Computing and Informatics, Vol. 40, 2021, 249–276, doi: 10.31577/cai_2021_2_249

UNIFIED ABSTRACT MECHANISM TO MODEL LANGUAGE LEARNING ACTIVITIES

Gabriel Sebastián

Albacete Research Institute of Informatics
University of Castilla-La Mancha
Campus Universitario, s/n, 02071 Albacete, Spain
e-mail: gabriel.sebastian@uclm.es

Ricardo Tesoriero, Jose A. Gallud

Faculty of Computer Science Engineering
University of Castilla-La Mancha
Campus Universitario, s/n, 02071 Albacete, Spain
e-mail: {ricardo.tesoriero, jose.gallud}@uclm.es

Abstract. Language learning applications define exercises that are pedagogical tools to introduce new language concepts. The development of this type of applications is complex due to the diversity of language learning methodologies, the variety of execution environments and the number of different technologies that can be used. This article proposes a conceptual model to develop the activities of language learning applications. It defines a new abstraction mechanism to model these activities as part of a model-driven approach to develop applications supporting different language learning processes running on different hardware and software platforms. We define a metamodel that describes the entities and relationships representing language learning activities as well as a series of examples that use the proposed abstraction mechanism to represent different language learning activities. The modelling process is simplified using a common representation that does not affect neither the visual presentation, nor the interaction of each activity. The article includes an evaluation that analyses the product correctness, robustness, extensibility, and reusability of the obtained code. These results conclude that the code generated using the proposed approach overcomes the code generated following a traditional approach.

Keywords: Model-driven development, languages learning methodologies, web technologies

1 INTRODUCTION

Language learning applications development is complex due to the diversity of learning language methodologies, the variety of execution environments (Web, mobile and desktop) and the number of different technologies that can be used [9].

Besides, the development of learning exercises to implement a language learning application is a repetitive and tedious process. The process involves repeating the same resource management tasks many times having duplicated code, which is difficult to maintain. Moreover, the problem of duplicated code and task repetition is multiplied by the number of different target platforms, making the overall process more complex and prone to errors. This approach, also known as the traditional approach, can be improved using a Model-Driven Architecture (MDA) to capture common features in Computation Independent Models (CIMs).

To capture this common features, we performed an analysis of the Lexiway[1] language learning methodology and we experienced the following problems. Initially, the client requested the development of a mobile application for the iOS platform. Later on, they requested a Web version of the same application. Therefore, the development team adapted software resources to produce a new source project for the new version of the application.

This new development scenario consisted of two independent branches for the same project which leads to divergent resources and a source code. The resulting environment was difficult to maintain, impacting negatively on subsequent projects. For instance, the development of an Android version of the application was abandoned due to the high development costs.

The aim of a learning activity is teaching a concept by means of an interactive experience. Learning activities employ different interaction mechanisms such as fill-in the gaps (see the right side of Figure 4), joining the lines, Drag & Drop images (see the right side of Figure 5), and so on.

The traditional software development approach usually forces us to manually generate the different learning activities or exercises, with their corresponding media resources (audio, images or video). For example, the JUNIOR 1 level of the Lexiway learning methodology consists of 6 different blocks composed of 4 units each. In addition, each unit consists of 2 lessons containing 12 words. Managing all these learning activities manually involves a high level of resource duplication which leads to software validation difficult to manage.

[1] `https://www.facebook.com/LexiwayLearning`

Thus, following the traditional approach to develop multimedia interactive learning applications, the idea is to develop several prototypical games. For each activity type, a configuration file or a database register is manually defined. This configuration file defines the data and resources required by the logic layer of the activity to be executed.

Finally, another conventional aspect in the development process of interactive learning applications is that the navigation among activities and the progression of the level of difficulty is controlled by complex conditional structures that are difficult to manage and maintain which usually became a source of problems.

From the experience achieved during several years of developing language learning applications, we have learnt that the use of software artefacts (i.e. components, frameworks) and performing repetitive tasks reduces considerably the development time and costs. There are two conceptual tools that software engineering has traditionally employed to accomplish these challenges: increasing the level of abstraction and reuse.

The development of learning activities requires the specification of a great variety of aspects. Among the most relevant of them, we mention:

- the concept structure to be learnt,
- the media resources employed to represent these concepts,
- the mechanisms to manage, link and present these concepts,
- the activity workflow that should be followed to learn these concepts.

From the development perspective, all learning activities are different; however, they could share common aspects. For instance, different activities could employ the same interaction mechanisms (e.g. fill-in the gaps or joining concepts) to teach completely different concepts (e.g. fruits, vegetables, transportation, etc.).

This article proposes a conceptual model to develop activities in language learning applications. In particular, the article presents a new abstraction mechanism that allows designers to use (and reuse) the same model to represent many different learning activities, which have been taken from the learning methodologies under study. It defines the "fill-in the gaps" activity as universal abstraction to represent different kinds of activities. This abstraction is the basis of a model-driven approach to develop activities for different language learning methodologies.

This article also shows a series of examples where the "fill-in the gaps" abstraction is used to represent different kinds of activities for different learning language methodologies. Thus, the modelling process is simplified, since every activity is modelled using a common representation which favours the reuse of models. It is worth to note that this abstraction does not affect neither the visual presentation, nor the interaction of each activity.

This paper is organized as follows. Section 2 describes the research context together with the related work. Section 3 briefly describes the metamodel where

the "fill-in the gaps" abstraction is defined. Section 4 analyses a set of additional modelling capabilities derived from this proposal. Section 5 describes the users' evaluation carried out to evaluate the quality in use of our proposal as well as the product quality. Finally, we present conclusions and future works.

2 RESEARCH CONTEXT

This section contains the research context regarding the development of language learning applications which includes two main elements: the essential concepts and features extracted from different learning methodologies to abstract the language learning process, and the most relevant related work in the field of the model-driven development which was employed to tackle the problems described in Section 1.

Learning a foreign language is a process involving different methods, techniques and tools, each one appearing to be more effective than the others. In this paper, we focus on methodologies that offer some type of technological support (Web site, mobile applications or similar).

We have analysed the following methodologies: Lexiway[2], Duolingo[3], Babbel[4] and Busuu[5].

Although these methodologies take different approaches, it is possible to identify some common elements.

The Model-driven Architecture (MDAs)[6] approach proposed by the Object Management Group (OMG) in 2011 presents a set of tools to abstract these common elements to improve the software development. This solution gives a leading role to models in the software development during all phases (i.e. inception, design, building, development, and maintenance).

The main reason behind this approach is the constant evolution of the software technologies. Following a traditional development approach, the functionality code and the implementation technology code are interweaved. Consequently, when the technology is enhanced, the functionality is rewritten using the new technology.

Under these scenarios, MDAs introduce abstraction levels to promote the software reuse by emphasizing the design-time interoperability [20]. This kind of interoperability is possible due to the specification of Platform Independent Models (PIMs) that enable developers to separate the specification of the application functionality from the technology that implements it.

[2] https://www.facebook.com/LexiwayLearning
[3] https://www.duolingo.com/
[4] https://www.babbel.com/
[5] https://www.busuu.com/
[6] http://www.omg.org/mda/

Thus, it is possible to reuse the specification of the application functionality for different implementation technologies. Moreover, this functionality can be executed on different hardware and software platforms only with minor changes.

The source code of applications is automatically derived from models using model transformations [16].

In summary, the use of the MDA technology enables the generation of multi-platform applications from PIMs. This fact leads to several advantages; for example, let us assume we want to develop a learning activity using the fill-in the gaps interaction mechanism for different platforms (e.g. iOS, Web and Android). Following a traditional approach, we should develop 3 different and independent source code projects. Following an MDA approach, we specify only one PIM to generate the source code for the 3 platforms.

The core of the MDA infrastructure is defined in terms of the following OMG standards: the Unified Modeling Language (UML)[7], the Meta Object Facility (MOF)[8], XML Metadata Interchange (XMI)[9] and the Common Warehouse Metamodel (CWM)[10] which were successfully used in the modelling and development of modern systems.

From the Human-Computer Interaction perspective, we can find different approaches that make use of models to generate user interfaces.

Hence, since our work focuses on the development of interactive systems, the Model-based User Interface Development (MbUID) provides useful elements to analyse based on the CAMELEON Reference Framework (CRF) [5].

In recent years, other approaches such as [11] have also encouraged the use of models to develop multi-modal user interfaces.

The use of Model-driven Development (MDD) for learning applications was applied in different works, such as those exposed in [8, 2, 18]. However, none of them formalizes the definition of language learning activities using OMG compliant metamodels. Nevertheless, there are several works that use MDD techniques based on MDAs to develop Web applications [17, 3].

An interesting approach that defines a MDA to develop music learning applications is exposed in [26]. In particular, in this approach a MDA-based System Development Lifecycle is defined, three Learn-Models are built, and the important developing phases are described. The idea of using submodels in a complex metamodel has been applied in our work. And a methodology to model e-learning applications can be found in [10]; however, this methodology does not focus on developing language learning applications. Unlike this approach, our approach presents

[7] `http://www.omg.org/spec/UML/2.5/PDF`
[8] `http://www.omg.org/spec/MOF/2.5.1/PDF`
[9] `http://www.omg.org/spec/XMI/2.5.1/PDF`
[10] `http://www.omg.org/spec/CWM/1.1`

a set of models at different abstraction levels providing different points of view of the application depending on the level of abstraction.

