorded from a recording distance between 2-3 meters without using an optical zoom. That being mentioned, an adversary can obtain the similar quality video from a much longer distance by using a more sophisticated camera with high optical zoom capability and achieve equivalent performance. For example, 2 meters will translate to 2×10 =20 meters with a camera having ×10 optical zoom.

**Angle of recording**

The angle of recording is the angle between the line of sight of the camera and opposite of the line of sight of the user's eye (see Fig. 1). Comparison of the performance of our prototype attack model on the videos recorded from various angles showed that the video recordings obtained from an angle of -100° to 0° resulted in an average pin prediction accuracy of higher than 65%. In our experiments, the videos recorded from the angle between 0° to +30° could predict the users’ pin with an average accuracy of 50% to 65%. Table 1 summarizes the results on the angle of recording settings to obtain good (an average accuracy of 65% or higher), acceptable (an average accuracy of 50% or higher), and low accuracy results (see row 3). Figure 2 (g)-(i) shows the effect of change of angle of recording on the pin prediction accuracy.

**Lighting Conditions**

In general, the well-lit room is preferable to record the video for the prototype attack model to work to infer the pin. We performed the attack on the video recordings obtained under four different light conditions; (1) 250 lumens, (2) 500 lumens, (3) 750 lumens, and (4) 1000 lumens. We observed that the prediction accuracy decreases significantly under low light conditions (see Fig. 2 (d)-(f)). We observed that the videos captured under illumination of 750 lumens or higher resulted in an average accuracy of 65% or higher. Also, video clips captured under light conditions with illumination less than 600 lumens could predict the users’ pin with an average accuracy of 50% or less. Table 1 summarizes the results on the light illumination settings to obtain good (an average accuracy of 65% or higher), acceptable (an average accuracy of 50% or higher), and low accuracy results (see Table 1, row 4).

Table 1 Good (an average accuracy of 65% or higher), acceptable (an average accuracy of 50% or higher), and low prediction accuracy (<50%) results under the settings for video recording of user’s typing on mobile phone.

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Parameter Values (Corresponding to Prediction Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distance of Recording</strong> (No Optical Zoom)</td>
<td>Good: &lt; 3 meters, Acceptable: &lt; 5 meters, Low: &gt; 5 meters</td>
</tr>
<tr>
<td><strong>Distance of Recording</strong> (×2 Optical Zoom)</td>
<td>Good: &lt; 6 meters, Acceptable: &lt; 10 meters, Low: &gt; 10 meters</td>
</tr>
<tr>
<td><strong>Angle of Recording</strong></td>
<td>Good: [-100°, 0°], Acceptable: [-100°, +30°], Low: &gt; +30°</td>
</tr>
<tr>
<td><strong>Lighting Conditions</strong></td>
<td>Good: &gt; 750 lumens, Acceptable: &gt; 600 lumens, Low: &lt; 600 lumens</td>
</tr>
<tr>
<td><strong>Recording Resolution</strong></td>
<td>Good: 1080p, Acceptable: As low as 720p, Low: Lower than 720p</td>
</tr>
</tbody>
</table>

**Recording Resolution**

We performed the prototype attack on the video recordings obtained under two different recording resolutions: (1) 720p, and (2) 1080p. We obtained an average pin prediction accuracy of 50% using the video recordings of 720p resolution whereas videos with 1080p recording resolution could predict the users’ pin with an average accuracy of 65% or higher. Table 1 summarizes the results on the light illumination settings to obtain good (an average accuracy of 65% or higher), acceptable (an average accuracy of 50% or higher), and low accuracy results (see row 5).

Although we did not notice much difference in the performance for videos of 720p and 1080p recorded from a short distance (i.e. 2-3 meters), performance decreases for the videos of 720p recorded from longer distances.

**Recording Frame Rate**

We performed the prototype attack on the video recordings obtained under two different recording frame rates: (1) 30fps, and (2) 60fps. Although results obtained from videos recorded at 30fps and 60fps were comparable, we observed that the videos with higher frame rate converge faster in the key touch frame detection. We believe that the higher frame rate reduces the chance of missing the detection of key touched frames in the video as there is more number of the frame corresponding to each key touch, and hence converges faster.

The prototype attack keeps looping in key touch frame detection step in case of missed key touched frames and hence sometimes results in several executions of the same module.
DISCUSSION

Figure 2 shows the effect of lighting conditions, recording distance, and angle of recording on the performance of our prototype attack model for the smartphone users’ pin inference. Fig. 2(a)–2(c) show the effect of recording distance on the performance of our method. Optical zoom capability of the camera was not used to record the videos. The figure shows the prediction accuracy decreases with the distance of the camera from the user. We observed that the attack could infer the users’ pin with an average accuracy of 65% when the recording was done from 2-3 meters. With a more sophisticated camera, an attacker could obtain a similar quality video from a remote distance using optical zoom capability of the video recorder.

Figure 2(d)–2(f) show the effect of different lighting conditions on the performance. The method works better under better light conditions. In general, a well-lit room is ideal to obtain a good video recording. We observed that no significant results were obtained with video recording under low illumination less than 250 lumens.

Figure 2(g)–2(i) show the effect of angle of recording on the performance of our method. We observed that the prediction accuracy was very low when the videos were recorded from the right-front angle from the user. We believe that the reason behind low accuracy is probably the typing behavior of users in our dataset. Most of the users in our dataset typed the passwords using their right hand and hence any part of the palm of the typing hand was not visible. In such scenarios, the attack requires tracking any visible point on the forehead which resulted in a less accurate estimate of fingertip motion due to a relatively long distance from the hand anchor point. The method works better with video recordings captured from the angle between left-front to center-front from the user.

CONCLUSION

To evaluate the practicality of video-based attack model to infer the smartphone users’ pin, we recruited 30 volunteer participants for the study. By performing the analysis on 240 short video clips from these participants of pin entry process obtained in five different recording settings, we show that the attack remains effective however the accuracy is affected by various parameter settings such recording resolution, video frame rate, recording distance, typing behavior, etc. Although the users’ pin prediction accuracy did not show any significant variation with recording frame rate, our analysis shows that given a video clip with higher frame rate an attacker could infer the users’ pin faster.

On the Pin entry process typed on an HTC One phone keypad, we show, the attack could correctly infer 86% of the pins using video recording captured from three meters from the user in first 10 attempts which is a significant reduction in search space compared to random search where the attacker would need to try $10^9/2$ candidate Pins on average (Song, Wagner, & Tian, 2001) before finding the right one.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their detailed comments on the first draft of the paper which helped us improve the final version of the paper.

REFERENCES


INVESTIGATION OF RAPID HARDWARE STUDIES ON ANALOG CIRCUITS FOR DESIGNING FILTER SYSTEMS

Ozgur AKSU
Computer Engineering Department, Erciyes University
{oaksu@uab.edu}

ABSTRACT

Specific area based analog systems are still designed by human experts because automatic analog circuit synthesis tools are available only on a limited range. Developed for filter systems that are based on analog electronic automation system for practically, not possible to use direct analog circuit design. However, performance and technological development of digital systems, but also depends on the development of analog circuits that are the foundation. At this point, this work was prepared, comparison and information obtained through the operation of the system model with the suggestion that full automation of analog circuits, and for the specific purpose of a rapid prototyping system is recommended.

INTRODUCTION

The design of AE (Analog Electronic) circuits is a complicated and hard design problem (Middlebrook, 1990; Toumazou, Moschytz & Gilbert, 2002; Zebulum, Pacheco & Vellasco, 2001; Quentin & et al, 1999). AE design approaches are not considered as having reached adequate level when this problem, which is defined as the Analog Dilemma, is compared with the digital solution approaches (Toumazou, Moschytz & Gilbert, 2002; Quentin & et al, 1999). In order to receive or produce any physical signal, it is absolutely necessary to have electronic circuits, which clearly indicates that there will always be a demand for AE circuits at basic level for such circuits (Middlebrook, 1990; Quentin & et al, 1999; Sarapeshkar, 1998). The EHW automations developed by the researchers gain importance with the solution approach in Analog Circuit design works (Aggarwal, Berggren & O’Reilly, 2007; Toumazou, Moschytz & Gilbert, 2002). EHW ensures that electronic circuits that are beyond the reach of manual design works of humans are designed by using evolutionary algorithms (Koza & et al, 1998; Sarapeshkar, 1998). The EHW Architecture brings Evolvable Hardware, Artificial Intelligence, error tolerance and automatic design systems together (Zebulum, Pacheco & Vellasco, 2001). The Law of Moore defines that the intra-integrated circuit capacity will double in every 18 months (Franz, 2000). In our present day, this law continues to confirm itself, and complicates the design processes with exponential increase. The EHW Fast Prototyping approach provides an important acceleration in traditional AE Design Process with FPAA Solution Model (Schlottmann, Abramson & Hasler, 2012). This acceleration provides indirect solutions in problems to which humans cannot provide direct solutions with their skills and capacities with fast prototyping models. The use of the proposed solution methodology in the design process is given in Figure 1 in the scope of the problem detected.

Fig. 1 Custom Analog IC Design (a) FPAA Design Flow (b) (Schlottmann, Abramson, Hasler, 2012)

In this study, firstly, the AE design process problem is defined. The studies conducted on ADA Fast Prototyping for the defined problem were examined in the literature for the years 1980-2016, and are given in a comparative manner in Table 1. These studies are examined in terms of the proposed solution methods and sample applications on the defined problem in the literature, and the missing points in the process and the problems are highlighted. In this study, as the final item, a new solution model and a sample application are given based on the motivation areas obtained. There are many similar solution offers on the proposed topic and on the solution to the problems detected in similar solution offers in the scope of the GENAN Model. It is targeted that the innovative potential of the developed solution approach on EHW AE automation is revealed again for today’s researchers (Aksu & Kalinli, 2012; Williams, 1991; Aksu, 2016).

PROBLEM DESCRIPTION

ACAD/CAD software and tools have been used actively as of early 1970s (Toumazou, Moschytz & Gilbert, 2002; Platonov, 2014). The solution models used provide solutions that are specific to the target by bringing the proper circuit models and functions together (Koza & et al, 1998; Koza & et al, 1997). In case there are no models or functions for the solution intended to be developed, it is necessary that firstly the subordinated circuit models are developed for the targeted aim (Middlebrook, 1990; Franz, 2000; Toumazou, Moschytz & Gilbert, 2002; Balkir, Dundar & Ogreneci, 2003). The subordinated circuit used today has been produced based on the requirements in the electronic libraries (Toumazou, Moschytz & Gilbert, 2002). Active and passive AE circuit elements are used in order to form a circuit model that will give the targeted aim in the development process (Middlebrook, 1990; Toumazou, Moschytz & Gilbert, 2002; Koza & et al, 1998; Koza & et al, 1997; Sarapeshkar, 1998).
However, the issue of which of the electronic circuit elements that are plenty in number and type will be used for the targeted aim is a basic research problem (Middlebrook, 1990; Toumazou, Moschytz & Gilbert, 2002). In addition to the AE Element Selection problem; The question of which connection structure the circuit elements would be connected to each other and in which order constitutes an extremely difficult engineering problem (Franz, 2000; Ellis & et al, 2002; Middlebrook, 1990; Toumazou, Moschytz & Gilbert, 2002). In order to resolve this problem, scientists separate the frame of electronic science design process into two parts; AE and DE. The basic reason of this separation is developing a new electronic language and library in which it is possible to conduct design/application with basic rules and which has less complicated structure for the purpose of defining and computing the problems of AE (Middlebrook, 1990; Williams, 1991; Nair, 2002). The DE that is developed over the rules is fictionalized completely in a structure that works on binary number system (Middlebrook, 1990; Williams, 1991; Nair, 2002). Although DE works on AE rules and elements, its design and computing processes are in a simpler structure (Williams, 1991; Nair, 2002; Hornby, Sekanina & Haddow, 2008). The EHW viewpoint has been developed in electronic science as a result of the development of DE-targeted automatic design tools and as a result of the artificial intelligence support (Toumazou, Moschytz & Gilbert, 2002; Aggarwal, Berggren & O’Reilly, 2007; Hornby, Sekanina & Haddow, 2008). EHW solution processesand models ensure that humans develop practical solutions and perform automatic design processes in the complicated design process (Toumazou, Moschytz & Gilbert, 2002; Aggarwal, Berggren & O’Reilly, 2007; Hornby, Sekanina & Haddow, 2008). The solution proposal that is obtained with EHW Models require that the solution is renewed as the targeted aim details increase and as more demands for details develop over the result (Toumazou, Moschytz & Gilbert, 2002; Dobkin & Hamburger, 2014; Hornby, Sekanina & Haddow, 2008). The developing details of the solution offer increase the number of the elements of the silicone circuit that constitute the solution in an exponential manner (Toumazou, Moschytz & Gilbert, 2002; Eick & Graeb, 2012; Hornby, Sekanina & Haddow, 2008). The number of the silicone circuit elements that increase as a result of the practical application of SOC Architecture cannot provide adequate development in power consumption and working conditions (Eick & Graeb, 2012; Balkir, Dundar & Ogrenci, 2003). The basic reason of this problem is the lack of an efficient ADA Approach in order to tolerate the DE acceptances (Dobkin & Hamburger, 2014) by AE (Middlebrook, 1990; Hornby, Sekanina & Haddow, 2008). The Black Zone and Hook Up Delays are some of these basic problems in this topic (Dobkin & Hamburger, 2014; Hornby, Sekanina & Haddow, 2008; Balkir, Dundar & Ogrenci, 2003). The problem stems from the increase of the number of the elements of the analogue circuits in an exponential manner due to the increasing DE door/element number without ensuring a development in AE Library Approach on which DE is working (Middlebrook, 1990; Toumazou, Moschytz & Gilbert, 2002; Glelen & Sansen, 1991; Dobkin & Hamburger, 2014). The numbers of the elements that increase in an exponential manner cause that the points that are not considered as errors in theoretical design process. But this expectation return in the form of physical quantity and energy format and as a difference in the temperature because of not being tolerated (Advantage Analog) (Glelen & Sansen, 1991; Sarpeshkar, 1998). When the DE solution development is considered in basic terms, all the physical quantities that are present in the universe gain a meaning at macro level and in an analogue manner (Middlebrook, 1990; Malcher & Falkowski, 2014). In the world in which we live on an analogue axis, it is observed that, firstly, analogue physical data must be converted into numerical form, and the losses and errors must be tolerated and the entry data must be transferred into digital medium in order to develop a digital solution that is proper for the targeted aim (Toumazou, Moschytz & Gilbert, 2002; Balkir, Dundar & Ogrenci, 2003). Then, the digital data that is transferred is processed and interpreted in a manner that is proper for the target, and the result is thus formed (Zebulum, Pacheco & Vellasco, 2001). The digital result data obtained is re-produced as an analogue output by caring for the possible losses and errors (Toumazou, Moschytz & Gilbert, 2002; Balkir, Dundar & Ogrenci, 2003). This defined cycle fictionalizes an extremely long, inconvenient and non-productive work intended for a simple decision-making process (Hall, Twigg, Hasler & Anderson, 2004; Hall & et al, 2005; Lohn & Colombano, 1999). This solution approach that is defined above and that is used in today’s world must be made to become more efficient, which is a necessity for the electronic science (Zebulum, Pacheco & Vellasco, 2001; Malcher & Falkowski, 2014; Sarpeshkar, 1998). For the solution approach, firstly, the AE and DE are examined in terms of the design and production processes, and the results are presented in Table 1 in a comparative manner. The positive and negative sides of AE are summarized below over the data obtained (Aksu & Kalinli, 2010; Sarpeshkar, 1998).

**Advantages of Analog Circuits:**
- Due to simpler structures, circuits have lower power consumption.
- Due to simpler structures, circuits have more flexible working conditions.
- Re-programmable analog integrated circuit design tools exist.
- Lower production costs and production processes are repeatable.

**Disadvantages of Analog Circuits:**
- Model library construction and the mathematical approach are difficult.
- There is a complex relationship among the circuit elements, values, and numbers.
- Some Analog Electronic components cannot be used in integrated circuits.
- There is a complex relationship between component placement and optimization, which creates a problem.
- Library approach is needed to standardize on VLSI Design.

It is suggested that the Gene Law (Franz, 2000), which is known by his own name, is used in the definition of the road work released by Gene Franz in 2000 with the target of...
reaching comparative power consumption for AE and DE by comparing the developments and expectations on digital sign processing as the bases. In Gene Law, the mW/MIPS rate decreases in half every 18 months (Hall, Twigg, Hasler & Anderson, 2004; Hall & et al, 2005). The proposed Gene Law covers and confirms the Moore Law (Ellis & et al, 2002). According to Gene Law, the information is provided claiming that analogue sign processing units will have lower power consumption 20 years ahead when compared with digital signal processing units (ADC-supported) in the power consumption decrease curve for years given in Figure 2.

![Figure 2: Comparison of power consumption trends in DSP (Franz, 2000) and analog chip developed by the CADSP team (Ellis, et al., 2002; Hasler, et al., 2002; Smith, et al., 2002; Hall, et al., 2004)](image)

The confirmation work targeted for the proposed law is confirmed practically by the CADSP team (Hall, Twigg, Hasler & Anderson, 2004; Hall & et al, 2005). The target of less power consumption being achieved in analogue systems 20 years before the achievement of the same target in digital systems created an inclination to AE-based circuit models in many fields, mainly in the sign processing field (Middlebrook, 1990; Smith & et al, 2002; Aksu, 2016).

The FPGAs were developed in order to perform modular production with fast prototyping and to allow the designers confirm practically in terms of DE development process (Aggarwal, Berggren & O’Reilly, 2007; Hornby, Sekanina & Haddow, 2008). When AE is considered, it is observed that a FPGA-like approach is developed (Malcher & Falkowski, 2014; Balkr, Dundar & Ogrenci, 2003). The FPAA was developed for AE for the purpose of practical confirmation and modular production (Aksu & Kalinli, 2012). With the developed FPAA proposal, it is foreseen that similar automation systems may be established like in DE (Malcher & Falkowski, 2014). However, it is observed that AE includes more variables than DE, and the form of simple circuit approach is definitively not in the binary system form, and has exponentially increasing complexity (Koza & et al, 1998; Bennett III, 2000; Koza & et al, 1996). When the design process is considered in terms of calculation-based approaches without practical confirmations, it is observed that the DE Theoretical Calculation Approaches allow that real-time research may be conducted on a separate and continuous time axis (Hornby, Sekanina & Haddow, 2008).