A work where the experience of different research groups working in formal and informal learning language design using mobile devices is presented in [1]. Among the most relevant works regarding the development of e-learning applications we can find those presented in [13, 19, 21, 24, 9].

With regard to the use of models to build Web sites, some methodologies (e.g. [7]) and models (e.g. RMM [14], WebML [6] have a direct impact on this research because they focus on modelling applications at the software level. However, our interest is focused on higher level of abstraction where the learning activity is the centre of our modelling interest.

WebML enables to define the high-level description of Web sites considering several orthogonal dimensions. For instance, the Web site contents (i.e. structural model), the Web pages that compose the Web site (i.e. composition model), the link topology among Web pages (i.e. presentation model), and the personalization characteristics enabling the one-on-one content delivery (i.e. personalization model).

The standard Interaction Flow Modeling Language (IFML)[11] is designed for expressing the content, user interaction and control behaviour of the front-end of software applications in general. In [4], authors describe how to apply model-driven techniques to the problem of designing the front end of software applications (i.e. the user interaction).

Our approach proposes a domain specific language based on the definition of a set of models of a higher level of abstraction presented in [23]. These models represent the interaction techniques used in language learning activities to maximize code reuse and minimize maintenance costs. This article presents the abstraction mechanism that allows designers to use (and reuse) the same model to represent many different learning activities, which have been taken from the learning methodologies under study.

3 METAMODEL TO DEVELOP LANGUAGE LEARNING APPLICATIONS

The goal of this article is to create the definition of a common representation to model language learning applications. To accomplish this goal, this section describes a metamodel that supports the representation of this type of applications. This metamodel is based on the metamodel presented in [22]. The description presents the modelling concepts to define applications as well as a set of examples showing how these concepts are assembled. The formalization of the concepts and the relationships among them were defined in ECORE[12] (Essential MOF dialect[13] enriched

[11] http://www.omg.org/spec/IFML/1.0/

[12] https://wiki.eclipse.org/Ecore

[13] http://www.omg.org/spec/MOF/2.5.1/PDF

with expressions in OCL[14]. As a result of the analysis of different methodologies, we have found a set of common elements.

The first element in common is the definition language concepts (e.g. words, sentences, etc.) that are hierarchically organized in lessons, units, etc.

The second element is the definition of different representations for these concepts. These representations are media resources (e.g. images, audio recordings, videos, text, etc.) that can be associated to methodology concepts. For instance, the "house" concept can be associated to an image of a house, a video of a house or an audio recording that contains the voice of a person pronouncing the word "house".

The third element is the definition of activities enabling users to interact with the application. There are several types of activities; for instance, multiple choice, filling the gaps and sentence composition activities.

The fourth element to take into account is the order in which the activities should be performed by the user. For instance, some methodologies only enable students to start a lesson if they have passed the previous one. The order in which activities are carried out is known as the methodology workflow. In summary, the common elements of language learning applications are:

1. the language concepts and concept hierarchy,
2. the media resources,
3. the learning activities,
4. the activity workflow.

To model these common elements, we leverage the level of abstraction and reuse following the MDA principle of interoperability at design time. Therefore, we define four concerns regarding the modelling of language learning methodologies.

Thus a learning methodology (represented by an instance of the *Methodology* metaclass that is part of the *Methodology* package) is composed by 4 models.

The language concept model, representing the language concepts, is modeled by an instance of the *ContentContainer* metaclass that is part of the *Content* package. The media resource model representing the resources used in the presentation (or view) of learning activitites, is modeled by an instance of the *MediaModel* metaclass that is part of the *Media* package. The model that defines the set of learning activities is represented by an instance of the *ViewModel* that is part of the *Presentation* package. And the activity workflow model is represented by an instance of the *Workflow* metaclass that is part of the *Workflow* package.

Figure 1 shows the proposed metamodel exposing the packages of metaclasses that represent the language learning methodology common elements. The pack-

[14] `http://www.omg.org/spec/OCL/2.4`

age structure for this metamodel is based on the *Methodology* package which uses the *Commons* package containing the *Entity* and *Property* metaclasses that are used as super-metaclasses for all metaclasses in the metamodel. The *Methodology* package is the core package of the model architecture, and contains the Methodology metaclass, and a set of packages that contains the metaclasses to represent all models (*Workflow*, *Content*, *Media*, *Activity* and *Presentation*). The sixth package, aka the *Commons* package, provides metamodel entities with extension features. Next paragraphs explain the most relevant metaclasses of each package.

The language concepts and the concept hierarchy are represented by instances of the metaclasses defined in the *Content* package. Figure 1 presents the *Concept* and *ContentContainer* metaclasses of *Content* package. While *Concept* metaclass instances represent simple concepts, such as nouns (e.g. orange, apple, peaches) and verbs (i.e. stare, watch, glance); *ContentContainer* metaclass instances represent sets of related concepts (e.g. fruits, ways of looking, etc.).

For example, in a given methodology, a level can be composed of units and, a unit can be composed of lessons.

The *Media* package contains the metaclasses to represent the resources used in the presentation (or view) of learning activities. It enables developers to define 4 types of media resources: audio, text, video and image. These types of media are represented by *Audio*, *Text*, *Video* and *Image* metaclass instances. Combinations of this type of media can be combined to represent complex media resources such as text and speech. Again, the Composite design pattern [12] is applied to create a tree-based structure of media resources. The *MediaContent* metaclass plays the role of Component, the *ComposedContent* plays the role of Composite and the *Audio*, *Video*, *Image* and *Text* meclasses play the role of Leaves. The media relationship between the *Concept* metaclass and the *MediaContent* metaclass associates language concepts to media resources to provide these concepts with a concrete representation to activity presentations.

The definition of the activities of a learning methodology is organized into the *Activity* and the *Presentation* packages. On the one hand, the *Activity* package enables developers to parametrize the functionality of the activities conducted during the learning process. Every *Activity* metaclass instance provides users with information to perform learning activities. While *Ground* metaclass instances define activity statements represented by *MediaContent* metaclass instances; *Gap* metaclass instances define the information to be introduced by students. *Gap* metaclass instances are enriched with *Option* metaclass instances to define the potential information (including the correct answer to the activity) to be introduced by students. *Option* metaclass instances are related to language concepts that are linked to instances of the *MediaContent* metaclass which provide the presentation of the concept in the activity.

On the other hand, the *Presentation* package defines the *ViewModel* and the *Slide* metaclasses to represent the user interface and interaction mechanism of the activities offered to the students during the learning process. This package defines

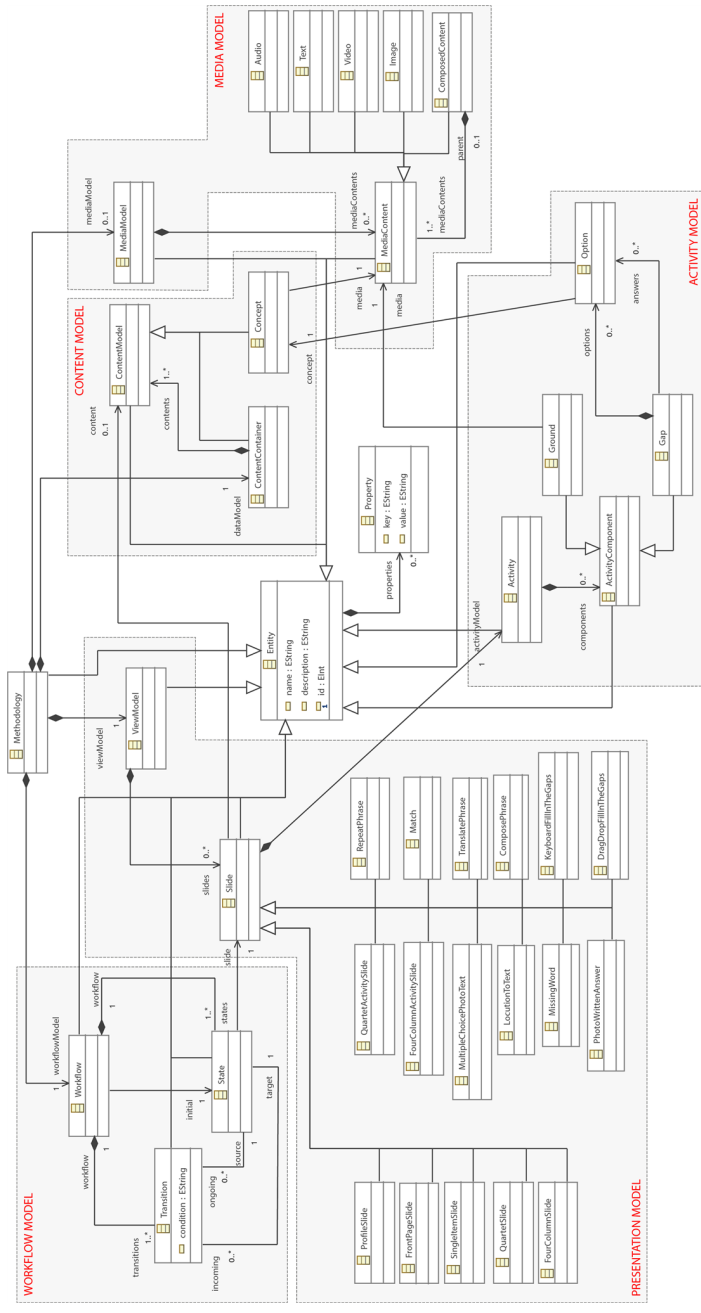


Figure 1. Language learning methodology metamodel

the concept of slide. It is defined by *Slide* metaclass instances that associate activities represented by *Activity* metaclass instances defined from the *Activity* package to specific interaction mechanisms. Consequently, students can interact with the same activity information using different interaction mechanisms (modalities), and vice-versa.

For instance, users can identify fruits matching images using the drag and drop or joining with lines interaction mechanisms. While an *Activity* metaclass instance contains the text for the statement of the activity; the options to be presented to the user and the option that solves the statement are represented by instances of the *Option* metaclass. And a *MultipleChoicePhotoText* metaclass instance defines the interaction mechanism.

The activity workflow defines the order in which learning activities should be performed by students, which is a crucial issue in the definition of learning methodologies. The *Workflow* package is responsible for representing this aspect of learning methodologies. This package defines the *Workflow* metaclass, whose instances define graphs. The nodes of the graph are defined by instances of the *State* metaclass, which are associated to instances of the *Slide* that represent them. The edges of the graph represent transitions (i.e. *Transition* metaclass instances) between states leading to transitions between learning activities.

Finally, the *Methodology* package defines the *Methodology* metaclass representing all the elements that define learning methodologies.

All the metaclasses in this metamodel inherit from the *Entity* metaclass defined in the *Commons* package which provides identification (i.e. *Entity* metaclass) and extension (i.e. *Property* metaclass) features to the rest of the metaclasses. This package is designed to take into account variable aspects of the activitites (aesthetical customization, look and feel of the user interface, structural limitations defined by a methodology, etc.).

Additionally, we have developed a language learning methodology model editor to create, edit and verify models according to the proposed metamodel. It was developed as an Eclipse plugin employing the Eclipse Modeling Framework (EMF)[15] to follow the MDA OMG standards. The metamodel was defined in OclInEcore[16] which is a dialect of the OMG Essential Meta-Object Facility (EMOF) enriched with OCL. This language is used to define the model invariants and queries that are the foundations for model verification.

3.1 Analysis of Language Learning Activity Modelling

In [23] we illustrate the flexibility and adaptability of the proposed metamodel to represent different language learning methodologies. Figure 2 shows the correspondence among the different elements of the model of a multiple-choice learning activity Lexiway as well as Duolingo. Thus, Figure 2 illustrates the expressiveness power of

[15] `http://www.eclipse.org/modeling/emf/`

[16] `https://wiki.eclipse.org/OCL/OCLinEcore`

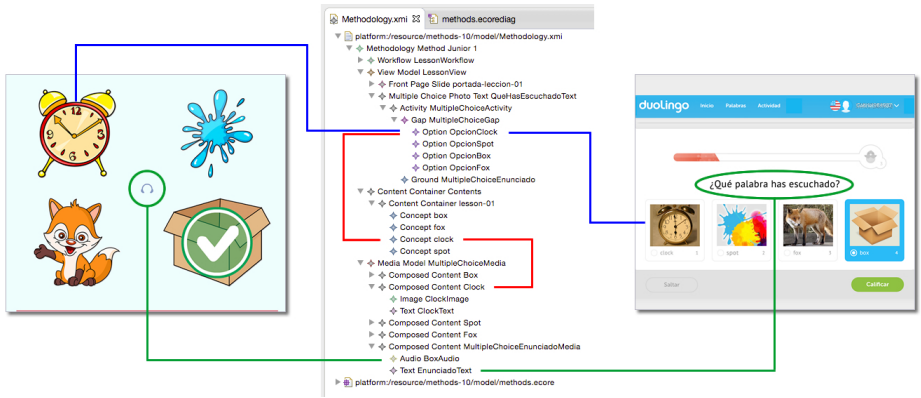


Figure 2. Reusing of the multiple-choice learning activity model for Lexiway and Duolingo methodologies

this modelling tool, since the same model represents two similar activities in two different learning methodologies.

As we have mentioned, the user interfaces of learning activities are defined in the presentation model which is an instance the *ViewModel* metaclass. Each type of user interface is defined by a *Slide* sub-metaclass.

Learning activity user interfaces are customized with information provided by the activity model represented by an instance of the *Activity* metaclass. *Activity* metaclass instances define two types of *ActivityComponent* metaclass instances that composes the definition of an activity model. *Ground* metaclass instances represent fixed parts of the activity (e.g. parts of sentences, audio recordings, videos, etc.). *Gap* metaclass instances represent the user inputs to introduce information in the activity (e.g. an input field to type a word, an input area to type a sentence, a combo box to select a word, a list to select an image, etc.). In any case, *Gap* metaclass instances define a set of *Option* metaclass instances that represent the options that are associated to list or combo box items as well as references to the set of options that are considered correct answers to the activity. *Gap* and *Option* metaclass instances are linked to *Concept* metaclass instances to provide *Slide* sub-metaclass instances with multi-modal user interface representations.

This modelling approach enables developers to reuse different media models in different activities as well as provide customized different learning activities with different looks. For instance, you could provide customized media resources to adapt the learning activities to colour-blind people.

Besides, this approach also enables developers to reuse the same activity model in interaction mechanisms. For instance, the *Match* and *MultipleChoicePhotoText Slide* metaclass instances reuse the same activity model to present the same activity employing different interaction techniques.

The learning methodology activity workflow model enables developers to reuse learning activities in different learning paths. For instance, developers reuse the learning activities defined for the workflow of the Standard version of Lexiway in the workflow of the Junior version of Lexiway because the main difference between these two versions lays on the number of concepts that are presented to students on each lesson.

Finally, we expose different ways to extend the proposed metamodel. Firstly, the *Slide* metaclass can be extend to introduce new interaction mechanisms to learning activities. Secondly, metamodel entities represented by sub-metaclasses of the *Entity* metaclass defined in the *Commons* package can be extended by adding *Property* metaclass instances.

4 THE NEW ABSTRACTION TO MODEL LANGUAGE LEARNING ACTIVITIES

This section explains how to use the "Fill-in the Gaps" activity model as a universal abstraction to represent different language learning activities. Figure 3 depicts the *Activity* package which contains the "Fill-in the Gaps" Activity metamodel used to model language learning activities.

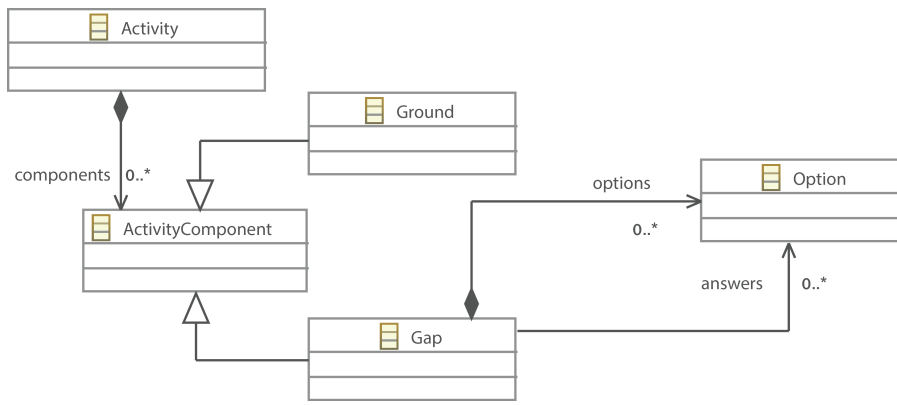


Figure 3. Activity metamodel

As we have mentioned, this model links media resources (i.e. *MediaContent* metaclass instances) to *Slide* sub-metaclass instances using *Concept* metaclass instances as the glue between these two aspects.

4.1 Fill-in the Gaps Learning Activity

The fill-in the gaps learning activity is a straightforward application of the fill-in the gaps abstraction. Figure 4 depicts the model and presentation of a fill-in the

gaps activity in the Busuu methodology. While the right side of the figure shows the actual user interface for the activity; the left side of the figure shows the model that represents the activity.

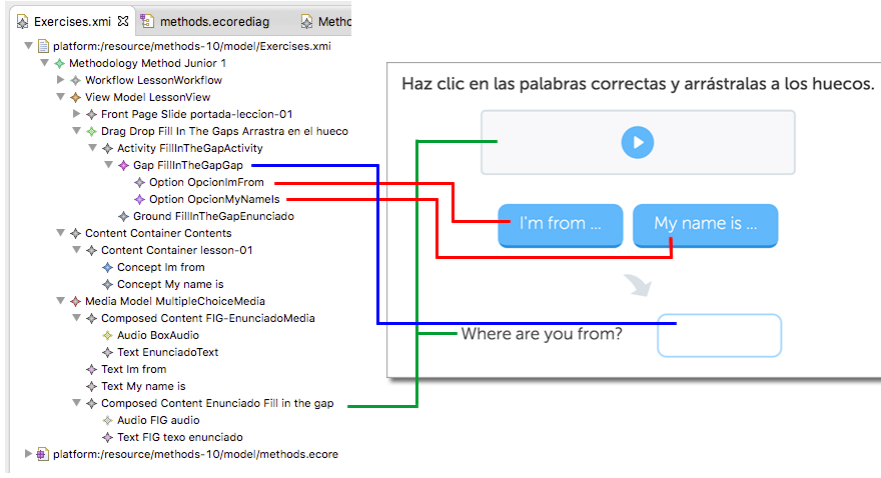


Figure 4. Fill-in the gaps activity in Busuu

4.2 Word Ordering Learning Activity

The word ordering learning activity asks students to order a set of words to compose a meaningful sentence. Figure 5 shows the Busuu methodology version of this activity. The presentation model of this activity is defined by an instance of the *ComposePhrase* metaclass. The information to customize the activity is defined by an instance of the *Activity* metaclass.