Table 1 Analog Circuit Design Challenges (Aaserud and Nielsen, 1995; Rutenbar, 1993; Goh and Li, 2001; Lohn and Colombano, 1999)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Digital Electronic</th>
<th>Analog Electronic</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circuit Design</td>
<td>With the Binary System and Logic Equations</td>
<td>Fully in the actual physical quantities</td>
<td>Analog Circuit Design is difficult</td>
</tr>
<tr>
<td>Circuit Elements</td>
<td>Logic Gates and Libraries ($10^3$ pieces)</td>
<td>Passive and active discrete components ($\approx$ n units)</td>
<td>Number of Analog Circuit Element is not known even now</td>
</tr>
<tr>
<td>Circuit Size</td>
<td>No any factors within an integrated approach</td>
<td>This is a factor because circuit cannot beget integrated</td>
<td>Analog circuit must be taken into integrated</td>
</tr>
<tr>
<td>Placement</td>
<td>Automatic placement and routing tools are available</td>
<td>Placement and routing criteria include multi-parameter</td>
<td>Analog circuits have more difficult layout problem</td>
</tr>
<tr>
<td>VLSI Design</td>
<td>Is based on a very practical and easy for the library</td>
<td>For there is no any library that covers all elements solution is specific</td>
<td>Analog Circuits need to standardize for VLSI design</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>Logic components contains many analog sub-elements, therefore power consumption is high and the complex structure</td>
<td>For used less and simple circuit element, power requirement is lower</td>
<td>Analog circuits have low power consumption</td>
</tr>
<tr>
<td>Working Conditions</td>
<td>For high speed requirement, working conditions is specific</td>
<td>Operating conditions terms is capable of independent operation</td>
<td>Working conditions of Analog Circuits are more flexible than digital circuits</td>
</tr>
<tr>
<td>Practical Design Tools</td>
<td>FPGA</td>
<td>FPMA, FPTA, FPAA</td>
<td>Integrated design tools for analog circuits has been developed within the dynamically programmed</td>
</tr>
<tr>
<td>Production Costs</td>
<td>Library-based approach for the production large number of sub-circuit unit requires</td>
<td>Less cost-effective because they have less contain circuit elements and less circuit size</td>
<td>Much more suited to the production costs of Analog Circuits</td>
</tr>
</tbody>
</table>

It is also observed that the same situation is not valid for AE (Malcher & Falkowski, 2014). For subordinated circuit models and circuit solutions in AE, many mathematical approaches that are still used today were developed (Dobkin & Hamburger, 2014; Zebulum, Pacheco & Vellasco, 2001; Balkr, Dundar & Ogrenci, 2003). In the developed models, when the circuit size increases, the number of the variable-elements in the calculations and the relations between each other on the time axis also develop (Malcher & Falkowski, 2014; Sarpeshkar, 1998). The requirement for calculation, which shows increase, also increases the processing time in an exponential manner (Toumazou, Moschytz & Gilbert, 2002; Gileen & Sansen, 1991; Dobkin & Hamburger, 2014), (Malcher & Falkowski, 2014). For the solution of this problem, there is various software that were developed over the computer units with high processing and data capacities, which ensure that the results of analog circuits are observed with simulation calculation (Grimbleby, 2000). This software ensures that the result is calculated over the mathematical equations that are proposed and modeled for the AE that is
desired to be simulated (Malcher & Falkowski, 2014). The resolution used during the calculations and the error tolerance values of the calculation models influence the result in a direct manner (Malcher & Falkowski, 2014; Sarpeshkar, 1998). It is observed that the simulation data and the results that are obtained by the practical testing of analogue circuits are not the same, and as the circuit size and sensitivity develop, the error rate reaches incredible levels (Brambilla & D’Amore, 1993). Although the aforementioned analogue circuit design and application problems are reported, design and production engineers still continue to develop analogue circuits in an intense manner and design more productive subordinated circuit models (Balkır, Dundar & Ögenci, 2003; Sarpeshkar, 1998). On the basis of this labor-intensive effort, there is the fact that analogue circuits provide less productive, lower power consuming, wider spectrum and high performance abilities when compared with numerical solutions (Middlebrook, 1990; Hall, Twigg, Hasler & Anderson, 2004; Balkır, Dundar & Ögenci, 2003; Sarpeshkar, 1998).

SIMILAR STUDIES IN THE LITERATURE

Since 1980, the studies in the literature conducted on ADA and AE fast prototyping in the scope of EHW have been examined and a general summary is given in Table 2. The solution models among the ones listed in Table 2 that come to the forefront on the basis of scope and study targets are examined below in detail. In determining the studies, the ADA solution approaches and the confirmation models are used as defining factors (Aksu, 2016).

An Online Evolvable Chebyshev Filter Based on Immune Genetic Algorithm: In the study, a new analogue evolvable hardware platform based on AN231E04 FPAA was proposed and a fourth stage Chebyshev filter was developed. With the solution model, more flexible circuit development is presented using flexible dynamic reprogrammable and function design method. The ANEHP-Alpha model is said to be a high performance realistic evolvable hardware platform. The Immune Genetic Algorithm (IGA) used in the model is used as the main evolution algorithm. It is emphasized that the preliminary evolution and poor solution candidate elimination methods included in the study have achieved high performance and have prevented the candidates from harming the appliances to which they are applied. As a result, the use of the developed ANEHP-Alpha model in the development of analog systems is exemplified by the development process on the filter circuit. Suggested system model in Figure 3, the core circuit model is given in Figure 4 to improve the application results. The system model contains a research and optimization system consisting of FPAA - ADC elements. The proposed model is compared with the simulation-based approach and presented by Chebyshev filter circuit application (Zhang, Li & Liu, 2008).

LAYGEN II—Automatic Layout Generation of Analog Integrated Circuits: The basic definitions given over analogue integrated circuit structures with the name LAYGEN II and the innovative design automation that was created by using evolutional calculation techniques is introduced in the study. The LAYGEN II Design Tool developed solutions with the existing expert data and analogue data library. It is capable of defining the designer circuit size, the technology to be used, and the production type before the actual design. It is possible to perform high-quality layer planning and the placement of circuit elements with B-Tree notation in ADA application targeted for production with intra-integrated circuit; and by using NSGAII Algorithm, analogue circuit structure is fictionalized for more than one targeted aim. The LAYGEN II ADA Application has the internal assessment structure for each VLSI circuit layer. As a conclusion, the analogue circuit solution is presented in the GDSII format for the Calibre Software in the industrial production process. The proposed ADA VLSI system model is presented in Figure 5.
MISSING DETAILS IN THE LITERATURE

When the studies in the literature are examined, more than 20 solution models, which were developed since 1980 for the same targeted solution, are observed. When the relevant models are examined, it is observed that they brought proposals for the solution of AE circuit design problem, and for the solution of ADA design automation solutions. When the approaches developed by the solution models are examined, the following results are concluded;

- Practically, the application and confirmation of the result of the study within the automation system (AE Offset and Validation problem),
- Making the system become parallel for a higher system success (High Calculation Effort),
- Using simulation software in the design process (Pre-emulation),
- Using an artificial intelligence algorithm that is adapted for the purpose (Adaptive Artificial Intelligence), and
- Online assessment of the design result (Online Simulation for total span).

The requirements of ADA, which were determined as five items, are given in a comparative manner with solution approaches in Table 2. A new innovative solution model is proposed in Figure 6 in the light of the problems observed in literature scan by considering the missing points determined and the innovative approaches. No other studies were observed in the literature that proposed the comprehensive and application-focused AE fast prototyping system model solution model.

![Fig. 6 GENAN ADA System Model (Aksu & Kalinli, 2010; Aksu, Kalinli & Tanik, 2012; Aksu, 2016)](image_url)

It is observed that some parts of the proposed system model are included in some studies, and results are produced and released by using the relevant parts that were examined. This information provides confirmation on the requirement of the sub-units placed in the proposed system model, and is selected by considering the system target and ADA automation system requirements. It is targeted that the system is placed in the literature as an authentic model with the joining of the model parts presented in Table 2 and by making them work together to receive results. It is considered that the process parts and steps for the proposed ADA system model being present in similar studies in the literature will not prevent the system model; on the contrary, it will confirm the success of the system. The solution approaches in the proposed system model for the analogue circuit problems are given in Table 1.

### Table 2 Analog Circuit Design Problems and System Model Solution Approach (Aksu, 2016)

<table>
<thead>
<tr>
<th>Detected Future Problems</th>
<th>Suggestion for Solution in the System Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult in terms of basic library and mathematical approach to circuit design.</td>
<td>Using the analog circuits simulation software HSPICE for containing the most recent and innovative solutions with the primary outcome analysis.</td>
</tr>
<tr>
<td>A combination of numbers and values of the possible elements of the circuit element is an infinite number of elements.</td>
<td>A growing set of research space, does higher number research with parallel computing and narrowing the focus of research space on the basis of the practical limitations of the hardware verification.</td>
</tr>
<tr>
<td>Integrated in-house design and manufacturing process must be supported with the SOC approach. But some analog circuit elements cannot be taken into an integrated.</td>
<td>Instead of analog circuit elements, which are not included in the practical verification unit, use corresponding solutions developed in FPAA.</td>
</tr>
<tr>
<td>Placement and optimization approach for circuit, complex relationship between the circuit elements is a problem.</td>
<td>With the working on practical application unit the connection matrix, more complex relationships keep out of the study.</td>
</tr>
<tr>
<td>VLSI Design for the standardization, the library approach is needed.</td>
<td>With storage obtained solutions to database, make VLSI model library.</td>
</tr>
</tbody>
</table>

PILOT APPLICATION: BAND STOP CIRCUIT

Filters are placed in many electronic systems as basic components (Haji Ali, Shaker, & Salih, 2008). Traditional and fast resolutions are fictionalized in the field of analogue structures. However, adaptation of the analogue filter solutions developed for the system and solution-based requirements is an extremely difficult engineering problem. FPAA provides adaptive solutions in this topic with its dynamic concept. The basic difference between adaptive filter and classic filter circuit is the fact that the filter circuit keeps the error coefficient at minimum level by adapting itself on the bases of the requirements that will occur in necessary conditions (Visan & et al, 2010). The microprocessor in the FPAA supports the adaptation of the filter circuit developed with the controlled signal producer. The analogue circuit model used for the targeted aim is given in Figure 7. The comparison of result-quality values are given in Figure 8. S1-S8 Switching Elements change the pass frequency, and thus, the adaptive quality factor of the filter may be changed. The equations given by the FPAA producer company Anadigm Corporation are given in Equation 1 and 2. The switching values in the scope of fostart=10 Khz / fostop=11 Khz value selected for targeted circuit aim are given in Table 3.
\[ f_o \approx \frac{f_c}{2\pi} \sqrt{\frac{C_2 C_3}{C_A C_B}} \]  

\[ Q \approx \frac{C_B}{C_A} \sqrt{\frac{C_2 C_3}{C_A C_B}} \]  

CONCLUSION

There is a need for autonomous filter design and production with green power consumption. In addition to this, the circuit model that will be produced must be enabled to work under hard and various conditions. The aim and scope of this study is to propose a solution model that makes it possible to develop a circuit model in the required targets. Three basic targets are supported with the proposed solution model: adaptiveness, low cost, and wide usage area. As a conclusion, the developed system model fictionalizes a pilot solution model for high-capacity production models for similar aims in terms of high-capacity and critical resolution.

REFERENCES


ACKNOWLEDGMENT

We would like thank Prof. Dr. Murat M. Tanik and UAB ECE in providing support in the department as a visiting scholar and facilitating the usage of the UAB Cheaha grid system.


TRAFFIC SIGNS AUTOMATIC RECOGNITION USING CONVOLUTION NEURAL NETWORK

Rina Komatsu and Tad Gonsalves
Department of Information & Communication Sciences
Faculty of Science and Technology, Sophia University, Tokyo, Japan
{t-gonsal@sophia.ac.jp r_komatsu@outlook.com}

ABSTRACT

Automatic recognition of traffic signs is necessary for self-driving vehicles to drive smoothly and safely. It can also serve as a valuable aid to handicapped and elderly drivers who find it difficult to check traffic signs in a short time and take a corrective action immediately. This study focuses on implementing an automatic traffic signs recognition system using a Convolution Neural Network (CNN), which is fast becoming a standard for image recognition. Using a deep learning framework, this study created three CNN models: an original handcrafted CNN model, and two models from another study which have become renowned for their exceptional performance at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), and compared their respective classification accuracies. The performance of our simple handcrafted CNN model is found to be comparable to that of the two reputed models. It did not show any signs of over-fitting. Moreover, comparing the grayscale and the RGB images classification results, we find that the number of channels does not affect its prediction accuracy.

INTRODUCTION

Recent Japanese automobile research focuses on developing “Autonomous Driving by the application of Artificial Intelligence Techniques” for the 2020 Olympics in Tokyo. If this development is successfully implemented, high merits may be expected such as the support for handicapped and elderly drivers whose physical abilities are on the decline, automatic transportation network for sparsely populated areas and so on.

One of the most basic requirements for autonomous driving is the ability to read and recognize traffic signs automatically. Despite all the significant advances in road sign detection brought by computer vision for driving assistance, it is still a challenging problem. One reason is the extremely varying lighting conditions, namely daytime and nighttime (S. Hamdi et al., 2016), complex backgrounds and different viewing angles (Malik, 2014). For quite some time there was no standard to test the performance of traffic signs recognition algorithms. Fortunately, the German Traffic Sign Detection Benchmark presented as a competition at the International Joint Conference on Neural Networks (IJCNN 2013) has now become one of the standards (S. Houben, et al., 2013).

Support Vector Machines (SVM) and Histograms of Oriented Gradients (HOG) features were the major techniques in the image recognition systems for quite some time (Ardianto, Chen & Hang, 2017; Sugiharto & Harjoko, 2016). Although they showed promising results, the main drawback was the manual extraction of the relevant image features.

Neural Networks or Multi-Layer Perceptrons (MLP) learning algorithms have made great strides and have solved many engineering problems. Their rather shallow learning has been transformed into Deep Learning by an additional set of hidden layers, which has found applications in image recognition. Neural Networks are also successfully used in traffic signs recognition (Ali, Kalis & Safei, 2014; Yamamoto, Karungaru & Terada, 2014). However, the hand-tuned feature extraction necessary for Neural Network computation has several disadvantages – it needs expert knowledge, is laborious and time-consuming, and does not generalize to other domains.

The latest research in image recognition is through Convolutional Neural Networks (CNN). These nets are modeled after the image processing activity in the visual cortex of the mammalian brain. Preservation of the spatial relationships in the data, simple structure, shared weights, fewer learning parameters, and adaptability are some of the factors that make CNNs suitable in image and voice recognition applications. In MLPs, the flattening of the 2D pixel information of the image into a 1D vector loses most of the spatial information. The CNNs preserve the spatial relationship between the pixels, since the 2D pixel information is directly fed into their input layer.

CNN methodology is fast becoming a standard in image recognition and, in particular, in traffic signs recognition (Hu, et al., 2017; Jin, Fu & Zhang, 2014). AlexNet and GoogLeNet CNN models have become renowned for their exceptional performance at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). We have designed an original handcrafted CNN model to learn to classify traffic signs. This paper reports the results of a comparative study on the implementation and performance of the above two models and our original handcrafted CNN model.
implementation is through the chainer framework freely available as a Python module. Our handcrafted models outperforms the ILSVRC in the training results on the German Traffic Sign Detection Benchmark. In the test results, its performance is comparable to the ILSVRC models.

This paper is organized as follows: Section 1 introduces the subject matter of this study; Section 2 describes the traffic signs recognition implementation; Section 3 does a brief CNN overview; Section 4 describes our comparative study on the three CNN models and Section 5 presents the experimental results; Section 6 concludes the study.

TRAFFIC SIGNS RECOGNITION

Conventionally, traffic signs recognition is done using Multi-layer Perceptrons (MLP) (Lorsakul & Suthakorn, 2007). The MLP method of recognizing traffic signs from landscape pictures usually follows the following 4 steps (CNN method follows along the same line):

**Step 1: Image Extraction**

Get landscape images by means of a video camera mounted on the self-driving vehicle.

**Step 2: Sign Detection and Extraction**

Extract the area in the picture likely containing the traffic sign and generate a small image.

**Step 3: Form Recognition**

The small images are converted into a set of real-value parameters which serve as an input to the Neural Network. When using chainer as the framework for deep learning in Python, the real-value parameters need to be set as type numpy.float32, while the integer-value parameters need to be set as type numpy.int32 (Shino, 2016).

**Step 4: Traffic sign Recognition**

Input the parameters created in step 3 to the Neural Network and classify the small image. This study uses CNN to learn to classify images. The 4 steps leading to traffic sign recognition are depicted in Fig. 1. Consider the “stop” traffic sign image picked up by the video camera along with the background. The traffic sign is extracted from the picture, pre-processed into a set of real-valued parameters and fed into CNN, which identify the traffic sign as “stop”.

![Fig. 1. Traffic signs classification using CNN](image-url)

CNN OVERVIEW

Neural Networks, in general, imitate the function of biological neurons whose excitation signals can be on/off depending on the strength of the input stimulus. The well-known Multi-Layer Feedforward Neural Network which includes the combination matrix operation and nonlinear activation function performs quite well in solving classification problems. For example, the error in classifying the 0-9 digit images in the MINST dataset is as low as 3% (Glorot & Bengio, 2010). However, in the classification of the CIFAR-10 dataset which has ten classes of images such as dogs, birds and so on, the test error increases to about 50%. The poor performance of MLPs in classifying images has given rise to a new kind of a Neural Network, called Convolutional Neural Nets (CNN). A CNN, for example, with three convolution layers interlinked with max pooling layer showed only 16.6% test error in classifying the images in the CIFAR-10 dataset (Hinton, et al., 2012).
Fig. 2 How to classify images using CNN

Fig. 2 describes the structure of CNN used in this study. First, the values of the input image are convoluted by Convolution layer to extract the features. Next, the Pooling layer thins out the convoluted values to extract the strongest feature. After passing through many combined layers consisting of Convolution and Pooling, the values are output as channels which are different from the RGB channels. These channels are input to the fully connected forward propagation type neural network. The outputs from the neurons in the output layer become the elements to classify the image fed into the units of the CNN input layer.

COMPARATIVE STUDY

We have made a comparative study of three different CNN models to learn to recognize standard traffic signs: our original handcrafted CNN model, AlexNet and GoogLeNet. Each of this model is described below.

Model 1: Handcrafted CNN model

Fig. 3 describes the structure of the handcrafted CNN model. First, the input images are resized into 100x100 pixels and are fed as input to the CNN model.