In this case, the activity defines 4 gaps (represented by instances of the *Gap* metaclass) composed by 4 instances of 4 options (represented by instances of the Option metaclass) for each gap. These options are linked to the same text strings (represented by instances of the *Text* metaclass) to enable users to choose one string (i.e. *are*, *Where*, *from?*, *you*) in any position of the sentence. The order of the gaps defines the order of the words in the sentence. And each gap defines only one option as the correct answer to set only one word ordering as correct. Finally, the *Ground* metaclass instance defines the learning activity statement. *Ground* metaclass instances can also be used as part of the sentence to include fixed words that cannot be modified by the user (e.g. punctuation marks, words, etc.).

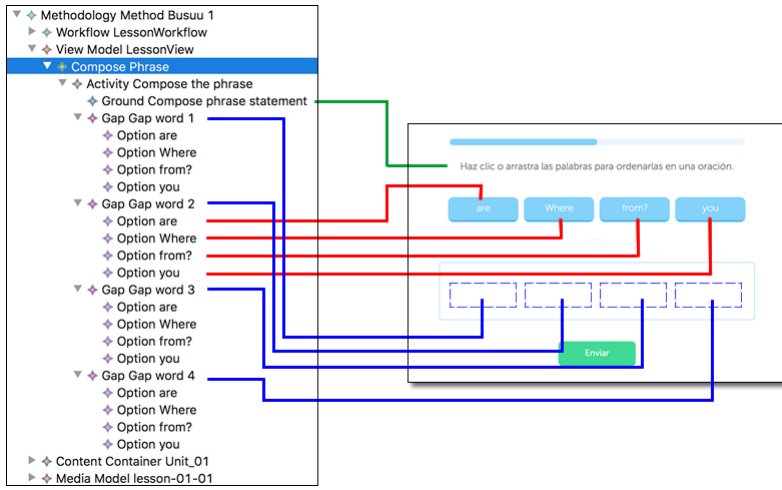


Figure 5. Word ordering learning activity in Busuu

4.3 Match-Up Learning Activity

This section describes how to model a match-up learning activity in the Busuu methodology using the fill-in the gaps abstraction. Figure 6 depicts the learning activity user interface which consists in locating and matching up the 3 elements on the left with the corresponding 3 elements on the right.

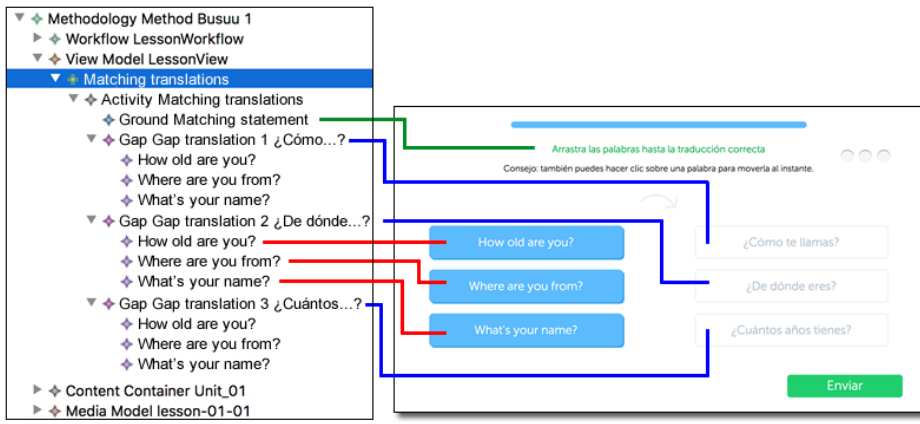


Figure 6. Match-up learning activity in Busuu

In this case, the presentation of the activity is defined by an instance of the *Match* metaclass. The activity information is modelled as a fill-in the gaps exercise

including 3 gaps representing the 3 elements depicted on the right side of the figure (i.e. *¿Cómo te llamas?*, *¿De dónde eres?*, *¿Cuántos años tienes?*). Each of gap presents the same 3 options to be linked to the 3 elements on the left side (*How old are you?*, *Where are you from?*, *What's your name?*) where only one of these options is defined as the correct answer. As in the previous example, the options (represented by instances of the *Option* metaclass) are linked to instances to the *Text* metaclass.

4.4 Locution to Text, Multiple Choice and Translate Phrase Learning Activities in Duolingo

Figures 7, 8 and 9 depict 3 learning activities that illustrate the similarities of the activity models in different learning activities in the Duolingo methodology. All the activity models of these learning activities define only one gap (instance of the *Gap* metaclass) including several options (instances of the *Option* metaclass) where at least one of them is set as the correct one.

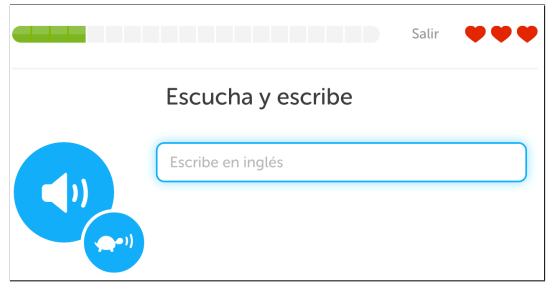


Figure 7. Locution to text learning activity in Duolingo using one gap

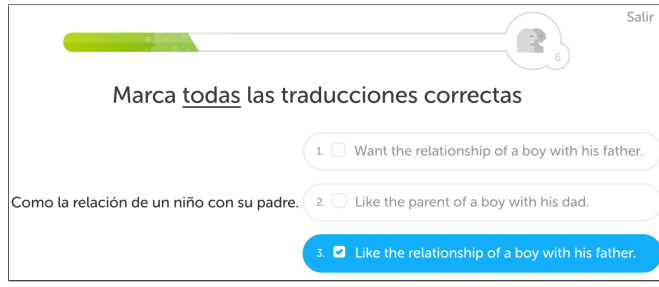


Figure 8. Multiple choice learning activity in Duolingo using one gap

These models also define instances of the *Ground* metaclass to represent the statement of the activity (e.g. *Escucha y escribe*, *Marca todas las respuestas correctas*, *Traduce este texto*). The *Ground* and *Option* metaclass instances are linked

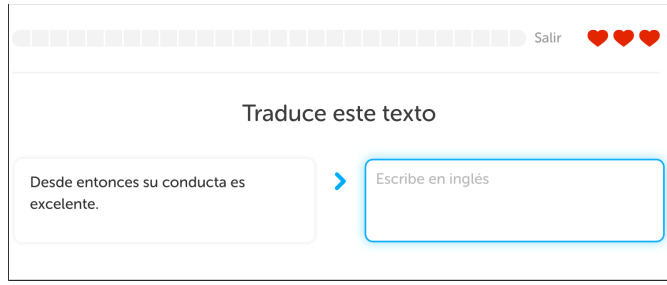


Figure 9. Translate phrase learning activity in Duolingo using one gap

to media resources represented by instances of the *ComposedContent* or *Text* metaclasses. The *ComposedComponent* metaclass instances enables developers to provide different types of media resources (e.g. *Text* and *Audio* metaclass instances).

Finally, the locution to text, multiple choice and translate phrase learning activities are represented by instances of the *LocutionToText*, *MultipleChoice* and *TranslatePhrase* metaclasses, respectively.

4.5 Multiple Choice Learning Activity in Babbel, Busuu and Duolingo

The modelling process can be analogously applied to model the information or content of the same learning activity for different methodologies.

Figure 10, Figure 11 and Figure 12 depict 3 examples of multiple choice activities in 3 different learning methodologies. (i.e. Babbel, Busuu and Duolingo). All these examples represent the activity exposed in Section 3 and depicted in Figure 2.

Figure 10. Multiple choice learning activity in Babbel

All these 3 examples define one gap with 3 options. However, each exercise defines different kinds of media resources to represent the activity statement and options. For instance, while Figure 10 and Figure 12 depict Babbel and Doulingo

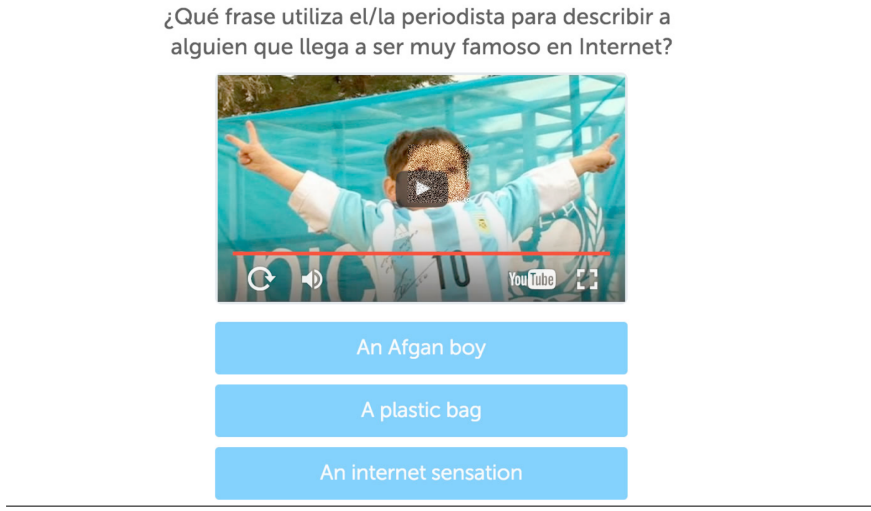


Figure 11. Multiple choice learning activity in Busuu

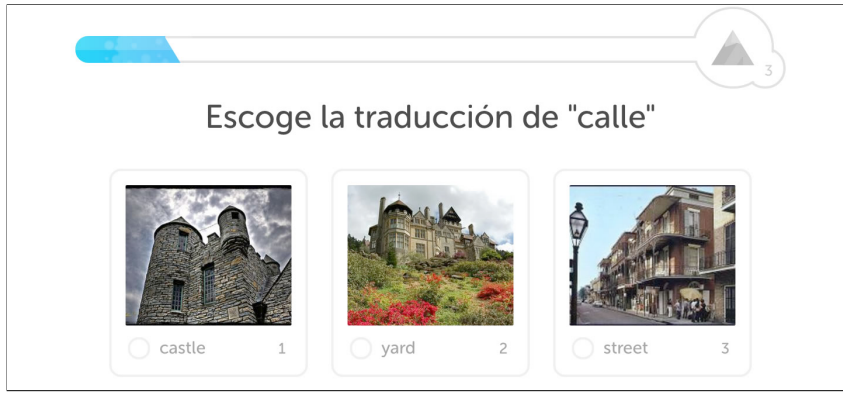


Figure 12. Multiple choice learning activity in Duolingo

version of a multiple choice activity using 3 pictures to represent gap options; the Busuu version of the activity depicted in Figure 11 uses a video to represent the activity statement.

Therefore, the idea of employing the fill-in the gaps activity as an abstraction for all kinds of learning activities simplifies the modelling process since all activities are modelled in the same way which favours the reuse of models.

5 EVALUATING THE PROPOSAL

The main goal of the evaluation section is to evaluate the quality of the code obtained after applying the MDA approach proposed in this article (product quality evaluation). The quality of the code is evaluated by comparing the code obtained using the MDA approach against the code generated by an expert (called traditional approach). The comparison is performed using a set of well-known quality factors. Before performing the product quality evaluation, we have to prepare the artefacts using both approaches. The next subsection describes the preparation process.

5.1 Preparation Process

This section describes the process followed to generate the artefacts (learning activities) that will be compared in the next subsection.

There are two different mechanisms to obtain the learning activities: the proposed approach and the traditional approach.

5.1.1 Objectives of the Preparation Process

The objective of the preparation is to develop a set of learning activities considering both the proposed and traditional approaches.

5.1.2 Participants of the Preparation Process

This evaluation is carried out by two participants of different profiles.

The first participant (male, age 23, university graduated) is expert in HTML, CSS and JavaScript technologies and develops learning activities following a traditional approach (HTML expert).

The second participant (male, age 41, Ph.D. student) is an expert in the Eclipse Modeling Framework (EMF) technology designed for model-driven development and develops learning activities following the proposed approach (MDA expert).

5.1.3 Computing Environment

Each participant performed the preparation test in the ISE Research Group Interaction laboratory located in the Albacete Research Institute of Informatics (I3A) building in Albacete, Spain. This location is equipped with computers and multimedia equipment (e.g. video cameras, microphones, and so on) that makes it suitable for performing HCI (Human-Computer Interaction) interaction evaluations.

Both development processes were performed using the same computing equipment. The hardware consists of a MAC Book Pro 13" Retina laptop computer with 8 GB RAM and 256 GB SSD. This computer runs the High Sierra iOS, SublimeText (ver. 3.0), and the Eclipse Modeling Tools NEON 3 IDE including ATL (ver. 3.6.6),

ACCELEO (ver. 3.7.0) and the proposed approach reflexive model editor (ver. 1.0.0) and transformation (ver. 1.0.0) plugins.

5.1.4 Tasks of the Preparation Process

Participants receive the same list of requirements proposing the development of four Lexiway learning activities to review student language vocabulary. These activities look like the learning activity depicted on the left side of Figure 2 that presents four images and plays the audio file of a word when it starts. Learners should click on an image corresponding to the word that was played. When an image is clicked, they receive the result of the matching in terms of negative or positive reinforcement. If the clicked image does not correspond to the audio played, the file is played again and the user is asked to click on an image again until they choose the correct image.

Each participant followed a different path to develop the learning activities. The MDA Expert had to define a model for each activity, validate the model and generate the code. The HTML Expert had to use his favourite HTML editor, locate the resources, write the code, and test the solution.

However, it is possible to identify some general tasks, no matter the tools used to get them. The development of each learning activity represents a task. Each task is divided into the sub-tasks, which are defined in Table 1.

5.1.5 Review and Testing

Both participants knew that the learning activities they developed, would be analyzed by a group of experts. Therefore, they spent some time to check the product obtained. This process was carried out in different ways by both participants, since the tools used in each case were different. In the case of the HTML Expert, this process includes activities like refactoring, refining, testing, and so on. In the case of the MDA Expert, this process consists on model review, validate the OCL restrictions, operate the transformations and review the results.

5.2 Product Quality

This evaluation analyses the quality of language learning activity source codes generated with the proposed and traditional approaches, obtained in the previous section. To carry out this task, we propose an heuristic evaluation where a set of 5 experts evaluates 4 software attributes related to software quality characteristics.

5.2.1 Objective

This heuristic evaluation compares the source code quality of the language learning application depicted on the left side of Figure 2 generated with the traditional and

T	Description
1	Define the graphic design of the user interface for the learning activity presented on the left side of Figure 2 using the set of images defined in a specific folder. 1. Slide option images (e.g. clock, spot, fox and box), 2. Common images (e.g. headphones, correct, wrong).
2	Define the audio modality of the user interface for the learning activity presented on the left side of Figure 2 using audio files defined in a specific folder. 1. Possible slide statement locutions (e.g. clock, spot, fox and box), 2. Common sounds (e.g. correct and wrong answers).
3	Define the learning activity statement linking the click event on the headphones image to one of the possible locutions for the activity.
4	Define the learning activity answer linking answer images to option images according to the selected statement locution for the activity (e.g. the correct answer image for the learning activity depicted on the left side of Figure 2 is box and the rest of options are linked to the wrong image).
5	Define the learning activity answer linking answer sounds to option images according to the selected statement locution for the activity (e.g. the correct answer sound for the learning activity depicted on the left side of Figure 2 is box and the rest of the options are linked to the wrong sound).
6	Define the learning activity behaviour when learners click on the wrong answer (i.e. play the statement locution again).

Table 1. Common tasks performed by the participants

proposed development processes in terms of software *correctness*, *robustness*, *extensibility* and *reusability*; where software *correctness* refers to the Functional Correctness sub-characteristic of the Functional Suitability characteristic defined in the Product quality model of the ISO 25010:2011(E) standard [15], software *robustness* refers to the Fault tolerance and Recoverability sub-characteristics of the Reliability characteristic defined in the Product quality model of the ISO 25010:2011(E) standard [15], software *extensibility* refers to the Modularity and Modifiability sub-characteristics of the Maintainability characteristic defined in the Product quality model of the ISO 25010:2011(E) standard [15], and software *reusability* refers to the Reusability sub-characteristic of the Maintainability defined in the Product characteristic quality model of the ISO 25010:2011(E) standard [15].

5.2.2 Participants

This evaluation is performed with 5 experts, whose profiles are exposed in Table 2. Participant profiles include information such as gender, age, and experience in HTML, JavasScript, Language Learning Applications and Software Quality.

Participant	Gender	Age	Experience			
			HTML	JavaScript	L. L. Apps.	Soft. Quality
1	M	38	5	5	3	5
2	F	35	5	4	5	4
3	M	39	5	5	3	4
4	M	45	4	5	5	4
5	F	48	5	5	4	5

Table 2. Participant profiles

5.2.3 Computing Environment

This evaluation was carried out in the ISE Research Group interaction laboratory located in the Albacete Research Institute of Informatics (I3A) building in Albacete, Spain. This location is equipped with computers and multimedia equipment (e.g. video cameras, microphones, and so on) that makes it suitable for performing HCI interaction evaluations.

The evaluation was performed on a Dell XPS 702x laptop computer running Microsoft Windows 10. The Internet browser used to run both implementations is Chrome version 64.

5.2.4 Metrics

The metrics to evaluate software *correctness*, *robustness*, *extensibility* and *reusability* are scored from 1 to 5 according to experts' criteria where 1 and 5 represents the lowest and highest scores of the software product for a specific attribute, respectively.

5.2.5 Procedure

The evaluation procedure starts when participants receive the source codes of the language learning activity (both traditional and proposed) presented on the left side of Figure 2, and a form to score these source codes in terms of selected software attributes as well as an extra section to justify the software product scoring. It is worth to highlight that how source codes were generated is unknown to participants.

The W3C Validator [17] and the JSHint[18] tools are available to participants to help them to score the software product.

5.2.6 Results

The overall results of the comparison between the traditional approach and the proposed approach are exposed in Table 3.