Model 2: AlexNet

AlexNet is a CNN model created by Alex Krizhevsky. (Krizhevsky, et al., 2012). This CNN model has 5 Convolution layers, 3 Pooling layers and 3 fully connected layers. The salient feature of this model is that using RELU as activation function in the fully connected layer, it can prevent gradient loss which is generally the cause of low quality machine learning.

Model 3: GoogLeNet

GoogLeNet is the CNN model created by Christian Szegedy (Szegedy, et al., 2014). The depth of the layers is 39. Its specialty is the inception module containing sets of convolution and pooling layers.

AlexNet and GoogLeNet are used to classify images belonging to 1,000 different classes. However, since our study used more than 1,000 classes, each of the CNN models is modified at the fully connected output layer to accommodate all the classes.

EXPERIMENTAL SETUP

1. Creating images for Training and Testing

This study tried using “chainer” as the framework for deep learning. The Python chainer framework (Shino, 2016) is used to build the program which learns from the training data and the various parameters of the function which is used in the Neural Network.

The traffic sign data is from the German Traffic Sign Detection Benchmark (http://benchmark.ini.rub.de) which includes landscape images, and documentation about the coordinates of the traffic signs and their identification. Using
this data, this study tried classifying 43 classes of traffic signs.

After extracting the area of the traffic sign from the landscape image, image processing such as adding noise to the traffic sign image and changing the contrast is done. This is because Okatani (2015) points out that learning will not be successful if the training sample is small. Image processing is an effective way to obtain quality learning results by creating a large and varied images dataset. There are 500 images corresponding to a single traffic sign. Accordingly, the entire collection of $43 \times 500 = 21500$ images is split into a training and a testing sub-set. Further, each sub-set has two categories of images: grayscale and RGB.

2. Training

Each pixel value in the images of the training dataset prepared in phase 1 is normalized. To prepare for training, we created small sample groups called mini batches from the training data. Using these mini batches, training was held through the cycle of forward propagation calculation and back propagation error correction. The number of epochs used in training was set to 10.

3. Evaluating the Accuracy through Testing

Using the images in the test dataset as input to the CNN learned model, we evaluated the accuracy rate from the output values using the softmax cross entropy function.

RESULTS

Fig.4 shows the traffic signs classification accuracy in the training phase of the grayscale images, while Fig. 5 shows the accuracy in the training phase of the RGB images.

<table>
<thead>
<tr>
<th>Table 1 Classification accuracy in the Training phase of grayscale images (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handcrafted CNN</td>
</tr>
<tr>
<td>97.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2 Classification accuracy in the Training phase of RGB images (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handcrafted CNN</td>
</tr>
<tr>
<td>97.21</td>
</tr>
</tbody>
</table>

In both the training cases, our handcrafted CNN model reached 90% classification accuracy as early as 4 epochs of training. This is faster than AlexNet as well as GoogLeNet. Moreover, the overall accuracy of 97.09% achieved by our handcrafted model is superior to that of the other two reputed models. In conclusion, the handcrafted CNN model performs better in training for the classification of grayscale as well as RGB images.

Using the CNN learned model, we evaluated the classification accuracy in the test phase. Table 3 shows the classification accuracy of our CNN model in the test phase of the grayscale images. It is comparable to those of the AlexNet and GoogLeNet models.
Next, Table 4 shows the classification accuracy of our CNN model in the test phase of the RGB images. Here, too, it is comparable to those of the AlexNet and GoogLeNet models.

<table>
<thead>
<tr>
<th>Handcrafted CNN</th>
<th>AlexNet</th>
<th>GoogLeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>93.77</td>
<td>95.90</td>
<td>94.44</td>
</tr>
</tbody>
</table>

To sum up, although our handcrafted CNN model is the simplest among the three deep learning models that we implemented in our comparative study, its high prediction accuracy proves that the model is competitive. Moreover, comparing the grayscale and the RGB images classification results, we find that the number of channels does not affect the prediction accuracy.

DISCUSSION

In this comparative study, CNN was able to get over 93% accuracy in classifying 43 kinds of traffic signs even though each of the input images is different from one another due to preprocessing like rotation, addition of noise, change in contrast and so on.

Why do we get such excellent results? According to Okutani (2015) the secret lies in the Pooling layer of the CNN. The Pooling layer makes the outputs thinned out and invariant despite the translation in space of a given feature through image processing. As a result, each of the CNN models which included Pooling layer could catch the features from the images invariant without over-fitting.

Comparing the results in Table 1 and Table 2, the number of channels did not affect the classification accuracy. This would imply that the traffic signs can be detected without relying on color. For instance, the gray image of speed limit 30 is perfectly understandable only if we could see the value 30 (and not the surrounding coloration).

In this study, the traffic sign image was extracted from the landscape image in advance by providing the relevant coordinates. Another effective method of detecting the area is by extracting Red, Blue and Yellow colored regions (Ruta, Li & Liu, 2010). Equation 1 gives the strength of the colors: Red: $f_R(x) = \max(0, \min(x_R - x_G, x_R - x_B) / s)$, Blue: $f_B(x) = \max(0, \min(x_B - x_R, x_B - x_G) / s)$ and Yellow: $f_Y(x) = \max(0, \min(x_R - x_B, x_G - x_R) / s)$.

The OpenCV library for programming supports Hough Circle Transform function to detect circular shapes as in a typical traffic sign; however, this function depends on the manually set parameters (Bradski & Kaehler, 2013). The above study also introduces the Cascade method for classification. Cascade classifies the object as likely positive (this object means traffic sign) or as likely negative (this object means part of car, pole, etc).

As for the computational cost, despite using a high spec GPU, it took over 35 minutes to train the CNN model using 100 epochs. To learn the traffic signs from massive real-life data sets, the computational time will be colossal. We suggest using a series of high spec GPUs to handle the traffic signs recognition deep learning task in a reasonable amount of computational time.

CONCLUSION

Automatic traffic signs recognition system is handy for elderly drivers who may have a difficulty in recognizing traffic signs as they steer on. Secondly, such a system is a must for future self-driving vehicles. In this study, we compared the classification accuracy of three different CNN models in recognizing traffic signs. The dataset employed was the German Traffic Sign Detection Benchmark. The performance of our simple handcrafted CNN model is found to be comparable to that of AlexNet and GoogLeNet, both of which have acquired a reputation for their outstanding performance at the ImageNet Large Scale Visual Recognition Challenge. Moreover, comparing the grayscale and the RGB images classification results, we find that the number of channels does not affect the prediction accuracy.

Our current method consumes a lot of time extracting the area from the landscape images with the likelihood of traffic signs. In our future research, cascade is a promising candidate for extracting images from landscape pictures. The biggest challenge, however, is to train the CNN themselves to extract the area of traffic sign automatically and identify the traffic sign without any external aid.

REFERENCES


FEYNMAN’S DRUMS: A STEM EXERCISE ON PERMUTATIONS

Michael Lipscomb, William Monroe, Karthikeyan Lingasubramanian
UAB, Department of Electrical & Computer Engineering
{mml005@uab.edu, wsmonroe@uab.edu, klinga@uab.edu}

ABSTRACT

This paper sets out certain efforts to bring advanced engineering concepts to K-12 students in the context of STEM education. The current proposal involves the use drumming in the classroom to give students, quite literally, hands-on experiences with topics relevant to engineering.

The paper begins with a short explanation of musical notation. It is next proposed that the patterns of non-Western drumming can be understood in terms of set theory, permutations, complex systems, communication theory, and DNA coding. From this, students can be lead to present such patterns in the form of mathematical notations applicable to each topic. Next is set out how classroom drumming experiences can be applied to the topic areas. Finally, the use of Mathematica is suggested with examples given.

INTRODUCTION

Sets, permutations, systems and communication are important topics, particularly in the domain of electronic and computer engineering. As such, these topics would seem to be a natural fit for STEM education. Yet the abstract nature of the material can make it difficult to approach for some students, particularly those who have not had a history of excelling in math. In this paper is presented a proposed presentation solution for STEM education.

The authors are engaged in ongoing STEM education research for the development of curricula, materials and teaching tools for K-12 classrooms. Current attempts to bring engineering knowledge to K-12 students, as can be found in the Next Generation Science Standards, tend to focus on the daily practices of engineers (National Science Teachers Association, 2013).

Previous work by the project team involved combing this approach with others to develop a free-standing engineering course for high school students. In the material engineering was portrayed as systematic problem solving (Lipscomb, 2015). More recent efforts have included presenting engineering as a transdisciplinary design and process science (Lipscomb & Lingasubramian, 2016). Such work has been influenced by the core concepts of SDPS.

The purpose of the instant paper is to investigate how advanced topics in engineering can be delivered to K-12 students. Herein, the focus will be on permutations, complex systems, communication theory, and DNA coding. Student participation in drumming patterns is used as a teaching tool. These experiences can lead to the employment of different forms of mathematical representation, and the use of Mathematica.

A different approach

It is not uncommon to present difficult topics by relating them to something familiar, and liked, by students. Yet analogies can be stretched to the point of nonsense, and dressing up a topic is merely decoration. In chess instruction, for example, it might be possible to hold a student’s attention longer by using pieces that look like cartoon characters. Yet in the end, it’s still chess. This is the “science is fun!” approach.

The project team proposes that this model be flipped. What if the things students liked were not the dressing but the substance? Instead of turning instruction into games (gamification), we could turn games into instruction. Instead of trying to put the fun into science, it is revealed that science is in the fun. The present project presents music not a costume but as the raw material from which engineering concepts are explained.

The ideas presented herein are simple, and perhaps because of that, easy to dismiss. Yet engineers know that simple components can have big impacts. Richard Feynman famously demonstrated this when he pointed an accusatory finger to the common O-ring for the 1986 space shuttle failure (Feynman 1988).

Dr. Feynman is known for his accomplishments in physics. He is also known for his eccentricities. One was his love for drumming. According to various accounts, including his own, Feynman was quite proficient with various hand drums. He was particularly adept at identifying drum patterns (Feynman, & Leighton, 1985).

Without an explanation from the man himself, one can only speculate as to why world rhythms were so fascinating to him. However, as will be shown, connections can be made from non-European rhythmic patterns to science and engineering. Perhaps this was a hook that lured Feynman into drumming. This hook can be the basis for exercises in engineering STEM education.

MUSICAL NOTATION

To set out the project, it will be helpful to provide a brief explanation of musical notation. Traditional western musical notation is comprised of symbols and words placed on and near a set of five parallel lines called a “staff”. Symbols called “notes” can indicate the pitch, duration and timing of sounds to be performed on a musical instrument.

For the purposes of the instant project, drum music notation will not need to represent pitch or duration. Instead,
notes or symbols will represent timing and what type of drum hit, or “stroke”, is required.

Players of hand drums are capable of producing a number of sounds by means of different hand gestures. Through these techniques, a player can produce, for example, a bass tone, an “open” tone and a slapping sound on a hand drum. In Western drum notation, the vertical placement of a note on the music graph indicates what type of stroke is called for.

The horizontal position of a note symbol on the staff indicates timing: the point during a standard cycle of time (“bar” or “measure”) that a drum stroke should be made. Bars of music are delineated by vertical lines on the graph. Bars can have different time lengths, counted in “beats”. Presented below is an example of Western drum music notation (Dwosky, 2012):

![Fig. 1 Standard drum notation](image1)

The relevance of this notation to set theory is that each drum stroke can be understood as an element of a set. When drum strokes are set out in a sequence, as they are in musical notation, they can be seen as a permutation of a set of drum strokes. The appearance of each distinct element can be heard as it is performed, and can be read in the notation.

While Western musical notation has existed for hundreds of years, and is widely understood by musicians, some experience is needed to read it effectively. This is unwelcome in a STEM exercise intended to direct students toward engineering topics, not music theory. Further, such notation does not readily lend itself to representation as permutations.

The project team’s solution was to use an alternate, simplified notation for hand drums created by Alan Dworsky (2000). In this notation, bass tone strokes are represented by asterisks, open tones by capital “o”s, and slap tones by triangles. A beat in which no stroke is made (“rest”) is represented by an empty box. In terms of set theory, this notation results in a $N$ set of 4 elements.

Drum music often involves a short pattern of strokes that repeat in a cycle. This can be seen in Sunguru Bani, celebratory drum music from Mali, West Africa. Sunguru Bani is comprised of three such patterns (“parts”) played simultaneously by three drummers.

Presented in the next figures are the three parts of Sunguru Bani as represented in Dworsky’s notation. Brackets delineate the beginning and ending of a repeated pattern (“phrase”). For purposes of permutation representation, a rest used to separate repetitions will not be considered an element of the permutation.

![Fig. 2 Sunguru Bani part 1](image2)

![Fig. 3 Sunguru Bani part 2](image3)

![Fig. 4 Sunguru Bani part 3](image4)

By performing these drum parts, students can gain an experiential encounter with permutations. Yet two problems for implementation exist: technique and equipment. The ability to produce different tones on a hand drum takes practice. Although such may be enjoyable for students, it would encroach on the time available to provide engineering instruction. Further, the ability to provide a hand drum to every student might not exist.

Fortunately, neither technique nor equipment is necessary. Students can perform drum patterns by using their hands to strike body parts. This allows for the cost-free production of different sounds to distinguish events in a pattern. Conversion is a simple matter of representing the drum strokes as body strikes. In the instant project they are a thigh slap, foot stomp and clap. The following notation was created by Alan Dworsky (2002):

![Fig. 5 Sunguru Bani part 1](image5)

![Fig. 6 Sunguru Bani part 2](image6)
APPLICATION TO SET THEORY

Now it can be explained how drum patterns can be connected to set theory and permutations. It can be said that the Sunguru Bani patterns are permutations of $N$ taken $K$ at a time, with repetitions, where $K$ equals 5. The rest taken between repetitions of the permutation is not counted as part of the permutation. In Augustin-Louis Cauchy’s two-line notation, Sunguru Bani pattern 1 for hand drums would be represented as follows:

```
O  *  Δ  □
*  O  O  *  Δ
```

Pattern 2 would be represented as follows:

```
O  *  Δ  □
Δ  *  O  O  *
```

Pattern 3 would be represented as follows:

```
O  *  Δ  □
*  O  O  □  Δ
```

As explained above, drum music usually involves a short pattern of strokes that repeat in a cycle. Thus, drum music could be represented in cyclic notation. Pattern 1 would be represented as follows:

```
*  →  O  →  O  →  *  →  Δ
```

It is a simple matter to represent these permutations in standard, numerical notation. Let $1 = O$, $2 = *$, $3 = Δ$, and $4 = □$. The patterns of Sunguru Bani would then be represented as follows:

```
1 2 3 4
1 0 1 0 0
2 1 0 0 0
3 1 0 0 0
4 0 1 0 0
5 0 0 1 0
```

Similarly, pattern 1 of Sunguru Bani can be represented as a circulant matrix as follows:

```
2 1 1 2 3
3 2 1 1 2
2 3 2 1 1
1 2 3 2 1
1 1 2 3 2
```

APPLICATION TO COMPLEX SYSTEMS

Polyrhythms such as Sunguru Bani possess a distinguishing characteristic that suggests another, non-obvious, educational opportunity. When individual permutations are repeated and played simultaneously, new patterns emerge. This effect is highlighted by three factors: different tones are produced by different drum strokes; patterns do not necessarily begin at the same time; and rests occur at different times in the patterns. This can be seen in Sunguru Bani. Different strokes produce different tones. Pattern 1 begins on the first beat, while patterns 2 & 3 begin on the third beat. When the players of parts 1 and 3 are resting in beat 6, the player of part 2 is producing an open tone.

A performance of Sunguru Bani is a demonstration of the emergent pattern effect. This can be readily seen in a graph combining all three patterns (Dworsky, 2000).

```
1 2 3 4
1 0 1 0 0
2 1 0 0 0
3 1 0 0 0
4 0 1 0 0
5 0 0 1 0
```

Similarly, pattern 1 of Sunguru Bani can be represented as a circulant matrix as follows:

```
2 1 1 2 3
3 2 1 1 2
2 3 2 1 1
1 2 3 2 1
1 1 2 3 2
```

Encoded within Sunguru Bani is the polyrhythm known as “3 over 2”. A pulse of 3 is created by the three evenly spaced slaps (Δ) in each bar. This pulse emerges as a cumulative effect of the patterns. A pulse of 2 is created by the two evenly spaced bass strokes (*) that appear in beats 1 and 4 of parts 1 and 2.

The polyrhythm of 3 over 2 can be more readily seen in a graph representing it in isolation. In this form, 3 over 2 can be performed by two students making a clapping sound at the symbol “x” (Dworsky, 2002).
APPLICATION TO COMMUNICATIONS THEORY

There is another aspect of drumming that has engineering STEM value: drums can be used to communicate. This fact was noticed by Europeans in the early Eighteenth Century. For them it was a topic of interest and concern (Gleick, 2011), (Feldman, 2006).

Drum communication is not an attempt to convey messages by means of a coded alphabet. Instead, the vowels and rhythms of the words are transmitted. Many African languages are tonal, meaning that the pitches of vowel sounds have significance for the meaning of words (Odden, 1995). The pitch of drum tones can represent the pitch of vowels. Drum communication, then, is a reproduction of the tonal pitches and rhythmic patterns of speech (Arom, 2007).

The drumming of the Banda is instructive. The Banda are the main ethnic group of the Central African Republic. Their language, like many tonal languages, contains three pitches: low, medium and high.

The Banda use log drums to communicate quickly over significant distances. The hollowed-out logs have convex sides of different thickness allowing for the production of two tones. Logs of different size can produce different pitch sets. To communicate a message, two log drums are used. One is struck to produce the low and medium pitches, while the other is struck to produce the high pitch. Messages are conveyed through a succession of drum strokes separated by pauses. To make reception easier, there is much redundancy (Arom, 2007).

Of relevance to STEM engineering is the fact that tonal-language drum communication can be expressed in terms of set theory. In the system of the Banda, \( N = 3 \) taken \( K \) at a time with repetitions. \( K \) will equal a range of numbers corresponding to the length of a given message.

The idea that permutations can be used as coded messages gives the STEM instructor a way to move into a discussion of Shannon’s communication theory. Next, an exercise could be presented for representing drum transmissions as communication channels and error content graphs. Presented in figure 17 is an error content graph (also known as the polygonal representation of a communication channel) for the Sunguru Bani pattern.