[17] `https://validator.w3.org/`

[18] `http://jshint.com/about/`

Part.	Approach	Correctness	Robustness	Extensibility	Reusability	Average
1	Traditional	3	4	2	3	**3**
	Proposed	5	5	5	5	**5**
2	Traditional	3	3	3	3	**3**
	Proposed	5	5	5	5	**5**
3	Traditional	5	2	1	2	**2.5**
	Proposed	5	5	4	4	**4.5**
4	Traditional	5	5	3	5	**4.5**
	Proposed	4	5	5	5	**4.75**
5	Traditional	5	4	3	4	**4**
	Proposed	5	4	4	5	**4.5**
Average	Traditional	**4.2**	**3.6**	**2.4**	**3.4**	**3.4**
	Proposed	**4.8**	**4.8**	**4.6**	**4.8**	**4.75**
Std. Dev.	Traditional	1.09	1.14	0.89	1.14	
	Proposed	0.44	0.45	0,55	0.45	
Max	Traditional	5	5	3	5	
	Proposed	5	5	5	5	
Min	Traditional	3	2	1	2	
	Proposed	4	4	4	4	

Table 3. Heuristic evaluation results

According to experts, the proposed approach based on models overcomes the traditional approach because while the proposed approach scores 4.75 out of 5 in the overall scoring, the traditional approach only obtained 3.4 out of 5. Moreover, the scores of the proposed approach are higher than 4 out of 5 in all evaluated software attributes.

The standard deviation on the proposed approach also delivers scores on all attributes are close to the average score which is above 4.6 out of 5 showing an homogeneous consensus on the experts.

The proposed approach does not only overcomes the traditional approach in the overall results; it also overcomes all evaluated software attributes.

The score is even more significant when evaluating the *extensibility* of the software product (it almost doubles the score of the traditional approach). Participants also highlighted the quality of the code using the traditional approach is more difficult to extend than the code generated using the proposed approach avoiding the great impact on code modifications.

Although the proposed approach obtains a higher score than the traditional approach in terms of *correctness*; the average score in this subject for the traditional approach is really good obtaining a difference less than 0.6 with a standard deviation of 1.09 with respect to the proposed approach.

Comparing both approaches in terms of *robustness*, participants state that conditional structures in the code generated using the traditional approach are not as well-structured as in the proposed approach.

About the code *reusability*, they mention that the code generated using the proposed approach organizes multimedia resources (i.e. media files, JavaScript and Cascade Style Sheets) more efficiently than the code generated using a traditional approach, because the proposed approach groups common resources to all activities encouraging their reuse. Moreover, all participants highlight the code structure generated using the proposed approach because it defines parametrized functions encouraging their reuse too.

6 CONCLUSIONS

This article proposes the fill-in the gaps abstraction to model language learning activities in the context of an MDA to develop language learning applications.

It shows, by means of a series of examples, how the fill-in the gaps abstraction is used to represent different learning activities in different language learning methodologies. These examples show how the modelling process is simplified since activity models can be easily reused to:

1. customize the interaction mechanism of the learning activity,
2. adapt the application look to users' needs (i.e. colour-blind people),
3. model the same activity for different learning methodologies.

Traditional approaches force developers to build applications for different platforms (e.g. Web, iOS, Android, Windows, etc.) leading to different development branches which are prone to errors, difficult to maintain and test (e.g., changes and fixes on the application domain model should be addressed in all platforms).

Model-driven architectures decouples application functionality from technology which enable developers to create Platform Independent Models (PIMs) and Platform Specific Models (PSMs) to derive application source code semi-automatically using model transformations.

One of the main features of employing MDAs is the design time interoperability. This feature captures different application concerns in independent models which are integrated at the last stage of development (just before the generation of the application source code).

This proposal defines 5 concerns regarding the development of activities for language learning methodologies (i.e. learning methodology contents, learning activity workflows, learning media resources, learning activity interaction mechanisms, and learning activity model). The main advantage behind this feature is the capability of modifying the model with minimum impact on the others. For example, the representation of a concept (i.e. an image) can be changed by modifying only the media resource model without affecting the other models (i.e. content, workflow, interaction mechanism or activity model).

We performed a users' evaluation to analyse the product quality of our proposal. The product quality analyses the product correctness, robustness, extensibility, and reusability. These results conclude that the code generated using the proposed

approach overcomes the code generated using a traditional approach in the selected metrics.

The future work for this project includes the development of graphic modelling editor using the GMF framework to create, edit and verify learning methodology models generated with our metamodel. We are working on extensions to the proposed metamodel that allow improving the validation of the models through the definition of customized OCL expressions that are loaded dynamically (using the Complete OCL plugin [19]). These extensions are defined based on the subclassification of meta-classes, which allow introducing new features in a flexible and robust way.

Moreover, we are addressing the process of model-to-model (M2M) and model-to-text (M2T) transformations required for source code automatic generation (i.e. HTML and JavaScript) of language learning applications. To carry out these transformations, the ATLAS Transformation Language[20] (ATL) and ACCELEO[21] transformation languages are used respectively.

This transformation process requires 3 M2M transformations: (1) a M2M transformation to generate the Activity Model and Workflow Model instances that are part of the PIM layer of the model architecture using a Content Model instance that is part of the CIM layer; (2) another M2M transformation at PIM layer to generate Presentation Model and Media Model instances from Activity Model and Workflow Model instances; finally, (3) the last M2M transformation generates the TagML [25] PSM layer model from Activity Model and Workflow Model PIM layer instances.

In addition, the transformation process also involves 2 M2T transformations. While the first one generates HTML source code from TagML [25] PSM model to define the application UI structure; the second one generates JavaScript source code from Workflow Model PIM layer instance to define the application UI behavior. Thus, the whole process generates the Implementation Specific Model (ISM) of a fully functional application.

Finally, we are also exploring the adaptation of this model architecture to generate a wide variety of interactive multimedia applications including gamification features.

Acknowledgements

This work has been partially supported by the national project granted by the Ministry of Science, Innovation and Universities (Spain) with reference RTI2018-099942-B-I00 and by the project TecnoCRA (ref: SBPLY/17/180501/000495) granted by

[19] `https://marketplace.eclipse.org/content/eclipse-ocl`
[20] `https://eclipse.org/atl/`
[21] `https://www.eclipse.org/acceleo/`

the regional government (JCCM) and the European Regional Development Funds (FEDER).

REFERENCES

[1] Bárcena, E.—Read, T.—Underwood, J. et al.: State of the Art of Language Learning Design Using Mobile Technology: Sample Apps and Some Critical Reflection. In: Helm, F., Bradley, L., Guarda, M., Thouësny, S. (Eds.): Critical CALL – Proceedings of the 2015 EUROCALL Conference. Padova, Italy, 2015, pp. 36–43, doi: 10.14705/rpnet.2015.000307.

[2] Bizonova, Z.: Model Driven E-Learning Platform Integration. In: Maillet, K., Klobucar, T., Gillet, D., Klamma, R. (Eds.): Proceedings of the EC-TEL 2007 PROLEARN Doctoral Consortium. CEUR Workshop Proceedings, Vol. 288, 2007.

[3] Blumschein, P.—Hung, W.—Jonassen, D.—Strobel, J. (Eds.): Model-Based Approaches to Learning: Using Systems Models and Simulations to Improve Understanding and Problem Solving in Complex Domains. Brill, Leiden, The Netherlands, Modeling and Simulation for Learning and Instruction, Vol. 4, 2009, doi: 10.1163/9789087907112.

[4] Brambilla, M.—Fraternali, P.: Interaction Flow Modeling Language: Model-Driven UI Engineering of Web and Mobile Apps with IFML. 1st Edition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2014.

[5] Calvary, G.—Coutaz, J.—Thevenin, D.—Limbourg, Q.—Bouillon, L.—Vanderdonckt, J.: A Unifying Reference Framework for Multi-Target User Interfaces. Interacting with Computers, Vol. 15, 2003, No. 3, pp. 289–308, doi: 10.1016/s0953-5438(03)00010-9.

[6] Ceri, S.—Fraternali, P.—Bongio, A.: Web Modeling Language (WebML): A Modeling Language for Designing Web Sites. Computer Networks, Vol. 33, 2000, No. 1-6, pp. 137–157. doi: 10.1016/S1389-1286(00)00040-2.

[7] Conallen, J.: Building Web Applications with UML. 2nd Edition. Addison Wesley, Reading, Massachusetts, October 2002.

[8] Conn, S.—Forrester, L.: Model Driven Architecture: A Research Review for Information Systems Educators Teaching Software Development. Information Systems Education Journal, Vol. 4, 2006, No. 43, pp. 3–11.

[9] Dodero, J. M.—García-Peñalvo, F.-J.—Gonzãlez, C.—Moreno-Ger, P.—Redondo, M.-A.—Sarasa-Cabezuelo, A.—Sierra, J.-L.: Development of E-Learning Solutions: Different Approaches, a Common Mission. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, Vol. 9, 2014, No. 2, pp. 72–80, doi: 10.1109/RITA.2014.2317532.

[10] Fardoun, H.—Montero, F.—López Jaquero, V.: eLearniXML: Towards a Model-Based Approach for the Development of E-Learning Systems Considering Quality. Advances in Engineering Software, Vol. 40, 2009, No. 12, pp. 1297–1305, doi: 10.1016/j.advengsoft.2009.01.019.