APPLICATION TO DNA CODING

In the K-12 Next Generation Science Standards, engineering is presented as an embedded component of standard science subjects. Therefore, those who wish to present engineering concepts to American students will need to relate engineering to science course topics. The project team has made efforts to convey engineering design in the context of 9th grade biology (Lipscomb & Lingasubramanian, 2016). As will be shown, engineering concepts can be related to DNA through drumming.

A figure combining the three patterns of Sanguru Bani was presented \textit{infra}. It was used to suggest emergent qualities of complex systems. This same figure can be used in a discussion of genetic coding. Each column of the graph can represent a unit of genetic instruction – the codon. The progression of the music, then, can represent genetic sequencing.

A brief discussion will suffice to make the point. There are four letters of the genetic alphabet: A, C, G & U. They
always combine in threes, called a codon. A codon is a single word, or message, that takes the form of either an amino acid or a “stop codon”. Genetic instructions are a permutation of codons. It is known that codons are set out in a string, and that the order of codons matters. All living beings use this code to synthesize proteins.

There are 64 codons in the codon set. In other words, there are 64 permutations of the set [A, C, G, U] taken 3 at a time. Differences between codons can be described in terms of Hamming distance. For example, the difference between AAA and AAC is 1. This is true because there is one position change. This is the Hamming distance. Hamming distances and their geometric representation are tools of mathematical code analysis. This has relevance to code error detection and correction.

As was shown by Lenstra (2015), a mathematical model for analyzing the patterns of codons can be created. Although it is unknown what program Lenstra used to create the model, it can be done in Mathematica, similar to the way in which drum patterns were graphed.

Further, we can use the same set and information theory approaches to analyze DNA structures. Thus bridging the gap between the abstract concept and practical application. Presented below is an analysis of the “codons” produced by the Sunguru Bani musical notation using Mathematica.

Mathematica to produce the music of Sunguru Bani is similarly straightforward.

CONCLUSION

The project team proposes that instead of turning instruction into games, we can turn games into instruction. Drumming provides an opportunity for students to experience and get meaning from abstract concepts such as sets, permutations, systems, communication, mathematical representation and computing. Working with different mathematical notations provides an opportunity to present to a significant meta concept: the value of different representations is that each offers a another toolset for answering a question or solving a problem.

The ideas presented herein are simple. Yet the development of simple approaches for presenting complex material could have real impact on the life of a student. It could be the difference between a job in the service industry and a career in engineering.

ACKNOWLEDGMENT

Grateful acknowledgement is given to Murat Tanik, PhD for his significant contributions to this paper.

REFERENCES


ABSTRACT

In order to increase safety in human-robot interaction, it is important to find reliable methods and approaches allowing better detection of people and objects in the robot environment. One research area concerns the development of proximity sensing technology. This paper intends to address this challenging requirement by applying a machine learning model specifically dedicated to a new kind of capacitive tactile proximity sensor (CTPS) that has been developed by the Intelligent Process Automation and Robotics Lab (IPR-KIT) in Germany. Our research study focused on using machine learning approach to infer from the data collected by the sensor array in order to insure a safer human–robot interaction. To achieve our goal, we built a classifier and a regressor on the projected distance between objects and the sensor.

INTRODUCTION

With the increasing demand for new flexible robot based production concepts, close cooperation of humans and robots seem to be a promising strategy. This requires new sensor technologies to allow the robot the sensing of humans or obstacles nearby and correspondingly guarantee the safety respectively and avoid collisions. Tactile sensors are one interesting sensing modality, but they recognize a collision “too late” as a contact has to be established. Proximity sensors can specifically solve this issue.

At IPR ("Institute of Anthropomatics and Robotics (IAR) - Intelligent Process Automation and Robotics (IPR)." 2017) a versatile and modular capacitive tactile proximity sensor has been developed. The sensor combines a tactile and proximity sensing modality which supports the idea of more autonomous and safer robotic systems in their interaction with humans and the environment. The combination of tactile and proximity sensing closes the perception gap in the classic vision based sensing, which suffers from occlusion (Navarro, Hein, & Wörn, 2015).

But these sensors come also with some downsides. The main challenges with CTPS is their dependency on the type of the objects when sensing the proximity information, it behaves differently to different object size, grounding state and material types, (metal, wood, humans, etc.) (Navarro et al., 2013). Currently, the pure analytical methods used for the environment recognition are quite complex and they are making the modelling of the different objects a very laborious endeavor. Therefore, the focus has been set on the use of machine learning as an attempt to help identify different objects parameters during proximity sensing and learn the information from the sensors generated datasets from different objects.

The rises of machine learning techniques surely based on the explosion in computing power nowadays has allowed to tackle a wide range of complex problems. It is possible to gather and store the huge amount of data needed for the algorithms which made machine learning accessible for all kind of domains.

To our knowledge machine learning (ML) algorithms have currently not been investigated to process the data of proximity sensors. Our research focuses on investigating ML algorithms, specifically neural networks, for data processing of the proximity sensors. Distance estimation is targeted in the study as a start. Other parameters determination will be investigated in future work. We consider the problem as a classification and regression problem, which reflects supervised learning or learning from labeled data where desired outputs are fed to the neural network. The backpropagation technique will be used to train the network; which is broadly known as a powerful tool in artificial neural networks (Dai & Liu, 2012; Werbos, 1990).

CONTEXT OF STUDY

For the experimental set-up to generate real data, we used two CTPSs each having a special resolution of 2x4 electrodes. The used CTPS are a revised version of the technology presented in (Alagi, Navarro, Mende, & Hein, 2016). The multi-modality refers to tactile and proximity measurements. The tactility is measured by determining the distance changing between two electrodes stacked over each other (top and bottom electrode) when force is exerted at the electrodes (Figure 2). The proximity measurement could be performed in two modes: self-capacitive and mutual-capacitive modes. In self-capacitive mode, the Top electrode is driven with exciter signal generating an electrical field, which is at the bottom side, actively shielded by the bottom electrode. In the mutual-capacitive mode, the top electrode receives a generated exciter signal; here the receiver electrode is also shielded actively by the bottom electrode. In both modes, the distortion of the electrical field by an object in the near field of the sensor will be interpreted as proximity signals. The relationship between the distance x and the
recorded proximity signal by the sensor is inversely proportional and can be roughly approximated by $ax^{-b} + c$, where $a$, $b$ and $c$ depend on the geometry of the electrodes and the settings of the measurement units for a known object.

Figure 1 shows one sensor board with the controller and the analog/digital converter which reads 8 electrodes in a time multiplexed way.

Figure 1: Sensor board (controller and ADC) and 8 sensor electrodes

Two of those sensor boards and sensor electrodes are stacked together to build a 4x4 array.

Figure 2: The 4x4 matrix of a sensing element for the CTPS

This array is then attached to an end-effector, which is mounted to a 6DoF lightweight robot as shown in Figure 6.

Often, machine learning is considered as a collection of methods and algorithms that are used to create prediction models from a priori known data, though, machine learning experts identify it as a fertile field with very extensive themes and patterns, it provides great flexibility benefits.

Most machine learning algorithms are distinguished as three different types (Alsheikh, Lin, Niyato, & Tan, 2014):

1) supervised learning; where labeled training set (i.e., known outputs) is used to build the learning model, it serves to solve classification or regression problems.

2) unsupervised learning; where labels are not provided (i.e., no output vector), this category serves for clustering the sample sets to different groups (i.e., clusters) by finding similarity between input samples.

3) reinforcement learning algorithms; where the agent learns by interacting with its environment, takes an action and gets a reward. The goal is to receive a maximized reward.

Figure 3: Measurement curve in send mode of the (object-electrode qE00)

METHODLOGY

Data acquisition

a. Simulated data

To verify and evaluate the approach, in the early stage of the research, simulated data was generated based on the function $ax^{-b} + c$ which is used to simulate sensor values in relation to the distance of the object. 1500 different object positions were taken in a space box of 8x8x15 cm below the sensor pad. The input vector contains 16 values. The benefit of the simulated data is being free from any noise. So, we focus on first to set up an appropriate neural network model.

b. Real data

Experimental set-up is shown in figure 6. The data was recorded in an area of 8x8x15 cm below the sensor pad. The input vector contains 22 values (OEs+qEs) as illustrated in Figure 5.

Figure 5: Data structure
The robot was moved horizontally and vertically as shown in figure 7a and 21969 data samples were generated.

Figure 6: The Experimental set-up

Figure 7a: Movement of the sensor

Figure 7b: Collected proximity information (object-electrode qE00)

Machine learning process

![Diagram of the machine learning process]

Preprocessing

Cross-Validation

The success of machine learning algorithms is dependent on the quality of the data they operate on. Usually raw data may contain irrelevant information, noise, missing or inadequate values which make machine learning algorithms produce less accurate or not clearly consistent results. They may also fail to discover anything of use. In order to resolve these problems data preprocessing is an important step in the machine learning process (figure 8) (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Both simulated and real data went through the preprocessing process.

First, data was labeled based on the coordinate of z axis (Oz), into 15 classes for both simulated and real data, as for real data class 15 was removed to keep the number samples relatively similar to preserve the balance of the data. The remaining total is 18246 data samples. As for regression problem Oz coordinate is used as a target of the learning problem.

Normalization is an important process for neural network algorithms (Sola & Sevilla, 1997); Z-score normalization (Standardization) was performed where the features are scaled so that they have the properties of a standard normal distribution.

\[ Z = \frac{x - \text{mean}}{\text{std dev}} \]

Feature selection (FS) refers to the process of opting relevant features and removing irrelevant or redundant ones (Kotsiantis et al., 2006). FS helps enhance the performance of machine learning by identifying prominent features.
During this study manual selection was performed based on some experiments and analysis of results. Mainly choosing of the input vector size.

Data was shuffled and split into training and testing sets, cross validation technique (Kohavi, 1995) (Shao, 1993) was performed with 5 folds \(k=5\), each time, one of the \(k\) subsets is used as the test set and the other \(k-1\) subsets are combined to form a training set. The training algorithm has to run from scratch \(k\) different times. Every data point has been in a test set once but \(k-1\) is a training set. In the end, the average root mean squared errors was computed across all \(k\) folds.

Early stopping is used when results does not improve for a pre-specified number of epochs (Prechelt, 1998).

**Artificial Neural network**

Artificial neural networks (ANN) take input and transform them by a series of hidden layers. Figure 9 shows a feed-forward neural network with 3 hidden layers. Each hidden layer consists of a set of neurons, each neuron is fully connected to all neurons in the previous layer. The input layer neurons take the input vector, in our case up to 16 neurons in simulated data \(OE_s\) and up to 22 \(OE_s+qE_s\) in real data due to the measurements of the sensor combined pixels.

![Figure 9: feed-forward neural network model](image)

The open source framework ("Tensorflow.") was used to build and train the neural network. Additionally, other open source APIs were utilized in our application including Numpy, Pandas, Scikit-learn, and Matplotlib.

**RESULTS AND EVALUATION**

Through some trial and error, the architecture of the machine learning algorithm ended up in a feed-forward backpropagation network with 16 inputs, 14 outputs for the real data models, 15 outputs for the simulated data, and 3 hidden layers with 80-40-20 neurons respectively. All hidden neurons used the rectified linear unit \(relu\) activation function (LeCun, Bengio, & Hinton, 2015), the network was trained using backpropagation (Dai & Liu, 2012; Werbos, 1990) method and the \(adam\) (Ruder, 2016) optimizer. A pair of experiment results are presented to show the influence of feature selection on the learning results.

The classifier model was trained using vector input consisting of three features. Choosing only 3 features was based on the idea of triangulation which requires at least 3 measurements to identify location. The score is 51.49%. Figure 10 shows that the near distances to the sensor are easily distinguished compared to the further distances.

![Figure 10: The normalized confusion matrix on real data using 3 features](image)

Figure 11 shows the results of a classifier trained using vector input consisting of eight features, by using the parallel outside electrodes; the score is 83.64%.

![Figure 11: The normalized confusion matrix on real data using 8 features](image)

**Final classifier**

Many experiments were performed (around 50), to find an optimal ANN model based on the real data as it contains noise and is more significant for real world applications. The simulated model – as expected - yielded a higher performance without difficulties.

**Simulated data**

Simulated data classifier yields 90.33 % score with not difficulties. ANN architecture consists of feed forward network with 3 hidden layers 80-40-20 nodes respectively. The cross-validation method was not used. Figure 12 shows the confusion matrix presenting the true label on vertical axis and detected labels on the horizontal axis. It is normalized and shows greater color density closer to 1 for most classes.
Real data
Real data classifier built using the 22 features ($OEs+qEs$) yields the highest score of 94.87%, after using cross-validation and tuning manually the hyper-parameters. Figure 13 shows the confusion matrix normalized with greater color density close to 1 for most classes.

Final regressor

Simulated data
The simulated data regressor has an RMSE of 0.06, which is the square root of the variance of the residuals, it indicates the absolute fit of the model to the data and how good the true data points are to the model’s predicted values. Better fit is indicated by smaller RMSE values. Cross-validation was not used, with 3 hidden layers of 80-40-20 nodes respectively. Figure 16 shows the regression chart of the test sets samples on horizontal axis, and the expected and predicted distance value.

Real data
The real data regressor was built using the 22 features ($OEs+qEs$) and has the RMSE error of 0.39 after using cross-validation and tuning manually the hyper-parameters. Figure 17 Shows the regression chart of the test set samples on the horizontal axis, and the expected and predicted distance value. The model can predict closer distances much better than longer distances.
CONCLUSION

In this project, we have focused on building a machine learning classifier and regressor models to infer on CTPS sensor data. The study shows promising results, especially on successfully classifying far distances with high resolution despite the noise present in the real data. Feature selection process did show the effect of getting the relevant subset of attributes in the data, which can improve accuracy, reduce training time and reduce overfitting. In this study, all 22 features of the real data yielded better accuracy. This confirms to the building of the classifier as well as the regressor models to realize a high accuracy and the least RMSE error.

Further determination of other object parameters are now the subject of future works. Therefore, automatic fine tuning of the hyper-parameters is to be investigated. Especially hybrid architectures, such combining a classifier and a regressor seem promising.

Regarding the efforts of learning we will investigate whether the simulated data model could be used as training set to the neural network and then be successfully used on real data. This would drastically reduce the efforts for the training process.

REFERENCES


A HIGHLY EFFECTIVE TRANSFER LEARNING FRAMEWORK OF DCNNs APPLIED TO AUTOMATIC POLYP DETECTION IN COLONOSCOPY IMAGES

Qinghui Liu†, Antonio L. L. Ramos‡, and Sergio L. Netto‡
†University College of Southeast Norway (USN)
‡Faculty of Technology and Maritime Sciences – Kongsberg
‡Federal University of Rio de Janeiro (UFRJ), Brazil
{ samleogh@gmail.com antonio.ramos@usn.no sergioln@smt.ufrj.br }

ABSTRACT

Convolutional neural networks (DCNNs) have shown a great potential to solve many computational problems, with special emphasis on those where a specific analytical solution would be impractical. However, the usually huge amount of data required poses great challenges, especially in the health care sector where collecting and labeling a larger amount of data can be very difficult. This work proposes an effective transfer learning (TL) framework relying on pre-trained DCNNs using a large collection of natural ImageNet images. This has been achieved by evaluating various kinds of cutting-edge techniques including both traditional machine-learning methods by training feature-based classifiers from scratch and modern DCNNs algorithms with (TL) and fine tuning pre-trained models. We transfer learned ImageNet weights as initial weights, and then fine-tune this model combined with a new deep classifier called fully connected networks (FCNs) with data augmentation and patch-extraction of colonoscopy images for automatic detection of polyps. In case of insufficient colonoscopy images, patch-based data augmentation and deep features extracted using TL strategy can provide sufficient and balanced classification information. After optimizing the hyper-parameters of the TL framework, the system achieved overall 96% polyp detection precision and sensitivity, which outperformed the traditional machine learning classification methods in each defined performance metric. Moreover, the TL framework proposed is scalable and flexible so that it can be easily extended to include other types of diseases detection by integrating more DCNNs models.

INTRODUCTION

Early detection of polyps, protrusions from the colon surface, is vital for the prevention or treatment of colorectal cancers, most of which are curable when detected in early stages. To reduce the miss-detection of polyps caused by human factors and the cost and time of screening a large number of colonoscopy images, various techniques have been studied and exploited recently for the automatic detection of polyps in colonic images.

However, computer-aided automatic detection of polyps is still a difficult task due to the variety of shape, size, color, texture and size scale in the captured images. Additionally, the complex structure of the GI tract, similar color between polyp and non-polyp regions, poor image quality, and image variation of the same polyp caused by frequent camera angle changes impose additional challenges.

In the literature, the early work of Iakovidis et al. (Iakovidis, Maroulis, Karkanis, & Brokos, 2005) and Li et al. (Liu & Ramos, 2016), investigated texture extraction methods for the detection of gastric polyps by using an SVM-based classifier. Another method, as Hwang et al. (Hwang, Oh, Tavanapong, Wong, & Groen, 2007), (Hwang & Celebi, 2010) suggested, is to utilize curvature-based geometric information to detect the shots of polyps. The work by Tajbakhsh et al. (Tajbakhsh, Gurudu, & Liang, 2016) proposed a hybrid context-shape approach, which utilizes texture context information to remove non-polyp structures and shape information to reliably localize polyps. The suggested system had achieved a sensitivity of 88.0%.

More recently, Deep learning (DL) techniques have become state of the art for many image and signal processing tasks (Liu & Ramos, 2016). A great variety of deep learning architectures have been developed and extensively studied. Some example are deep belief network (DBN) (Hinton, Osindero, & Teh, 2006), auto-encoder (Vincent, Larochelle, Bengio, & Manzagol, 2008), deep convolutional neural network (DCNN) (Krizhevsky, Sutskever, & Hinton, 2012), recurrent neural network (RNN), region-based convolutional neural network (R-CNN) (Girshick, Donahue, Darrell, & Malik, 2014). In signal processing (Hui-Hui, Jia-Ming, Mingyu, & Guo-Zheng, 2015), (Santos-Mayo, Jose-Revuelta, & Arribas, 2016) they have been successfully applied in various subfields, such as natural language processing (Mnih & Hinton, 2009), (Weston, Ratle, Mobahi, & Collobert, 2012), (Abdel-Hamid, Mohamed, Jiang, & Penn, 2012), computer vision (Taigman, Yang, Ranzato, & Wolf, 2014), (Wu, He, & Sun, 2015), (Schroff, Kalenichenko, & Philbin, 2015), and so on. However, current trends in research have demonstrated that DCNNs are highly effective in automatically analyzing images. The main power of DCNNs lies in its deep architectures (Eigen, Rolfe, Fergus, & LeCun, 2013), (Szegedy et al., 2015), (Simonyan & Zisserman, 2014), which allows for extracting a great number of features at multiple levels of abstraction.
There are recent works (Ribeiro, Uhl, & Häfner, 2016) and (Tajbakhsh, Gurudu, & Liang, 2015) that have exploited CNNs-based methods for automatic detection of polyps in endoscopy and colonoscopy images. However, it is a great challenge to train DCNNs from scratch (full training). Because DCNNs require not only a large number of domain tagged dataset which is difficult to achieve in the medical field, but also the training DCNNs require a lot of computing resources, without which the training process would be very time-consuming. Additionally, training DCNNs is often complicated by over-fitting and convergence problems, and it is often necessary to optimize a large number of learning parameters and architectures of the network to achieve proper convergence which requires a great deal of expertise and effort to ensure that all layers are learning at a considerable rate.

In view of the above difficulties, a promising alternative to full training DCNNs is to transfer learning and fine-tune DCNNs pre-trained by a large labeled dataset from a different domain (e.g. ImageNet (Russakovsky et al., 2015), which contains 1.2 million images with 1000 categories). The pre-trained models have been applied successfully to various computer vision tasks as a feature generator or as a baseline for transfer learning (Sharif Razavian, Azizpour, Sullivan, & Carlsson, 2014), (Penatti, Nogueira, & Santos, 2015).

**METHODOLOGY**

In this paper, we present a high effective transfer learning (TL) framework (3) which makes use of ResNet50 (He, Zhang, Ren, & Sun, 2016) model with weights pre-trained on ImageNet combining with new designed but highly customized input layer and output layers (top-layer). Motivations for this pre-trained ResNet50 model were a simultaneously deeper as well as computationally inexpensive architecture, and safe to say, the ResNet is now the best single CNN architecture for object detection.

**TL-Framework**

In the proposed TL framework, we design new top layers to replace the FC-1000-d layer of ResNet50. The new top layers consist of two new FC layers (FC-512 and FC-2) and one dropout layer between the 2 FC layers, as shown in Figure 1. We transfer the learned ImageNet weights as the initial weights, and fine-tune the customized model with the new top layers by training and running back propagation on the built-in ResNet50 with our patch-balanced polyp dataset.

The new top layers use softmax as the output layer to predict a separate probability for each of our categories: polyp or no-polyp, and the probabilities will all add up to 1. This softmax function is a normalized exponential defined as:

$$ y_c = \rho(z)_c = \frac{e^{z_c}}{\sum_{d=1}^{C} e^{z_d}} \text{ for } c = 1 \ldots C (1) $$

Here the function $\rho$ takes as input a $C$-dimensional vector $z$ and outputs a $C$-dimensional vector $y$ of real values between 0 and 1. The denominator $\sum_{d=1}^{C} e^{z_d}$ acts as a regularizer to make sure that $\sum_{c=1}^{C} y_c = 1$. In our case, $C$ is initialized to be 2 ($C = 2$) since we have only two classes: polyp and non-polyp. In addition, we make use of the ReLU functions as the activation functions in our framework. The biggest benefit of ReLU is that it bypasses the vanishing gradient problem.

---

**Fig. 1 Transfer learning framework, by fine-tuning 50-layer ResNet model with a new designed top layer that consists of FC-512, Dropout, FC-2, and softmax output.**
ℒ(𝑋, 𝑌) = −\sum_{𝑖=1}^{𝑛} 𝑦(𝑖)\ln(\alpha(𝑥(𝑖))) + (1 − 𝑦(𝑖))\ln(1 − \alpha(𝑥(𝑖)))

In our work, the cross-entropy loss function is used for parameter estimation in training neural networks, which is defined as:

\[ L(𝑋, 𝑌) = −\frac{1}{𝑛}\sum_{𝑖=1}^{𝑛} 𝑦(𝑖)\ln(\alpha(𝑥(𝑖))) + (1 − 𝑦(𝑖))\ln(1 − \alpha(𝑥(𝑖))) \]

where \( X = \{x^{(1)}, ..., x^{(n)}\} \) is the set of input examples in the training dataset, and \( Y = \{y^{(1)}, ..., y^{(n)}\} \) is the corresponding set of labels for those input examples. The \( \alpha(\cdot) \) is the output of the neural network given input \( x \), which is typically restricted to the open interval \((0, 1)\) by using a ReLu. Besides, we choose Adadelta (Zeiler, 2012) as the gradient descent optimizer of the framework. Although adadelta algorithm strives to do away with learning rate tuning, in practice the issue is not completely resolved. The constant \( \varepsilon \) can be considered as the 'learning rate' of adadelta because it actually determines the update of \( \Delta x_t \) since \( RMS[\Delta x] = \sqrt{E[\Delta x^2] + \varepsilon} \), and \( E[\Delta x^2] = \rho E[\Delta x^2]_{t-1} + (1 - \rho)\Delta x^2_t \) (Zeiler, 2012). Here, RMS stands for root mean squared. Setting and tuning constant \( \varepsilon \) and decay rate \( \rho \) are still important and necessary in our work to achieve sound performance curve while the adaptation can effectively counter the learning rate with its own scaling if the optimization directs it in that direction.

Hyper parameters

Based on our TL-framework, we provide a set of hyper parameters that include learn rate (\( \eta \)), decay (\( \rho \)), batch size (\( Bs \)), input size (\( Is \)), epoch number (\( Te \)), dropout rate (\( Dr \)), and pooling size (\( Ps \)), which makes our system very flexible and scalable. There is a brief description of hyper-parameters listed as below:

- Learning rate (\( \eta \)): Learning rate is one of the most important and sensitive parameters that multiplies the computed gradient in the update and determines how much an updating step influences the current value of the weights.
- Decay rate (\( \rho \)): When training deep neural networks, it is necessary to lower the learning rate as the training progresses by setting a proper decay rate.
- Dropout rate (\( Dr \)): Dropout is a simple but quite effective way to regularize the neural networks and address the over-fitting problem.
- Batch size (\( Bs \)): In practice, batch size and learning rate are related. If a batch size is too small, then the gradients become more unstable and would need to reduce the learning rate. In addition, the higher the batch size, the more memory space we will need.
- Input size (\( Is \)): The size of image resized to feed the model, which is related to batch size, which depends on the GPU’s capability.
- Training epochs (\( Te \)): One epoch means one forward pass and one backward pass of all the training samples.
- Pooling size (\( Ps \)): In our neural networks, we utilized average pooling methods before the fully connected layers in order to reduce the resolution of the feature map but retain features of the map required for classification.

Fig. 2 Fine-tuning learning rate experimentation. We first only tuned learning rates and fixed other parameters (\( \rho:0, Dr:0.5, Ps:3\times3 \)) with fixed small input image size (100×100) and big batch size (20) in order to reduce the training time.
EXPERIMENTS AND RESULTS

Dataset preparation

We validated our proposed methodology on our patch-balanced dataset generated from CVC-ColonDB dataset (Bernal, Sánchez, & Villariño, 2012), which contained 300 colonoscopy frames with a total of 300 polyp instances extracted from 15 different colonoscopy video studies. We propose the following methodology for patch extraction:

- Positive patches: we extract a patch (300×300) which covers the whole polyp from every frame (574×500).
- Negative patches (non-polyp patches): we crop the region, which does not contain any part or only cover a little part of polyp from each frame.

After patch extraction, we make use of data augmentation techniques with horizontal and vertical flips, random rotations and so on to artificially boost the amount of positive and negative samples. Finally, we generate our new balanced dataset with 2200 training samples and 400 test samples. The positive and negative sets are equal in size.

Fine-tuning

Fine-tuning hyper parameters of deep neural networks is a tricky process. Though there are some automatic fine-tuning approaches such as grid search (Hsu, Chang, Lin, & others, 2003), random search (J. Bergrstra & Bengio, 2012), or Bayesian optimization (Mockus, 2012) and TPE algorithms (J. S. Bergrstra, Bardenet, Bengio, & Kégl, 2011), etc. All these methods either are too costly and time-consuming or too difficult to apply in unique deep neural networks. Therefore, experimentation with hand-tuning is still the best approach till now for fine-tuning deep learning systems, since a few training epochs and evaluations can give a good judgment which settings would be suitable or not, and then we can give better and better configuration for the next set of parameters.

Therefore, we fist conduct experiments to primarily establish a proper setting range for the most important 3 hyper parameters (learning rate, decay rate, and drop rate) of the TL framework, whereby we can avoid wasting time on poor settings which would likely have resulted in worse performance caused by over-fitting or non-convergent problems. However, for our project, fine-tuning hyper-parameter directly on the whole dataset is too costly and time-consuming considering the limitation of hardware. Therefore, we decided to take the following hand-tuning strategy:

- Set up a subset database (sub-sample 650 images from our patch-balanced dataset).
- Tune a hyper-parameter one time on the subset database with a small input size and big batch size to reduce training time.
- Figure out a rough setting range of the hyper-parameter by observing and analyzing the experimental results and related learning curves.
- Further, fine-tune parameters within the rough range over the whole dataset with increased input sizes.

- Determine a more accurate setting range of hyper-parameter by careful trade-offs in tuning among different combination settings.

By using a small dataset, we could quickly get rough but valuable information about the detail performance of our network against different parameter settings. Figure 2 illustrates experimentation for searching proper learning rates. We first only tuned learning rates and fixed other parameters (ρ:0, Dr:0.5, Ps:3×3) with fixed small input image size (100×100) and big batch size (20) in order to reduce the training time. It is clear that CFG1-1 configured with η:0.01 performed much better than the other two models (CFG1-2 and CFG1-3) with setting of lower learning rates at η:0.001 and η:0.0001 separately. The results showed that the proper range of learning rate should be above 0.01 for our system. Based on this general idea, we can avoid wasting time to try the learning rates lower than 0.01 for the next experiments. It only took us about 1 hour to get a valuable indication that the learning rate should not be set lower than 0.01 for our transfer learning system.

We then repeated a few similar experiments afterward with a set of different learning rates ranging from 0.01 to 0.09. At last we worked out reasonable learning rates that could be in the range of 0.04-0.07. Based on the same strategy, we then fixed the learning rate at 0.05 to tune the dropout rate, and the decay rate. The other hyper parameters (input size, batch size, epoch and pooling size) are linked and depend on the size of networks and GPU’s capability. We have to take some trade-offs and compromise on their settings based on our hardware configuration so that their varying ranges are not so wide as learning rate or drop rate. Thus, it is easier for us to optimize since there are only several optional combination settings for them. Figure 3 illustrates an experiment for tuning the batch-size. The results indicated the proper range of batch-size could be from 16 to 24 giving the input image resize of 100×100.

When we determined the rough range of each key hyper parameter on the subset database, we need to conduct a large number experiments on the whole dataset to further evaluate and adjust the setting ranges. Through observing plenty of experimental results, we finally established reasonable configuration ranges of the considered hyper parameters for our TL framework as shown in Table 1.

<table>
<thead>
<tr>
<th>Hyper Parameter</th>
<th>Setting Range From</th>
<th>Setting Range To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate-η</td>
<td>0.045</td>
<td>0.055</td>
</tr>
<tr>
<td>Decay rate-ρ</td>
<td>0.002</td>
<td>0.0027</td>
</tr>
<tr>
<td>Dropout rate-Dr</td>
<td>0.75</td>
<td>0.85</td>
</tr>
<tr>
<td>Input size-Is</td>
<td>100x100x3</td>
<td>224x224x3</td>
</tr>
<tr>
<td>Epochs-Te</td>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td>Batch size-Bs</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>Pooling size-Ps</td>
<td>2x2</td>
<td>7x7</td>
</tr>
</tbody>
</table>
K-fold cross validation

In our work, K-fold (K) Cross Validation (CV) approach is also employed to estimate the performance of various models configured by different setting of hyperparameters. K-fold CV method works by splitting the dataset into K parts (e.g. K = 3, 5 or 10). After running K-fold dataset that we can summarize using a mean and a standard deviation. The result is more accurate because the model is trained and evaluated multiple times on different K-fold dataset. Figure 3 shows the 3-fold CV learning curve of Model-8.

Results

We conducted a set of different experiments on a number of models with specific input sizes and optimized hyperparameters as shown in Table 2. We eventually achieved the best overall 96% polyp detection accuracy, precision, sensitivity, specificity, and F1-score with Model-8 configured by the optimized hyper-parameters (η:0.05, ρ:0.0025, Bs : 10, Dr : 0.8, Ps : 7×7, and Te : 50) as shown in Table 3, which outperformed the traditional machine learning classification methods.

We conducted a set of different experiments on a number of models with specific input sizes and optimized hyperparameters as shown in Table 2. We eventually achieved the best overall 96% polyp detection accuracy, precision, sensitivity, specificity, and F1-score with Model-8 configured by the optimized hyper-parameters (η:0.05, ρ:0.0025, Bs : 10, Dr : 0.8, Ps : 7×7, and Te : 50) as shown in Table 3, which outperformed the traditional machine learning classification methods.

Table 2 Models configuration and test results, where TP stands for True Positive, FN- False Negative, TN- True Negative, FP- False Positive.

<table>
<thead>
<tr>
<th>Hyper Parameter</th>
<th>Model-0</th>
<th>Model-3</th>
<th>Model-6</th>
<th>Model-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kfold</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Epoch</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Input size</td>
<td>100x100</td>
<td>150x150</td>
<td>224x224</td>
<td>224x224</td>
</tr>
<tr>
<td>Batch</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.8</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.049</td>
<td>0.049</td>
<td>0.049</td>
<td>0.05</td>
</tr>
<tr>
<td>Decay rate</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.0025</td>
</tr>
<tr>
<td>Pooling size</td>
<td>3x3</td>
<td>3x3</td>
<td>7x7</td>
<td>7x7</td>
</tr>
</tbody>
</table>

Table 3 Models performance comparison.

<table>
<thead>
<tr>
<th>Performance</th>
<th>Model-0</th>
<th>Model-3</th>
<th>Model-6</th>
<th>Model-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>75.25%</td>
<td>82.75%</td>
<td>92.75%</td>
<td>96.00%</td>
</tr>
<tr>
<td>Precision</td>
<td>77.60%</td>
<td>80.18%</td>
<td>90.91%</td>
<td>96.00%</td>
</tr>
<tr>
<td>Sensitivity/recall</td>
<td>71.00%</td>
<td>87.00%</td>
<td>95.00%</td>
<td>96.00%</td>
</tr>
<tr>
<td>F1-score</td>
<td>74.15%</td>
<td>83.45%</td>
<td>92.91%</td>
<td>96.00%</td>
</tr>
<tr>
<td>Specificity</td>
<td>79.50%</td>
<td>78.50%</td>
<td>90.50%</td>
<td>96.00%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>75.25%</td>
<td>82.75%</td>
<td>92.75%</td>
<td>96.00%</td>
</tr>
</tbody>
</table>
From what have been observed in these experiments, given a proper choice of hyper parameters, the larger the input size (Is) used to train the models, the better the performance achieved. For instance, Model-0 (Is: $100 \times 100 \times 3$), Model-3 (Is: $150 \times 150 \times 3$), and Model-6 (Is: $224 \times 224 \times 3$) have almost identical hyper parameters settings, but Model-6 achieved much better performance as compared to Model-3 and Model-0. Respectively, accuracy (93% vs. 83% vs. 75%), precision (91% vs. 80% vs. 78%), sensitivity (95% vs. 87% vs. 71%), and F1-score (93% vs. 83% vs. 74%). This is summarized in Figure 5.

Though larger input sizes improve the performance of DCNNs models, longer time is required per epoch during training, and larger amount of system memory is required as well. This may lead to out-of-memory issues which can be solved either by decreasing the input and batch sizes, or the system’s hardware configuration improved with larger memory and a more powerful CPU.

DISCUSSION

Impact of the input size

Deep neural network is known to over-fit easily due to the large number of parameters. In our experiments, the overfitting was expected to be significant since the dataset was small even if we utilized data augmentation. To tackle over-fitting problems, we introduced a dropout layer built-in the top FCNs layer in our proposed TL framework. We traded off three key hyper parameters- dropout rate, learn rate, and decay rate to effectively prevent over-fitting problems. Meanwhile, from our observation, we also found bigger batch size could mitigate over-fitting issues in some way, and greater input sizes commonly require a slightly higher dropout rate and decay rate to avoid over-fitting than smaller input sizes.

Tackling over-fitting

As we can observe in our experiments, the proposed TL framework generalizes quite well given that the training accuracy are almost all 100%. First, dropout strategy improves the generalization capability of our models by a large rate (at 0.8% for Model-8). Second, for the structure of ResNets, batch normalization applied in convolutional blocks also help improve both the training speed and generalization. Another important reason is that we replace the fully-connected layer of ResNet50 by a global average pooling layer, and 2 FC layers before the softmax output layer, which greatly reduces the amount of parameters. Thus, our TL DCNNs models demonstrate very strong generalization capability with the state-of-the-art performance.

Constraints
Since we are using the pre-trained model, our TL architecture then uses fixed convolutional filters, kernel size, and number of layers. We are also slightly constrained in terms of the model architecture. For example, we cannot arbitrarily take out certain convolutional layers from ResNet50. However, it is possible to customize the input layer with different image sizes due to parameter sharing.

CONCLUSIONS

In this paper, we present an effective transfer-learning framework that consists of a new FCNs classifier and input layer combined with a pre-trained 50-layer ResNet model. We therefore provide eight hyper parameters that include learning rate (\(\eta\)), decay (\(\rho\)), batch size (BS), input size (IS), epoch number (TE), dropout rate (Dr), k-fold number (K), and pooling size (Ps). These hyper parameters make our framework very flexible and scalable so that it can easily be extended to include other types of disease detection in future.

In addition, due to the high sensitivity to the setting of hyper-parameters, fine-tuning DCNNs is a tricky process. In our work, we creatively introduced a high effective random-hand-tuning strategy to search and select the best and most suitable setting of the hyper-parameter to obtain better performance. We finally achieved overall 96% detection accuracy and precision, 96% sensitivity and specificity, and 96% F1-score by using the proposed TL framework with optimized hyper-parameters, which outperformed the traditional machine learning classification methods in each defined performance metric.

For future work, we have only used colonoscopy images in this paper to evaluate our methods. It is therefore necessary to collect a greater number of capsule endoscopy images, or use a larger medical images datasets, in order to re-evaluate, qualify and respectively confirm the results of this work, and then we could further optimize the proposed framework to be able to detect all different polyp morphological types in future. In addition, though the proposed TL framework allows the use of more than one pre-trained model, we only tested the ResNet50 model in this work, so it would be valuable to add some other cutting-edge pre-trained models such as Google Inception, to achieve better performance and further boost its generalizing capabilities.