[11] Feuerstack, S.—Pizzolato, E. B.: Engineering Device-Spanning, Multimodal Web Applications Using a Model-Based Design Approach. In: Bressan, G., Sil-

veira, R. M., Munson, E. V., Santanchà, A., da Graça Campos Pimentel, M. (Eds.): Proceedings of the 18th Brazilian Symposium on Multimedia and the Web (WebMedia'12). Association for Computing Machinery, 2012, pp. 29–38, doi: 10.1145/2382636.2382646.

[12] Gamma, E.—Helm, R.—Johnson, R.—Vlissides, J.: Design Patterns. Elements of Reusable Object-Oriented Software. Addison-Wesley, Massachusetts, 1995.

[13] García-Peñalvo, F. J.: Advances in E-Learning: Experiences and Methodologies. Information Science Reference, 2008, doi: 10.4018/978-1-59904-756-0.

[14] Isakowitz, T.—Stohr, E. A.—Balasubramanian, P.: RMM: A Methodology for Structured Hypermedia Design. Communications of the ACM, Vol. 38, 1995, No. 8, pp. 34–44, doi: 10.1145/208344.208346.

[15] ISO. ISO/IEC 25010. Systems and Software Engineering – Systems and Software Engineering Quality Requirements and Evaluation (SQuaRE) – Systems and Software Quality Models, 2006.

[16] Kleppe, A.—Warmer, J.—Bast, W.: MDA Explained: The Model Driven Architecture: Practice and Promise. Addison-Wesley Professional, 2003.

[17] Koch, N.—Kraus, A.: Towards a Common Metamodel for the Development of Web Applications. In: Lovelle, J. M. C., Rodríguez, B. M. G., Gayo, J. E. L., del Puerto Paule Ruiz, M., Aguilar, L. J. (Eds.): Web Engineering (ICWE 2003). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 2722, 2003, pp. 497–506, doi: 10.1007/3-540-45068-8_92.

[18] Lahiani, N.—Bennouar, D.: A Model Driven Approach to Derive E-Learning Applications in Software Product Line. Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication (IPAC'15), ACM, 2015, Art. No. 78, doi: 10.1145/2816839.2816850.

[19] Lanzilotti, R.—Ardito, C.—Costabile, M. F.—De Angeli, A.: eLSE Methodology: A Systematic Approach to the E-Learning Systems Evaluation. Educational Technology and Society, Vol. 9, 2006, No. 4, pp. 42–53.

[20] Mellor, S. J.—Scott, K.—Uhl, A.—Weise, D.: MDA Distilled: Principles of Model-Driven Architecture. Addison Wesley, 2004.

[21] Dehbi, R.—Talea, M.—Tragha, A.: A Model Driven Methodology Approach for E-Learning Platform Development. International Journal of Information and Education Technology, Vol. 3, 2013, No. 1, pp. 10–15, doi: 10.7763/IJIET.2013.V3.225.

[22] Sebastián, G.—Tesoriero, R.—Gallud, J. A.: Modeling Language-Learning Applications. IEEE Latin America Transactions, Vol. 15, 2017, No. 9, pp. 1771–1776, doi: 10.1109/TLA.2017.8015084.

[23] Sebastián, G.—Tesoriero, R.—Gallud, J. A.: Model-Based Approach to Develop Learning Exercises in Language-Learning Applications. IET Software, Vol. 12, 2018, No. 3, pp. 206–214, doi: 10.1049/iet-sen.2017.0085.

[24] Tang, S.—Hanneghan, M.: A Model-Driven Framework to Support Development of Serious Games for Game-Based Learning. 2010 Developments in E-Systems Engineering, 2010, pp. 95–100, doi: 10.1109/DeSE.2010.23.

[25] Tesoriero, R.—Sebastián, G.—Gallud, J. A.: TagML – An Implementation Specific Model to Generate Tag-Based Documents. Electronics, Vol. 9, 2020, No. 7, Art. No. 1097, doi: 10.3390/electronics9071097.

[26] Tian, Y.—Yang, H.—Landy, L.: MDA-Based Development of Music-Learning System. In: Zhang, S., Li, D. (Eds.): Proceedings of the 14th Chinese Automation and Computing Society Conference, 2008, pp. 97–102.

Gabriel Sebastián received his Ph.D. and M.Sc. degrees from the University of Castilla-La Mancha (UCLM) and B.Sc. degree from Polytechnic University of Valencia (UPV) – all the three degrees in computer science. His main research interests are multimedia, human-computer interaction and software engineering. He was involved in the development of many projects related to distributed user interfaces and model-driven development of user interfaces focused on the web as the deployment platform. He published more than 25 research articles and book chapters in journals and international congresses. Currently, he works as Project Manager in the Interactive Systems Engineering Research Group of the Computing System Department (Faculty of Computing Science Engineering) in UCLM, and he works as Researcher in the Albacete Research Institute of Informatics (I3A) in Albacete, Spain.

Ricardo Tesoriero received his Ph.D. and M.Sc. degrees from the University of Castilla-La Mancha (UCLM), Spain and a B.Sc. degree from National University of La Plata (UNLP), Argentina – all the three degrees in computer science. His main research interests are model-driven development of user interfaces focused on the web as deployment platform and human-computer interaction on ubiquitous computing environments. He published more than 70 research articles and book chapters in journals and international congresses. He performed a post-doctoral stay in the Université Catholique de Louvain (UCL) in Louvain-la-Neuve, Belgium where he performed research activities on model-driven development of user interfaces. He was committee member in several scientific conferences and workshops, including DUI (Distributed User Interfaces), INTERACCCION, ISEC, IADIS/WWW (La Web), etc. He is Associate Professor of the Computing System Department at the Faculty of Computer Science Engineering of the UCLM teaching web engineering and services, and human-computer interaction subjects since 2008.

Jose A. Gallud received his Ph.D. degree from the University of Murcia and the M.Sc. and B.Sc. degrees from the Polytechnic University of Valencia – all the three degrees in computer science. His main research of interest focuses on human-computer interaction, development of interactive systems and distributed user interfaces. He has published widely in these areas. He has been Guest Editor for several international journals, such as JSS (The International Journal of Software and Systems), IJHCS (International Journal of Human Computer Studies), JUCS (Journal of Universal Computer Sciences). He has some books and chapters in the field of human-computer interaction. He is member of different national and international societies (ACM and AIPO). Currently, he works as Professor at the University of Castilla-La Mancha.

Computing and Informatics, Vol. 40, 2021, 277–297, doi: 10.31577/cai_2021_2_277

ISOLATED WORD RECOGNITION BY RECURSIVE HMM PARAMETER ESTIMATION ALGORITHM

Jūratė Vaičiulytė

Faculty of Electronics and Informatics
Vilniaus Kolegija – University of Applied Sciences
Vilnius, Lithuania
e-mail: j.vaiciulyte@eif.viko.lt, jurate.vaiciulyte@mif.vu.lt

Leonidas Sakalauskas

Vilnius University Šiauliai Academy
Šiauliai, Lithuania
e-mail: leonidas.sakalauskas@mif.vu.lt

Abstract. Automatic speech recognition (ASR) technologies enable humans to communicate with computers. Isolated word recognition (IWR) is an important part of many known ASR systems. Minimizing the word error rate in cases of incremental learning is a unique challenge for developing an on-line ASR system. This paper focuses on on-line IWR using a recursive hidden Markov model (HMM) multivariate parameter estimation algorithm. The maximum likelihood method was used to estimate the unknown parameters of the model, and an algorithm for the adapted recursive EM algorithm for HMMs parameter estimation was derived. The resulting recursive EM algorithm is unique among its counterparts because of state transition probabilities calculation. It obtains more accurate parameter estimates compared to other algorithms of this type. In our experiment, the algorithm was implemented and adapted to several datasets for IWR. Thus, the recognition rate and algorithm convergence results are discussed in this work.

Keywords: Hidden Markov models, likelihood method, on-line algorithm, recursive EM algorithm, isolated word recognition

Mathematics Subject Classification 2010: 68T10, 62H12, 62H30, 68T05

1 INTRODUCTION

Automatic speech recognition (ASR) is a pattern recognition task with the objective of classifying input data into classes based on certain features. It is a complex, multistep task in computer-aided speech processing and recognition. In other words, speech recognition can be defined as speech transcription using a computer [1]. ASR can be applied to numerous practical areas, such as controlling software [2, 3], dialing numbers [4], internet searches [5, 6], etc. It is a difficult problem, so several recognition techniques have been proposed, including linear-time-scaled word-template matching [7], hidden Markov models (HMMs) [8, 9, 10], deep neural networks [11], etc. HMM is widely applied to speech recognition systems because it provides accurate speech modeling.

Traditional speech modeling and learning methods such as deep neural networks, linear-time-scaled word-template matching and HMM require static training dataset to accurately learn speech model parameters. However, the complexity of these learning methods is at least of the second order, since the required number of calculations at each learning iteration depends on the size of the dataset.

The quality and quantity of speech training and testing material play an important role in correctly representing modeled language and its recognition rate. In contrast to the traditional learning methods, recursive learning could be the solution in cases when the number of speech samples for training is too small to be practically used in recognition systems. Recursive learning methods could provide practical real-time collection of speech data.