REFERENCES


227
HUMAN-LIKE AUTONOMOUS NAVIGATION FOR
COLLISION AVOIDANCE IN DYNAMIC ENVIRONMENTS

Daichi Kawamura and Tad Gonsalves
Department of Information & Communication Sciences
Faculty of Science and Technology, Sophia University, Tokyo, Japan
{ t-gonsal@sophia.ac.jp  d.kawamura0202@gmail.com }

ABSTRACT

One of the basic requirements for safe driving involving autonomous vehicles is the ability to avoid colliding with stationary and moving obstacles in dynamic driving environments. Current AI collision avoidance techniques are mechanical and brittle; they do not take into consideration the human-way of driving and interacting to uncertain situations. This paper deals with the development of an AI navigation system (AINS), which we plan to implement in an autonomous driving system. Using a Feedforward Neural Network driven by the Q learning algorithm, the AINS demonstrates a human-like ability to avoid collisions with moving objects in a dynamic environment. In our various experiments, the AINS readily achieves a 100% accuracy in learning to avoid colliding with a moving object, in a human-like manner. In another experiment, when faced with the challenge of collision avoidance with obstacles with priorities, it learns to avoid colliding with the obstacles with a high priority at all costs, even though this may entail a near-miss collision with another obstacle of lower priority.

INTRODUCTION

Artificial Intelligence (AI) has become a hotbed of research in recent years. Although still in a primitive stage, some of the recently developed AI systems are gradually approaching human cognitive and decision-making capabilities. AI is now poised to make a breakthrough in autonomous driving. Great strides have been made in developing self-driving vehicles capable of safely driving in controlled environments; however, making these vehicles functional in our day-to-day driving environment still remains a big challenge. This is because, day-to-day driving environment is extremely complex.

The outstanding feature of any self-driving system would be, without question, safety. And, the fundamental concept in safety driving is avoiding collisions with obstacles in a given environment. To date, most of the research in collision avoidance is in robotics and computer animation (Park, Park, & Choi, 2011; Shim & Li, 2017) or in an environment where only unmanned vehicles (UVs) are operating (Belkhouche, 2017; Rawat, Faridi & Kumar, 2016). Various scenarios and techniques for collision avoidance in case of a single vehicle automated driving are found in literature: integrated trajectory control (J. Verhaegh, et al., 2017), fuzzy danger level detection (Wang et al., 2017), side collision avoidance (Ou et al., 2017), collision avoidance at traffic intersections (Ahn, 2017), collision avoidance and vehicle stabilization (Funke, 2017). A couple of studies are also dedicated to developing collision avoidance systems to aid human drivers (Su, 2017).

According to a WHO report, pedestrians are involved in about 22% of the road accidents (WHO, 2015). Collision avoidance systems to safeguard pedestrians are found in (Bachmann, Morold & David, 2017; Gelbal, et al., 2017; Ho & Chen, 2017; Sun et al., 2017).

According to Osipychchev (Osipychchev et al., 2017), most UVs accident occur because they do not behave as human drivers expect them to behave while driving. In fact, almost all the studies in collision avoidance found in literature do not take into consideration that human drivers also share the same road as the UVs in real-life situations. Most systems are developed with the implicit assumption of an ideal “mechanized” driving where the erring human factors are not taken into consideration.

Currently, AI collision avoidance techniques are mechanical and brittle; they do not take into consideration the human-way of driving and interacting to uncertain situations. In our study, we go a step further and make our agent human-like in avoiding collisions. Since machines operate in millisecond time-scales, near-miss situations involving only UVs maybe accident safe. However, on the “slower” human time-scale, near-miss collision avoidance is almost always fatal. Humans innately try to keep out of danger as early as possible, from the moment they sense danger. We try to incorporate this innate human-like behavior in our AI navigation system (AINS) learning to avoid collisions while driving in dynamic environments.

In this paper, we present a machine learning model in which self-driving vehicles learn from human-like behaviors in avoiding accidents and thereby reduce risks. The autonomous navigation system for obstacle avoidance in dynamic driving environments learns the art of maneuvering through roads strewn with moving obstacles using the Q-learning algorithm. This is a well-researched and promising algorithm in reinforced learning, where learning is achieved by means of rewards and punishments (Mnih et al., 2017).
This paper is organized as follows: Section 1 introduces the subject matter of this study; Section 2 describes the human-like behavior a true AI collision avoidance system is required to have; Section 3 describes the salient features of Q learning; Section 4 explains the experimental results, and section 5 concludes the study.

HUMAN-LIKE BEHAVIOR

In this section, we define the human-like steady movements involved in reducing the risk of collisions. Our day-to-day over-all behavior in avoiding collisions can be classified into the following component sub-tasks.

1. Avoidance in advance: Humans do not wait till the last minute to avoid collisions. Relying on the visual information at hand, they try to avoid colliding with obstacles when viewed from far.

2. Near-miss situations: When humans happen to be in the vicinity of an obstacle, they avoid near-miss situations. In other words, they distance themselves when passing by the obstacle as much as the given situation allows.

3. Priority: When humans have to mitigate collision risks, they do so by prioritizing the risks to be avoided.

In Fig. 1, the red and the blue blocks represent the obstacle and an AI agent respectively. The AI collision avoidance learning system continues moving on its planned trajectory until very close to the obstacle. It moves away from the obstacle, when the latter is almost within reach.

In Fig. 2, the AI agent plans obstacle avoidance well in advance. It computes the driving trajectory when it is actually quite far from the approaching obstacle. However, it moves very close to the obstacle when passing it by, thinking that near-miss actions are also safe.

The scenarios depicted in Fig. 1 and Fig. 2 are naïve collision avoidance scenarios. In both cases, the AI agent successfully manages to avoid colliding with the moving obstacles. However, the action plan of the AI agent, is at best a “near-miss” scenario, which is potentially risky. In our daily life, especially when the obstacle is a moving object such as a human being, we do not take such a naïve avoidance action.

In Fig. 3, the AI agent moves in advance in the traveling direction to a position where there is no obstacle so that there
is a distance to the obstacle to some extent when passing by it. In addition to reducing the risk of front and rear collisions, the risks of left and right collisions when passing by the obstacle are also reduced. The AI agent avoids near-miss chances in a way very similar to what humans do unconsciously. In our simulation study, we have trained our AINS to imitate this human-like obstacle avoidance behavior.

**Q LEARNING**

Our AINS uses the Q Learning Algorithm and a Multi-Layer Neural Network to achieve the task of learning to avoid collisions while driving. The three major components of the learning system, Q-Learning, Neural Networks, and rewards and penalties are described in the following sub-sections.

**Q-Learning**

Q-learning, often referred to as “model-less” is a variation of reinforcement learning. In Q-learning, there is an agent having states and corresponding actions. At any moment, the agent is in some feasible state. In the next time-step, the state is transformed into other state(s) by performing some action. This action is accomplished either by a reward or a punishment. The goal of the agent is to maximize the reward gain.

The Q-learning algorithm is represented by the following update formula:

\[
Q(s_0, a_0) \rightarrow \\
Q(s_0, a_0) + \alpha (r_t - Q(s_t, a_t) + \gamma \max_{a'} Q(s_{t+1}, a')) \\
(1)
\]

where \(Q(s_t, a_t)\) represents the Q-value of the agent in state \(s_t\) and action \(a_t\) at time \(t\), rewarded with a reward \(r_t\). \(\alpha\) is the learning rate and \(\gamma\) is the discount factor. The \(\gamma\) parameter is in the range \([0, 1]\). If \(\gamma\) is closer to 0, the agent will tend to consider only immediate rewards. On the other hand, if it is closer to 1, the agent will consider future rewards with greater weight, thus willing to delay the reward. The Learning Rate and the Discount Factor, described below, are the two most crucial parameters influencing the performance of the Q learning algorithm.

**Learning Rate**

The learning rate determines the strength with which the newly acquired information will override the old information. A factor of 0 will make the agent not learn anything, while a factor of 1 will make the agent consider only the most recent information. In fully deterministic environments, a learning rate of is optimal. When the problem is stochastic, the algorithm still converges under some technical conditions on the learning rate, that require it to decrease to zero. In practice, often a constant learning rate is used.

**Discount Factor**

The discount factor determines the importance of future rewards. A factor of 0 will make the agent short-sighted by only considering current rewards, while a factor approaching 1 will make it strive for a long-term high reward. If the discount factor meets or exceeds 1, the action values may diverge. Even with a discount factor only slightly lower than 1, the Q-function learning leads to propagation of errors and instabilities when the value function is approximated with an artificial neural network. In that case, it is known that starting with a lower discount factor and increasing it towards its final value yields accelerated learning.

**Neural Network**

Artificial Neural Networks (ANN) are an imitation of the structure and function of the natural neural network of the mammalian brain. ANN are multi-layered networks consisting of at least three different layers: the input layer, the hidden layer and the output layer (Fig. 4). The number of neurons in the input as well as in the output layer is determined by the problem the ANN is learning to solve. However, the number of neurons in the hidden layers is arbitrary.

![Fig. 4 Neural Network model](image)

Each neuron in a given layer is connected to every other neuron in the successive layer. The connection between any two neurons has an associated weight. The learning carried on by the network is encapsulated in these interconnection weights. Each neuron calculates a weighted sum of the incoming neuron value, transforms it through an activation function, and passes it on as the input to subsequent neurons. The information processing proceeds from the input layer to the output layer via the hidden layer. Hence the name, Feedforward Network.

The AINS simulation is performed on a 10 x 10 grid structure. We need to transform the grid structure to make it compatible with the structure of the input layer of the ANN. This is done by “flattening” the 2D map structure into a one-dimensional array like a vector and input to the network. When the grid map has a size of 10 x 10, since all cells are...
10 x 10 = 100, the input data becomes a vector having 100 elements. Subsequently, these inputs are received by the hidden layer with 100 nodes and applied to the activation function. Finally, those outputs are received by the output layer with 3 nodes. The three layers of the output layer correspond to the actions that AI can take respectively. “DO NOTHING”, “TURN LEFT” or “TURN RIGHT” (Fig. 5).

Rewards and penalties
To facilitate exploration in learning, we allow the AINS to move randomly 10% of the time. It is given a reward of 1 for every successful trial and a final reward of 100 for each obstacle avoided, at the end of a training session. On the other hand, penalty is also imposed on the AINS which depends on its distance from the obstacle – the closer it gets to the obstacle, the greater is the penalty, given by:

\[ P_d = 10 \times (1 - \frac{d}{L}) \]  (2)

where,
d is the distance between AINS and an obstacle
L is the total length of the simulation grid

When the AINS learns to avoid colliding with two obstacles, each with a pre-assigned priority, the penalties are imposed as follows:

High risk Penalty = -300  \hspace{1cm} (3)
Low risk Penalty = -200

When the AINS ends us in a near-miss situation while learning collision avoidance, the penalties are computed as follows:

High risk Penalty = -150
Low risk Penalty = -100

EXPERIMENTAL RESULTS
This section presents the results of three different obstacle avoidance scenarios.

Single obstacle
The self-driving learning system is presented with just a single obstacle. The system senses the presence of the obstacle and learns to avoid it efficiently while driving. The experimental results obtained in training as well as in testing in three different sets is 100% (Table 1).

<table>
<thead>
<tr>
<th>Sets</th>
<th>Training (%)</th>
<th>Testing (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Moreover, in all the training as well the testing experiments, the agent depicted a true human-like collision avoidance maneuver (Fig. 3).

Two obstacles
We tried 3 different sets, each with 100 attempts and calculated their mean (Table 2).

<table>
<thead>
<tr>
<th>Sets</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two obstacles avoidance rate</td>
<td>90%</td>
<td>87%</td>
<td>91%</td>
<td>89.3%</td>
</tr>
</tbody>
</table>

Two obstacles with priorities
Finally, we allowed the AINS agent to drive in an extreme situation involving two obstacles with varying priority levels of collision avoidance. The Q-learning algorithm rewards the agent with greater score to avoid colliding with a high priority (high risk) object than with the low priority (low risk) object. The agent, accordingly learns to avoid the higher risk obstacle when faced with two hazardous obstacles with varying priorities. The following two kinds of human-like behaviors are demonstrated by the AINS:

(1) When the two moving obstacles do not come within the danger zone, the AINS passes smoothly in between the two keeping a fairly safe distance from them (Fig. 6).
(2) When the two moving obstacles come within the danger zone, the AINS chooses to avoid the
high-priority obstacle at all costs even though it may near-miss hitting the low-priority obstacle (Fig. 7).

From (1) and (2) we can conclude that the AINS successfully learns not only to avoid hitting the high-risk obstacle, but also to avoid getting into the its near-miss range. This has a great practical significance in autonomous driving.

Table 3 shows the results obtained in 5 test sets.

Table 3 Success rate in case of two obstacles with priorities

<table>
<thead>
<tr>
<th>Sets</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collisions with low risk obstacle</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Near-miss with low risk obstacle</td>
<td>95</td>
<td>99</td>
<td>93</td>
<td>94</td>
<td>100</td>
</tr>
<tr>
<td>Collisions with high risk obstacle</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Near-miss with high risk obstacle</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 8 shows the graph of AINS learning to avoid colliding with two obstacles with priorities in a total of 100,000 attempts.

**CONCLUSION**

This study addresses the safety issue of autonomous driving in dynamically changing environments by designing an AI Navigation System (AINS) that learns to avoid collision with moving obstacles. Current AI collision avoidance techniques are mechanical and brittle; they do not take into consideration the human-way of driving and interacting to uncertain situations. Using a Feedforward Neural Network driven by the Q learning algorithm, our AINS demonstrates a human-like ability to avoid collision with moving objects while driving. In our various experiments, the AINS readily achieves a 100% accuracy in learning to avoid colliding with a moving object, in a human-like manner. In another experiment, when faced with the challenge of collision avoidance with obstacles with priorities, it learns to avoid colliding with the obstacle with a high priority at all costs,
even though this may entail a near-miss collision with another obstacle of lower priority. The next step in our study would be to incorporate fuzzy control in the AINS to further smoothen its response and train it in more hazardous and risky driving environments.

REFERENCES


ABSTRACT

Cyber physical spaces and internet of things have triggered a new era in software development. They have overrode smart spaces, which could be seen as subsets of CPS. Mobile and smart devices, bring forward the idea of physical presence and should be designed according to well recognized architectural model and developed solutions for new types of software components, which results in reusable software solutions and contributes towards a higher level of automation. Our study shows an example of proximity based documentation management through beacons and filtering relevant content in a data set available in the aviation industry.

INTRODUCTION

Cyber Physical Spaces (CPS) (Raykumar et al., 2010) have their origins in the specification of ubiquity in software development which paved the way towards pervasiveness of computational spaces (Bacon, 2002), (Banvar et al., 2000), (Gregory and Elisabeth, 2000), (Lyons, 2002). In the last decade, the advances in mobile and wireless technologies forced us to revisit the vision of ubiquity and pervasiveness (Saha and Mukherjee 2003), (Satyanarayanan, 2001), (Eblinng, 2017), (Satyanarayanan, 2017), particularly in the light of Weiser’s vision of the computer for the 21st Century (Weiser, 1991). Today, we talk about smart physical spaces and internet of things (IoT), which are not dissimilar to well-known pervasive spaces, but they focus on connections of intelligent devices instead of pushing forward subconscious integration of computers with humans (Satyanarayanan, 2017) (Shojanoori, 2013). According to Satyanarayanan, 2017, IoT is silent about the roles of humans, as opposed to Wieser’s vision to place humans at the center of computing. However, our new era of transient smart spaces, underpinned with mobile and smart devices, bring forward the idea of physical smart spaces, which could be seen as subsets of CPS.

Today, CPS touch every aspect of our everyday lives and will remain present in the near future. They have overrode wireless sensor networks (Wu et al., 2011) and are seen as amalgamation of computational power and physical and engineered systems as in (Rho et al., 2016), which is a common perception of CPS. However, the other authors talk about cyber physical social computing (Sheth et al., 2013), cyber-physical social thinking (Huansheng et al., 2016), smart cyber society (Awais et al., 2016) and cyber physical human systems (Scheuremann et al., 2015). They also push forward the attempts to capture the semantics of data generated by devices in CPS and bring them closer to computational power which creates these spaces (Henson et al., 2012), (Shojanoori, 2013). A “prescription” for CPS from (Hartung et al., 2015) gives a philosophical look at the problem of designing CPS and raises awareness of “technology providers pushing burdensome features in to the unexpected users” in CPS. We have to be aware of constant technology advances, which penetrate our lives as we speak: the way we create, design and manipulate CPS is and will be adversely affected by technology changes.

There is no consensus in industry and academia on the modeling, designing and creating CPS. For readers interested in Cyber Social Physical Systems, we suggest reading a survey of their design methodologies in (Zeng et al., 2016). Otherwise, we could list a few articles which focus on CPS architecting (Canedo and Richter, 2014), (Tsikganos et al., 2016), (Bonnema, 2014), (Neema et al., 2014). In this paper, we are more interested in using the practices of the development of software intensive systems, and component based design principles, as elaborated in (Crmkovic et al., 2016), due to one reason. Modern CPS are software intensive systems. They would not exist without significant software presence and should be designed according to well recognized practices from the world of pervasive commutating (Shojanoori et al., 2012). CPS are not far away from the main principles of pervasiveness in modern software systems.

In this paper we illustrate the development of reusable layered and component based software architectural model for creating a CPS, which uses blue-tooth low energy (BLE) technology (Bluetooth, 2017) for the management of data set relevant to physical items in that CPS. Our example is taken from the aviation industry, but the reusability of our software architectural model allows its applicability in any domain when proximity based technology is essential for creating a
CPS. We use beacon technologies because of its wide applicability in numerous problem domains (Bouchard, 2016), (Filippopolitis et al., 2016), (Liu et al., 2016), (Shinotsuka et al., 2016), Chagas De Oliveira, 2017), (Sato et al., 2017), (Srinivasan et al., 2017), (Akinsiku and Jadav, 2016), (Jia et al., 2014), (Stetten, 2016). Our main research goals are

(I) Develop a reusable and layered component based architectural model, with UML modelling concepts, which can be used for the creation of software applications, and rely on BLE technologies for creating a particular CPS. Reusability should (a) allow the creation of a CPS in any other domain (not solely aimed at the aviation industry) and (b) address a separation of concerns typical in software intensive systems (Juric et al., 2004), (Gamma et al., 1994).

(II) Illustrate modelling principles typical of software intensive systems in which physical items are incorporated according to their role and data they generate in CPS;

(III) Elaborate on the issues of the deployment of the architectural model from (I) in Android and iOS: this particular CPS would not exist without smart and mobile devices with which we can perform computations.

The paper is organized as follows. In the next section we describe the Problem Domain and motivation for this research. In the section which follows we talk about Technologies and Standards essential for the deployment of the architectural proposal. This section is an answer to problems mentioned in (Hartung et al., 2015) and serves as a good example of “technologies, used for implementing a software intensive solution, which affect its creation”. In the Modeling section we focus solely on the modelling with UML, in which we specify which modelling principles are used and why. It is important to note that the creation of layered and component based architectural model, dictates the use of UML modelling concepts, abstracted from the problem domain and conceptualized in a selection of UML diagrams. The Implementation section gives only excerpts, which highlight the differences in the nature of programing behind the functionality of our CPS. There were numerous challenges in this work, but in the section which follows, we again focus on the impact of technologies on the deployment of our reusable architectural model. The Conclusion section overviews and evaluates our work, highlights our contribution to knowledge and practice and makes possible future pathway for creating practical and modern CPS.

PROBLEM DOMAIN

In this research we focused on the modelling and implementation of a CPS in which a mobile software application access documentation of nearby physical items. The most likely scenario is in the field of aircraft maintenance, in which technicians/engineers perform their job by consulting the documentation relevant to numerous physical items and their parts, which exist on the aircraft. The software application would work with special markers on the aircraft, called beacons, which give information about these physical items. This will enable accessing relevant documentation and instantly displaying it on the mobile device, operated by the technician/engineer. The information within the documentation would include anything relevant to the history, measurements and parts of these physical items and the way of maintaining them. It has been known that up to 40% of engineer/mechanic work time is used for finding the documentation and technical information required for performing maintenance works (Wang et al., 2009). Therefore, our solution will save time and eliminate errors in the aircraft maintenance process.

In this problem domain, the software application uses BLE beacons (Davidson, et al., 2011) which filter content of the documentation on a smart phone, in a web viewer, with a preloaded data set. The application must achieve a high-level of accuracy in the filtering of the documentation and should be location aware.

TECHNOLOGY AND STANDARDS

The goals of this research would require a careful consideration of technologies for the deployment of our software architectural model in wireless and mobile operating environments when creating a CPS. In other words, we must specify which technologies are needed, for generating a software application from the proposed software architectural model. There are numerous possibilities available for developers, but we need to look at integrated development environments, which allow the deployments of software components and include beacon’s functionality into our software solution. Our choice was based primarily on the availability (i.e. open source solutions) and simplicity of technologies, in terms of their adoption when modelling and implementing the solution.

We do not claim that our choices are the best possible on the market, but they served us well. In the constantly changing software development world, the choice of tools used for the application development must be regularly revisited. In any future attempt of repeating this work, we will have to check the viability of our chosen technologies. Therefore, this section just lists the technologies and standards used, and does not give any rationale behind our decisions.

Angular (Angular, 2017) has been used as a development platform for building web applications, but it also uses TypeScript (Microsoft, 2017), as a strongly typed language and NativeScript framework (NativeScript, 2016, 2017) for the development of mobile applications. Angular recognizes software components in our models and packages them into User Interfaces, Service, Communications and Modelling. It proved to be a very effective integrated development environment for this application. Routing in Angular enabled the navigation between the views of the application, without creating a new page, like in all modern browsers. An Angular router-object passes routes as arguments in order to enhance the navigation, and binds links to buttons and drop down menus.
NativeScript was used for creating native cross platform applications for Android and iOS, with Universal Windows Platform Support. NativeScript apps are written using JavaScript or Typescript, and directly support Angular. NativeScript supports plug-ins, which enable the development of cross platform components, in spite of them using different SDK. This was a particularly important because Estimote SDK (Estimote, 2017) for iOS and Android are different.

TypeScript, as a typed superset of JavaScript, is compatible with any browser, any host and any operating system. Typing in TypeScript enables static checking and code refactoring when developing JavaScript applications. Estimote’s API has been used in order to utilize the company’s goals to make iBeacons more accessible to developers (Jackson, 2014).

BLE beacons were used as they transmit a range of single and multiple packages, characterized by its lowest possible energy consumptions, complexity and costs. Additionally, the S1000D specification (ASD, 2017) has been used for the production of technical documentation, which is available as a modified version for the land, sea and with any commercial equipment. The loading of a desired data set in our application, has been done in the format of Pinpoint Neutral Package, which is very similar to S1000D. The files, which were originally in XML format, have been reformatted into JSON files. This was extremely important in the filtering of relevant documents, because it enabled the creation of the Table of Contents and navigation to the right document in the document folder.

It is important to note that in this domain we use the term Applicability Documents to denote a list all different filter statements for each physical item or their parts on the aircraft.

MODELLING

The main modelling principle was based on our adherence to UML modelling concepts and the illustration of the semantic of the problem domain in a selection of UML diagrams. Any methodology which allows to

- depict the main functionality of the problem domain in Use Case models,
- extract all abstractions from Use Case models in the format of object interactions in Sequence Diagrams in particular and
- conceptualize these objects in Class Diagrams, would guarantee that we will detect the main building blocks of our CPS in the format of reusable software components.

It is important to note that the role of class diagrams will be specific to the ultimate goal we wish to achieve. If we strive for the detection of reusable software components, in which we can model the presence of physical objects, through either the role they may have in, or data they generate in CPS, then the selection of UML diagrams from the bullets above will produce a desirable result. Class diagrams, which model the whole application, may not adhere solely to the data modelling part of our solution; they should directly generate our software architectural model.

The UML was standardized almost 15 year ago, before the pervasiveness of computing became a norm and CPS our reality. However, the level of stereotyping in UML, and powerful mechanisms of creating software components from abstractions in semantics of problem domains, are very safe mechanisms in modelling any software intensive solution including CPS.

Our modelling steps reflect the rationale above and include:

a) Definition of a simple Use Case model,
b) Discovery of main objects, which are seen as key abstractions within each of the use case from a), and the creation of dynamic connections between them through message passing in Sequence Diagrams,
c) Generation of Class Diagrams, which collect all types of objects from Sequence Diagrams: from user interfaces to computations and data, which are directly abstracted into software components of the architectural solutions for our CPS.

Obviously, a)-c) above dictate that all UML diagrams and the final software architectural model shows the thread of functionality defined in use cases. In other words, the functionality from the Use Case model must be “visible” in all UML diagrams, when modelling a CPS as a software intensive solution. This was one of the main principles behind the evolution and the semantics of UML.

Finally, in this research we do not adhere solely to any formal methodology available in software development for two reasons. Firstly, our goal is to create reusable software components which model user interfaces, computations and data storage at the same time and discover or define a level of sharing of any of them. Secondly, our model should depict exactly how we view “physical items” in our CPS and how they contribute towards the delivery of functionalities through computational software components. However, adhering to a)-c) above will bring us quite close to Agile modelling principles as defined in (Cockburn, 2007) and will allow an iterative approach in modelling, particularly when discovering abstractions within our Problem Domain.

Use case Model

The UML was standardized almost 15 year ago, before the pervasiveness of computing became a norm and CPS our reality. However, the level of stereotyping in UML, and powerful mechanisms of creating software components from abstractions in semantics of problem domains, are very safe mechanisms in modelling any software intensive solution including CPS.

Our modelling steps reflect the rationale above and include:

a) Definition of a simple Use Case model,
b) Discovery of main objects, which are seen as key abstractions within each of the use case from a), and the creation of dynamic connections between them through message passing in Sequence Diagrams,
c) Generation of Class Diagrams, which collect all types of objects from Sequence Diagrams: from user interfaces to computations and data, which are directly abstracted into software components of the architectural solutions for our CPS.

Obviously, a)-c) above dictate that all UML diagrams and the final software architectural model shows the thread of functionality defined in use cases. In other words, the functionality from the Use Case model must be “visible” in all UML diagrams, when modelling a CPS as a software intensive solution. This was one of the main principles behind the evolution and the semantics of UML.

Finally, in this research we do not adhere solely to any formal methodology available in software development for two reasons. Firstly, our goal is to create reusable software components which model user interfaces, computations and data storage at the same time and discover or define a level of sharing of any of them. Secondly, our model should depict exactly how we view “physical items” in our CPS and how they contribute towards the delivery of functionalities through computational software components. However, adhering to a)-c) above will bring us quite close to Agile modelling principles as defined in (Cockburn, 2007) and will allow an iterative approach in modelling, particularly when discovering abstractions within our Problem Domain.

Use case Model

Figure 1 illustrates our Use Case model. It is self-explanatory and models the semantics from the Problem
Within our CPS, we should be able to choose the document we need, select applicability (i.e. define filtering criteria) and have the swipe function available when browsing documents. While choosing a document or swiping it we must know exactly which publications comprises the document.

Discovering Abstractions

All relevant abstraction for the modelling could be found in the description of the functionality of Use Cases from Figure 1, which is summarised in the Problem Domain. The abstractions are discovered as objects, which appear in Sequence Diagrams and have two distinctive roles:

(i) They are grouped because they represent User Interfaces, Computational and Data objects, as it was suggested in early object oriented development practices from the 90s and

(ii) They are conceptualised in classes and their message passing may be used for defining operations in associated classes.

It is obvious that this division of objects from (i) is the first step towards creating reusable software components because they lead towards the deployment of the Mode-View-Controller (MVC) software pattern (Gamma et al., 1994), which would support layered and component based software architectural style (Juric et al., 2004), (Clemens et al., 2003).

Figures 2, 3 and 4 show three Sequence Diagrams, derived from their associated use cases (they are colour co-ordinated).

The reader should note that in our diagrams, we use terms “view” to associate the object with user interfaces and “controller” to define an object performs computations. Objects which represent data, could be named according to the semantic they store. Therefore in Figure 2 we can see that objects which store beacon data and statements related to the applicability (i.e. filtering) are essential for performing beacon’s functionality. For choosing documents in Figure 3 we need information on them and their Table of Contents.

Having two adjacent “controller” objects within a sequence diagram (as in Figures 2 and 3) is an indication of either frequent computations, which may be dictated by the presence of physical objects in our CPS, or data sharing between these computations. The former is a welcome feature of our models because it associates them with CPS. The latter will have direct impact on reusability of these computations if we wish to share data objects.

Our colour co-ordination between use cases in Figure 1 and sequence diagrams in Figures 2, 3 and 4 is self-explanatory. We tend to merge sequence diagrams from use cases which are related with <<include>> stereotyped dependencies as Shown in Figure 2. However Choose Publication use case has been included in two base use cases from Figure 1, but is incorporated in the objects from Figures 3 and 4.

The simplicity of our diagrams in Figures 1-4 does not necessarily mean that we did not find all abstractions, or UML modelling concepts, which will secure viable modelling of CPS diagrams. In contrary, it shows our pathway of following a)-c), focusing on (i) and (ii) and securing that the final model will be a layered and component based architectural style (Juric et al., 2005), which enables a high level of reusability of its software components. If for any reason we wish to increase the amount of semantics or details shown in these three diagrams, we will have to come closer to the implementation which follows the modelling stage. In such cases the choice of technologies, as discussed earlier, and the way our architectural models are deployed in integrated development environments will all dictate the level of precision of our UML diagrams.

Finally if we wish to achieve a certain level of reusability of our architectural model for CPS, then we should not insist on technology specific components at this stage of modelling.
Class Diagram

Figure 5 shows a complete set of classes which have been derived as a result of object conceptualisations and the semantic from the Problem Domain. Therefore our colour coordination is still very important because it clearly associates functionality from the Use Case model in Figure 1 with potential choice of software components we may have in our architectural model.

Therefore there are three important messages we can read from the Class Diagram in Figure 5.

Firstly, classes associated with user interfaces (UI) ("view" objects in Sequence Diagrams) have been omitted for practical reasons. Class diagrams might not be the most important place where we discuss UI because of the specificity of objects, which model them. It is likely that these classes may not have operations. The languages used in the design and implementation of UI tend not to have computational power and it is wise to bear this in mind while modelling.

Secondly, class diagram in Figure 4 is a static model which shows cardinalities and associations between the classes. This is very important in data modelling and would be of interest when we create software components which store data repositories. This is the first place where we can look at data sharing and reusability of potential software components across the model and across any problem domain.

Finally, the class diagram is layered, classes are grouped in potential software components and therefore they give a first glimpse of our architectural model.

Readers might notice that the naming of potential software components comes very close to the functionality within Use Case models. Therefore, we should recognize the main functionality of our CPS through vertical sets of classes from Figure 5. Their layering signals that we have different types of classes, which appear within each vertical set, and will definitively have an impact on our architectural model.

The precision in the Class Diagram from Figure 5 indicates that it was not difficult, at this stage, to find and specify attributes and operations of these classes. They might not be essential for creating a software architectural style, but might help when defining architectural components.

Taking into account the complexity of our CPS, diagram 5 is easy to follow and it shows relationships with other UML diagrams. This is an important outcome of our modelling because it defines the main building blocks of our CPS. They are all abstract modelling elements but physical items in this CPS have also their representatives in our model.

SOFTWARE ARCHITECTURAL MODEL

Figure 6 is an overall software architecture of our CPS by showing its main layered software components.

Figure 6 is a direct conceptualization of the Class diagram from the Figure 5, it includes all important components derived from initial abstractions (objects) in earlier diagrams, it distinguishes between interfaces, computations and persistence and thus follows the MVC pattern and addresses separation of concerns in software development.

The architecture also shows shared repositories across vertical functionalities and explicitly implies the level of reusability. Layered and component based architectural style, would allow the replacement of its components without significantly affecting software applications generated from it, if the technology used for the deployment of our architectural model will support it.
EXCERPTS FROM THE IMPLEMENTATION

In this section, we illustrate how excerpts from Figure 6 have been deployed and relevant parts of the Class Diagram implemented. The implementation may be a complex process. It would require the deployment of our architectural components from Figure 6 and Class Diagrams from Figure 5 within a chosen technology environment. The vertical division of the architectural model would also indicate that differences in implementations are possible in the functionalities modelled within each vertical set of classes and components in Figures 5 and 6.

Due to space limitation, we can illustrate the deployment of Publication Component and the filtering performed with the Beacon component. These two examples are rather different. The first one, described in the first subsection, solely uses excerpts from the class diagram in order to model data and produce the code, which would be essential in the publication management. The second example, from the subsection which follows, is focused on the physical part of our CPS: it describes the implementation of “filter service” as a part of the functionality of “beacon manager”. In other words, the classes implemented in the second example, model the behavior of software components associated with the physical item of our CPS (a beacon attached to an aircraft part).
of composition are self-explanatory: publications must have folders in which we keep documentation and it has to have Table of Contents, which helps in finding the correct documentation in the filtering process.

Figures 8, and 9 are sequence diagrams which replace the actual code, i.e. they help with the understanding the nature and level of coding, which was generated for finding the adequate documentation and loading it.

**Filtering the Documents**

As previously mentioned, the implementation of this software component is different. We allow for aggregation, composition and generalization to be used in the static model from Figure 10, but the way we produced the code for the implementation, shown in Figure 11, emphasizes that a “physical device will be involved in the filtering process” (**BeaconManager** object in the sequence diagram).

There is another important part of the implementation which was difficult to show with UML concepts, but had an impact on them. The standard for filtering and document applicability is described by the S1000D for technical documentations. Following its particular structure, we could design a class structure, which would enable the interaction with the applicability in the publication.

**CHALLENGES IN THE RESEARCH**

It is important to note that challenges in this research were not related to the process of modelling and finding abstractions, which would generate usable software solutions for CPS. To contrary, the challenges were mostly related to the problems in the implementation of the functionality from the architectural model because of technology requirements.

There are two separate issues here.

Firstly, if we accept that our software application, generated from the architectural model, would run in mobile and wireless environments, then the model should be deployable in both iOS and Android. This means that the implementation should comply with the rules, which exist in both of these environments.

Secondly, **beacons** carry their own “semantics” defined by their manufacturers (e.g. Estimote integrates beacons with their SDK!), and no modelling principles could address them. They are not even part of non-functional requirements in software development: they belong to the process of deploying abstract architectural models in constantly changing integrated development environments, which do not have to recognize the exact nature of SDK for physical items.

At this stage, developers have a choice of either
- defining technology specific components within the architectural model, such as the one in Figure 6, or
- creating code according to the class diagrams in Figure 5, which in turn generated the architectural model.

In this research, we chose the latter, but there is no consensus amongst software developers on how to proceed with the deployment of architectural models according to technology requirements.

In order to show an extent of an ad-hoc problems, which appeared without warnings and significantly delayed the implementation process, we describe a problems associated with the implementation of the **BEACON COMPONENT** from Figure 6.

A **plug-in** for Estimote for the NativeScript framework existed, but the wrapper (i.e. plug-in for NativeScript), which is important for the cross platform programming, was not updated. This might not seem to be a significant obstacle, but the job for remedying the problem was not trivial. Open source software providers do not offer warnings and if they do not update their software and documentation, this has become a norm in software development. The lack of updates in widely used open source software tools may sometimes seriously jeopardize any implementation.

In this particular case, the only option (as of April 2017) was to create a similar environment in which we could
manipulate data generated by the beacon and create our CPS. Before the data generated by beacons were prepared and “Beacon View” created, the integration of the Estimote SDK had to be performed in a newly generated environment, in which the ranging of beacons for iOS and Android was feasible. In the following 5 paragraphs we illustrate this part of the implementation.

Estimote Plugin - Estimote has their own SDK provided for both Android and iOS and uses the native libraries for both platforms. The lack of a cross platform wrapper would require to develop two separate applications, using the platform’s own tools. To avoid this, NativeScript, which is a cross platform framework, that wraps around the native libraries, has been used. A new plug was created, which integrates the Estimote SDK with Android and iOS.

Beacon Detection for iOS - An Estimote class for beacon ranging was created. The BeaconManagerDelegate class gets the updates for the beaconManager. When an Estimote object was initialized, beaconManagers sets delegate to BeaconManagerDelegate in TypeScript. This enables updates for the beaconManager. When an Estimote object ranging was created. The BeaconManagerDelegate class gets the updates for the beaconManager.

Before the data generated by beacons were prepared and “Beacon View” created, the integration of the Estimote SDK had to be performed in a newly generated environment, in which the ranging of beacons for iOS and Android was feasible. In the following 5 paragraphs we illustrate this part of the implementation.

CONCLUSIONS

This paper illustrates the feasibility of creating a CPS by using well established software engineering principles and modelling with UML. The initial goal, to create a reusable software architectures have been achieved because of the MVC pattern: the model separated computational components from the persistence, which in turn allowed changes in the content of Beacon data, Applicability Documents and Publications, without affecting applications generated from the architectural model. This means that any changes in

- the way we filter the documents,
- publications which comprise the documentation and devices/physical items which are part of the CPS

will not affect our architecture and its deployment.

We tested this model, and the application generated from it, in three different environments. The first test involved the documentation, which contained “mock data” from the aviation industry. This was done in order to detect problems triggered by the choice of technologies and open source software tools. At this stage we tested the feasibility of the

value (1675-12907 in Figure 13), important for the management of multiple beacons within one application, we used them for finding which beacon appears in our range. Everything inside the curly brackets are information about the exact beacon (formatted as JSON object).

Beacon View - The beacon view contains the UI of the beacon part of the application. A list of beacon objects are displayed in a list picker. When the beacon view is created, a list of nearby beacons from the BeaconRangingService is available every few seconds. The BeaconRangingService starts the Estimote ranging and updates a list of beacons for each discovery. When the user chooses a beacon from the picker list, the beacon view returns the selected beacon object, which in turn is used for filtering the document.

CONCLUSIONS

This paper illustrates the feasibility of creating a CPS by using well established software engineering principles and modelling with UML. The initial goal, to create a reusable software architectures have been achieved because of the MVC pattern: the model separated computational components from the persistence, which in turn allowed changes in the content of Beacon data, Applicability Documents and Publications, without affecting applications generated from the architectural model. This means that any changes in

- the way we filter the documents,
- publications which comprise the documentation and devices/physical items which are part of the CPS

will not affect our architecture and its deployment.

We tested this model, and the application generated from it, in three different environments. The first test involved the documentation, which contained “mock data” from the aviation industry. This was done in order to detect problems triggered by the choice of technologies and open source software tools. At this stage we tested the feasibility of the
solution and resolved all the challenges and obstacles detected during the implementation. It also served as a “proof of concept”. The second test included a real set of documentation applicable to the DC-9 aircraft. The last test was designed for a public demonstration of this research by creating an ad-hoc CPS space, within a university environment, in which we attached beacons to a mountain bike placed in a one of the room within the University buildings. In all three cases, the performance of the applications on both, Android and Apple phone was immaculate.

The message to research communities/industry, interested in the development of modern CPS is that

- CPS and physical items, which comprise them could be modelled and deployed as software intensive systems, and use software engineering principles with a set of UML models, if we wish to generate CPS, from reusable, layered and component based software architectures.

- Technology choices for the deployment of software architectures for CPS have and will have serious impact on the application development. The main culprits are the open source tools, integrated development environments, which are not necessarily updated, and mobile/wireless operating environments, such as Android and iOS, which are still competing with each other and insist on their heterogeneity.

All the challenges point towards the problems with technologies. It was difficult to integrate ESTIMOTE with iOS; iOS would not allow the styling of the documents, which might have been caused by problems between iOS and NativeScript, and the combination of tools triggered problems when accessing beacon’s data due to discrepancies and inconsistencies in SDKs, and in operating systems in which we created the application. This proves that the main outcome of this research is in line with finding of (Hartung et al., 2015).

There is a very short step between the current version of the application and its commercialization. Our future work should include the automation in generating the Beacon View component, addressing the size of documents and their folders when loading them in iOS and Android and the technology trends in creating SDKs of physical items, which may comprise CPS.

The most important outcome of this research still remains that the specificity of the semantics of physical components in CPS can be address through either data they generate, as explained in Figure 6, or technology requirements they carry within themselves, as demonstrated in the sections on the Implementation and Challenges. They were the main reasons behind the successful creation of this type of a CPS, following our modelling and development principles, which apply to any software intensive systems.

REFERENCES


Ebling “Satya Revisits “Pervasive Computing: Vision and Challenges” Maria R. Ebling, IBM T.J. Watson Research Center Roy Want, Google


Gamma, E., Helm, R., Johnson, R. Vlissides, J., Booch, G. (1994) Design Patterns: Elements of Reusable
Object-Oriented Software, Addison-Wesley Professional; November 10, 1994


AUTONOMOUS SECURITY SYSTEM SPECIFICATION WITHIN OIL AND GAS INDUSTRY FOR OFFSHORE VESSEL: REQUIREMENT AND CONCEPT

Hugues Broyde, Kristin Falk, Aurilla Aurélie Bechina Arntzen
Faculty of Technology and Maritime Science, Kongsberg, University College of SouthEast Norway,
{ hugh.adrien@gmail.com kristin.falk@usn.no aurillaa@usn.no }

ABSTRACT

The maritime industry is encountering stringent challenges. Most specifically, security systems design for offshore vessels in the oil and gas industry is quite a demanding task. The last decade, such security systems encompasses a dynamic positioning and a related emergency quick disconnect systems. The security system could disconnect the vessel from the reservoir in case of an emergency. The decision to perform a disconnection is a complex task; and might generate unnecessary costs for the drilling operation. This paper aims to provide general framework that will contribute to designing a decision system helping the operator to decide whether to activate the emergency quick disconnect system (EQD). In order to achieve this goal, we analyze previous incidents to delinate an architecture of such decision systems. We have conducted a systematic review of two decades of incidents reported by the International Maritime Contractor Association (IMCA) and Petrobras. Our results show that a high number of activations of the safety system was unnecessary, and were rather a result of human manipulation. Therefore, we recommend delineating concepts of an autonomous system that could help with the decision to disconnect.

INTRODUCTION

The tragic incident in April 2010 with the offshore drilling platform Macondo where a number of persons died, and oil spilled towards the coasts is a sad example that shows a need for more reliable systems. In 2007 an operator pushed the surge buttons accidently with the elbow leading to a disconnection of the reservoir, and therefore generate huge costs to reconnect the vessels to the well (Akgun, Byrne et al. (2005)). These examples among several others, demonstrated the need for an autonomous system that could protect the vessel of human manipulation errors. Such system should support the operator with the decision to connect or disconnect to the well in order to avoid unnecessary cost rise or disasters. Autonomous systems are the next technological revolution for the maritime industry (Hurmelinna, Kylaheiko et al. ). These systems incorporate four main features: Self-Configuration, Self-Healing, Self-Optimization, and Self-Protection (Yahya, Yahya et al. 2013). Such systems can answer the increasing demand for more secure systems both for operators, but also for the environment and other stakeholders. Dynamic positioning (DP) systems are control systems used to keep the floating offshore vessel in a fixed position during the drilling operation trough thrusters on the vessel. There are different classes of DP vessels (Hurmelinna, Kylaheiko et al.). During a drilling or intervention operation using a DP system, only a marine riser connects the rig (floating drilling unit) to the subsea well possibly containing hydrocarbons. See illustration in Figure 1. The main challenge for a DP system during the drilling operation is that the vessel could drift off location.

![Figure 1: Illustration of DP drilling operation (Displaying the operational limits: green, yellow, red, and purple. Note that figure is not to scale).](image-url)
The EQD system acts as a safety barrier. The aim of an EQD system is to prevent structural damage to the wellhead, environmental spills, human casualty and thus increase operational cost (Chen, Moan et al. 2008). The purple line in Figure 1, illustrates the operational limit when the riser or well can no longer withstand the forces. There can be different causes leading up to an EQD activation, including the environment, human error or problem with management of the power of the vessel (Marz, Friedrich-Nishio et al. 2006).

Another incident reported in IMCA (IMCA 2006), highlighted that bad weather such as high current and strong wind could lead a vessel to disconnect. In that specific case, the main issue was related to the fact that one of the generators was shut down. We will discuss the causes of this disconnection later in this paper.

The different mode of triggering an EQD system are listed as below (Su, Chen et al.):
- manual procedure,
- semi-automatic
- automatic with manual override
- and purely automatic.

Manual procedure is the most standard approach. A DP operator pushes a yellow or a red alert button to indicate to the rest of the crew that the vessel operates in the yellow or red emergency state, respectively. When the status indicates that there is red alert, a dedicated person (normally related to drilling or vessel safety) pushes the EQD activation button that automatically should lead to a safe disconnect.

The loss of position is not a common event. For DP vessels there is approximately $10^3$ loss of position per DP hours. Therefore the loss of position for a specific vessel can occur between $10$ to $100$ years and can be considered as an exceptional event. A DP operator will have not enough experience about the loss of position and might not take the most appropriate action (Chen, Moan et al. 2008). This can result in late activation of the EQD as decision should be taken in a very short time frame as illustrated in figure 6.

Taking the decision on whether to disconnect is not straightforward since, on one hand, there is a risk of not being able to disconnect if you are too late, and on the other hand, each disconnection has a significant cost. The cost is several million USD (Hall and Andriani 2003), and in the worst case a disconnection might result in losing the well with a cost of up to a billion USD. Related to special operations, there is also a risk to personnel and the environment if activating the EQD system. Our research study focuses first on identifying the possible causes that conducted to the incidents reporting whether the EQD system was activated or not. To this end, we analyzed the database of IMCA and Petrobas that gather annually knowledge on EQD and DP systems (Johannessen, Olsen et al. 1999). The main purpose of this analysis is to delineate a framework encompassing requirements for defining an autonomous system that could help the decision making for the activation or not of the EQD system.

METHODOLOGY

Our qualitative research is based on both literature reviews, case studies, and on careful analysis of disconnection incidents from DP vessels reported to both IMCA and Petrobas database. IMCA facilitates the sharing of information on potential hazards. These reports are publically available, but anonymized. Most oil companies will record their DP incidents. Yet, we could find only one collection of DP incidents from a single company, namely from Petrobas operating primarily in Brazil. Our study is thus based on the analysis of these two reports: IMCA and Petrobas. This might constitute a limitation of our study.

We used directly the classifications provided by IMCA and Petrobas in our research study. Note that some of the labels or categories were identical in the two databases such as human errors and environment, while other categories were different such as poor maintenance and power management. See details in (Johannessen, Olsen et al. 1999) and (Malone 2002).

IMCA database

The annual IMCA reports are based on voluntary and anonymous incident reports from DP vessels (Marz, Friedrich-Nishio et al. 2006). These reports constitute a database that the maritime industry can use in order to improve safety and its operations. In our study, we analyzed 17 years of reporting from year 2000 to the year 2016. We have investigated in depth the 1171 incidents reported to IMCA.

Each of the disconnection incidents in the IMCA reports indicates a main cause and a secondary cause (Marz, Friedrich-Nishio et al. 2006). The primary cause is the main context of the incident. An example is a storm hitting a DP vessel. The secondary cause is an event not directly responsible for the incident, but a cause that will make it worse such as a generator failing during a storm. The process of identifying the cause of the incident was described as follow: search for relevant keywords, discriminated the results, refine the search, and then repeat the process.

Petrobas database

The available data from Petrobas contained 571 incidents from year 1992 to 2005. The data is stored in a database specific to Petrobras (Johannessen, Olsen et al. 1999). This kind of company report is rare to find in the public domain. Most oil companies do not disclose this kind of reports.

RESULTS AND ANALYSIS OF STATISTICAL DATA

Table 1 shows the total number of reported incidents in the two databases IMCA and Petrobas. The term “State”, in the first column, refers to the severity of the incident. The row named “Degraded Status Criteria” is a subdivision of the Yellow alert used by Petrobas, but not by IMCA (Terziovski
and Morgan 2006). Petrobras has its own procedures and means to resolve situations of this category.

Table 1: Number of incidents reported through IMCA [2000-2016], and by Petrobras [1992-2005] classified in terms of state and reported in actual numbers, and average number of incidents per year (in parenthesis).

<table>
<thead>
<tr>
<th>State</th>
<th>Total number of reported incidents and average per year (in parenthesis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degraded status (Petrobras)</td>
<td>- 305 (21.8)</td>
</tr>
<tr>
<td>Yellow alert</td>
<td>172 (10.1) 69 (4.9)</td>
</tr>
<tr>
<td>Yellow alert &amp; Degraded status (Fusion of the two above data sets)</td>
<td>245 (14.4) 374 (26.7)</td>
</tr>
<tr>
<td>Red alert</td>
<td>72 (4.2) 105 (7.5)</td>
</tr>
<tr>
<td>Disconnected</td>
<td>44 (2.5) 29 (2)</td>
</tr>
<tr>
<td>Black-Out</td>
<td>50 (2.9) 45 (3.2)</td>
</tr>
<tr>
<td>Total incidents reported</td>
<td>1171 (68.9) 479 (34.2)</td>
</tr>
</tbody>
</table>

The first observation we can make based on this table is that Petrobras reports in average more yellow and red alerts than IMCA. For disconnection and blackout, the two datasets are similar. This might be a result of the philosophy applied by Petrobras; that they handle problems before they escalate, thus give alert more easily than other companies do.

Amongst the more than one thousand incidents reported in the IMCA reports, only 44 led to an actual disconnection of the vessel from the well. For Petrobras only 29 incidents out of 571 led to a disconnection.

Figure 2, Figure 3 and Figure 4 present Pareto diagrams of the main causes of incidents. Figure 2 presents the actual number of causes that lead to a real disconnect, while Figure 3 and Figure 4 present causes that lead to a red alert.

The chart in Figure 2 displays the three main causes from the IMCA data leading to a disconnect, that is environment (25%), power management (18%) and human error (16%). Here the environment refers to the weather condition or the state of the sea. The power management represents the problem related to the power of the rig. For example, a failure of a generator is a problem of power management. Moreover, secondary and main causes seem not to be correlated. Two of the secondary causes have very high importance, namely design and procedure (referring to the lack of procedure for a given situation).

Figure 2: Data from IMCA 2000-2016 displaying the different root causes (in percentage) resulting in activating disconnection of the rig. The blue bars represent the main causes and the orange the secondary causes.

Figure 3 is based on data from Petrobras, and displays root causes of red alarms. There are slight differences between the data collected in 1992/2005 and 2000/2005. The two most remarkable differences are the increase of human error and the decrease of environment incident. It is difficult to explain the decrease. Maybe the reason is a change of policy in Petrobras related to these incidents, or there might have been fewer storms and less weather during the second period. Petrobras reacted to the trend of increasing number of human errors by producing new policies (Johannessen, Olsen et al. 1999).

IMCA did also try to trend the different root-cause categories as a function of time (Huang, Ou et al.). Unfortunately, there was no structured trends, and the data was not possible to analyze. The data analysis is completely dependent on the input that the oil and gas companies give to IMCA, and that dataset is far from complete.

Figure 4 shows IMCA data displaying the different root causes resulting in activating red alert on the rig. There is a significant difference between Figure 2 and Figure 4 representing disconnections and red alerts from the same database.

The environment that was the most significant cause of disconnections is only the fifth most common cause of a red alert. To explain this we assume that a vessel is able to handle unwanted situations in calm weather. On the other hand, in bad weather with the same unwanted situation, there is less time to resolve the situation, and thus an increased risk of disconnection. Thus, even if there is only a few environmental conditions leading to a red alert, most of these can easily lead to a disconnection.
Comparing the Petrobras data with the IMCA data shows one problem of power management as the environment. This can bring a bias compared to IMCA.

**Power management** (In Figure 2 and Figure 4) relates to the power of the rig. For example, a failure of a generator is a problem of power management. The Petrobras data does not show this item. Maybe some of these problems have been included in the maintenance and component failure category. If that is the case, Power management is an important cause also in the Petrobras cases.

**Human error** is the only cause that is common to all the three diagrams above. Human error represents 24% and 30% of the root causes in the Petrobras datasets, while it represents only 17% of the main cause in the IMCA dataset, both representing red alerts. Why does the IMCA report contain relatively fewer human errors?

Authors of one of the IMCA reports (Adamides and Karacapilidis 2006) state that the DP operators try to hide their responsibility by highlighting others causes. In some cases, reports are not even submitted to IMCA. Therefore, it is probable that data of the IMCA might not fully accurate or are missing (Johannessen, Olaisen et al. 2001). Some regions are overrepresented in the IMCA database, and while others are rather underrepresented.

According to Deegan (Hall and Andriani 2003): “The frequency (of loss of position) is one of the areas of the greatest uncertainty because of the lack of good quality data associated with DP drilling vessels”. For example, the Petrobras data reports that seven disconnections failed (Johannessen, Olsen et al. 1999). But the counterpart, IMCA data, show only one failure (Akgun, Byrne et al. 2005). We can doubt that Petrobras is the only company that encounters these problems. Nevertheless, the data we have analyzed is quite extensive, and gives a good indication on the complexity of the systems and operations. Furthermore, the data show that the systems and procedures used on DP vessels do not prevent all incidents.

Moreover, the secondary cause highlighted by the IMCA data shows that the lack of procedure and problem of design are leading root causes, especially for the disconnection. We can try to explain this by the fact that when an emergency occurs, people do not have enough training/experience (emergency situation are rare event) and the lack of procedure will make the situation worse (Dougherty, Borrelli et al. 2000). For serious incidents such as a disconnection, it is difficult to create precise procedure. This is different to a red alert, and the diagrams indicates this difference. Procedures were responsible for only 11% of red alerts, but this number increased to 33% when the situation lead to a disconnection.

In summary, we can claim that one of the main causes for disconnection is the environment. Figure 5 below illustrates the secondary cause of disconnections. The term procedure includes wrong application of procedures, non-application of procedures, poor or lacking procedures, and represents thirty present of the incidents. Poor procedure and human error result in more than 50% of the cause leading to a disconnection. We conclude that the human factor contribute to unwanted situation even if the main cause is the environment.
The DP operator has a key role in the DP operation. Obviously, he or she plays an important role when an incident occurs. Therefore, in order to assist the DP operator, it is important to provide, in real time, information about the current situation such as the environment, weather forecast, status of the machines and so forth. A system that could contribute to help the operator to take the right decision could be designated and is part of the next step of this ongoing research study.

CONCLUSION

This paper presents an analysis of causes that has led to disconnections of dynamically positioned vessels during offshore drilling operations. To this end, we have carefully analyzed incidents reported in IMCA and Petrobras databases. The analysis suggests that human errors are the leading causes. In order to alleviate this issue, we suggest building a system that could support the operator in the decision of whether to disconnect from the well or not. This is an ongoing work, and is the first step that will lead to the specification of a future autonomous system. The study presents some limitation due mainly to the quality and accuracy of available data.

AKNOWLEDGEMENT

Markom2020 supports this work. Markom2020 is a development project with the intention to increase maritime competency in Norway.

REFERENCES