Recently, much attention has been paid to recursive model parameter learning methods [12, 13, 14, 15, 16, 17, 18, 19, 20]. However, there has not been enough exploration of recursive learning algorithms applied to real-time speech recognition systems which are based on HMM. Most on-line speech recognition systems use a static trained model, which is then used to identify words in the speech signal. If new speech data is provided for training, these systems cannot apply a new dataset to the speech model without being retrained with the aggregated data. This disadvantage would be avoided if training and model parameter adaptation were performed incrementally while processing and recognizing spoken words occurs. Such algorithms would lead to the creation of a speech recognition system that constantly adapts to new speech signals without decreasing recognition accuracy.

In this work, isolated word recognition (IWR), which is a subclass of ASR, has been performed using a recursive EM algorithm for HMM parameter estimation. In an IWR system, the input data are considered as words that are processed individually, and previously uttered words do not affect the recognition. The input data is a raw speech file that is converted into an acoustic feature vector and is processed over time. A HMM with a fixed number of states is used to model each word. The recursive EM algorithm for HMM parameter estimation is presented in this work. It consists of two main parts – model training, and recognition with re-estimation. In the training part, the feature vectors are extracted from input files, and HMM parameter estimation is performed for each word. In the recognition and

re-estimation part, each input is recognized and the model parameters of the word are updated according to the recognized word, which allows the algorithm to continuously estimate model parameters and perform recognition at the same time. The experimental results for the created algorithm are discussed in this work as well.

2 RELATED WORK

HMM parameter estimation algorithms can be classified in two main categories: batch [8] and recursive (on-line) [21]. Batch learning is a standard procedure for learning model parameters, and is known to be very robust. Batch learning algorithms process blocks of observations after they are stored in the computer's memory and they execute as many iterations on the training set as necessary for tuning such parameters. Generally, batch algorithms apply an offline Baum-Welch (EM) algorithm, which locally maximizes the likelihood objective function and applies an HMM forward-backward procedure [8, 22]. In cases of sequential data processing, the complexity of the batch EM algorithm is quadratic. This approach means that the number of calculations required to obtain the parameter estimates is proportional to the observation set size.

Recursive HMM parameter estimation algorithms calculate model parameter estimates incrementally with each new observation. Calculations are performed sequentially in time so that the algorithm does not need to store all observations in the computer's memory.

The Baum-Welch algorithm is successfully implemented in numerous offline speech recognition systems which are based on HMM parameter estimation. The popularity of Baum-Welch algorithm urged others to develop on-line (recursive) EM algorithms for real-time HMM parameter estimation. The recursive EM contains a maximum likelihood estimator (MLE) which is iteratively maximized.

The main difficulty in implementing recursive expectation-maximization algorithms has been calculating the required data statistics without the backwards recursion of the HMM Forward-Backward procedure. As a result, in [23], the authors proposed the on-line HMM parameter estimation algorithm with implemented forward recursion (only forward recursion can be efficiently implemented in on-line mode). However, as the backwards recursion is hard to apply in on-line mode, it was ignored. The authors of [24] presented an on-line HMM parameter estimation algorithm with an adapted Forward-Backward procedure and applied it to background modeling.

In [25], the authors proposed recursive HMM parameter estimator with online finite memory approximation to the forward-backward procedure. Also, various recursive MLE method modifications based on different optimization techniques can be identified in the literature: a numerical smoothing method (which replaces the forward-backward procedure) [26, 27], fixed-interval smoothing with an exponential forgetting factor [12], and HMM parameter estimation based on stochastic gradient methods [28, 29].

Whilst all of these methods closely resemble the offline Baum-Welch algorithm, their convergence properties are poorly understood. And it is difficult to find experiments that demonstrate the effectiveness of these algorithms when they are applied to solving real tasks.

3 SPEECH RECOGNITION AND HMMS

An ASR system (see Figure 1) gets a speech signal input, processes it and outputs the text equivalent to the input. ASR usually consists of two stages – primary processing and final processing. Primary processing involves extracting features from the speech signal, and the final processing consists of a speech recognition engine that has an acoustic model, language model and grammar. A grammar contains sets of predefined combinations of words. A language model contains the probabilities of sequences of words. If the system is applied to IWR only, then it does not require a language model and grammar. In this case IWR systems recognize single words separated by silence [10, 30]. Thus, the probabilities of sequence of words or the combination of words do not matter because the system analyse separate words. These systems have "listening/not listening" states through which the user has to wait (usually processing is performed during these pauses). Such systems are useful when the user has to pronounce single words or commands.

If all these parts – acoustic model, language model and grammar – are correct, the engine of speech recognition identifies the most likely match for the inputs that are received and returns the recognized words to the text (the decoding is performed). The selection for proper feature extraction and speech recognition methods has a significant impact on the accuracy of the recognition system.

We will only discuss the acoustic model (leaving out the language model and grammar) because we are applying the recursive EM algorithm to IWR. The task of the acoustic model is to evaluate the probability of the sequence of words. The distribution of the feature vector O is usually modeled on smaller phonetic units, such as phonemes, contextual phonemes or syllables. HMMs are used to model this distribution. The HMM can be pictured as a random process that travels through a set of state S and generates a feature vector O. It is a stochastic Markov process with unknown parameters that are unraveled based on observation [31]. In other words, there are two stochastic processes (see Figure 2). The first one is a Markov chain characterized by states S that are "hidden" and transition probabilities A. The second process produces observations depending on a state-dependent probability distribution B.

HMM is used to classify each feature vector sequence with a specific class, which is given as a sequence of objects (such as letters, words, etc.). A probability distribution over possible sequences of classes is calculated, and the best class sequence is chosen. HMM defines observed events (such as utterances in the input) and hidden events (such as utterance recognition and transcription). Each acoustic unit is modeled with one HMM, which is composed of several states. The HMM of the

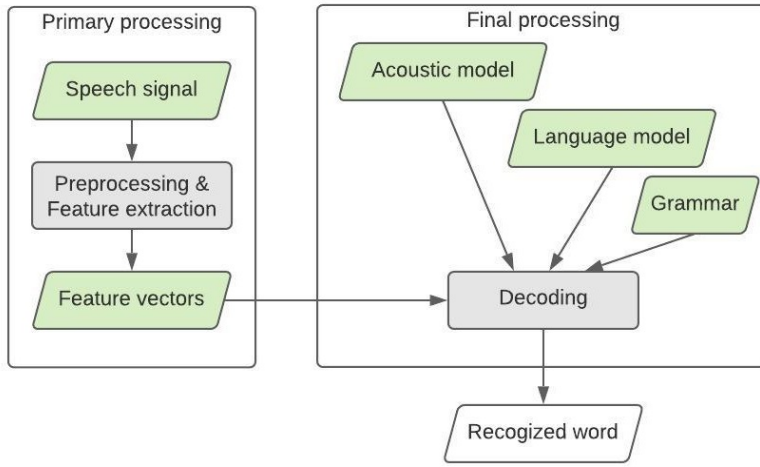


Figure 1. Structure of an ASR system

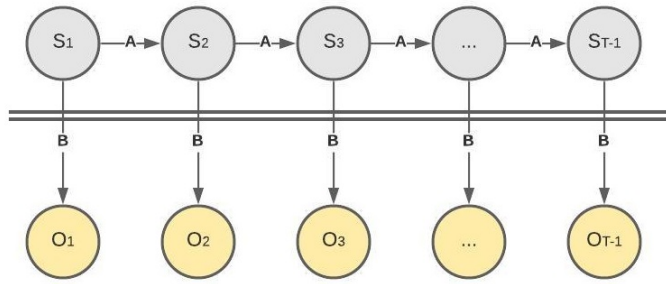


Figure 2. Hidden Markov model

three states (sound start, middle and end) are most commonly used. In the case of a large vocabulary, a static or dynamic network of words composed of many HMMs is formed, and the network is searched for a state sequence S that generates the feature vector O with the highest probability. The Viterbi algorithm is often used to find the best sequence.

Left-to-right HMM (see Figure 3) is mostly used in speech recognition. In this case, the state transition probability matrix has non-zero values for diagonal and neighboring states, and the values of other states in the matrix are set to zero:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & a_{3,2} & a_{3,3} & a_{3,4} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & a_{4,3} & a_{4,4} & a_{4,5} & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & a_{5,4} & a_{5,5} & a_{5,6} & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{6,5} & a_{6,6} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & a_{N-1,N-1} & a_{N-1,N} \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix},$$

N is a number of states and the sum of each matrix row elements is equal to one.

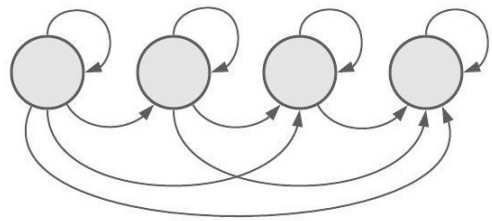


Figure 3. Left-to-right HMM

When constructing HMM, the three main problems that need to be addressed are:

1. Given the model parameters, compute the probability that the HMM generates a particular sequence of observations, solved by the Forward-Backward algorithm;
2. Given a sequence of observations, find the most likely set of model parameters, solved by statistical inference through the Baum-Welch algorithm, which uses the Forward-Backward algorithm;
3. Find the path of hidden states that is most likely to generate a sequence of observations, solved using a posteriori statistical inference in the Viterbi algorithm.

In this paper, we propose the recursive EM algorithm for HMM parameter estimation. This way, the incoming data can be processed recursively and HMM parameters can be updated as soon as new data becomes available.

4 RECURSIVE EM ALGORITHM

The recursive EM algorithm allows us to perform estimation in a sequential way and to re-estimate model parameters in real-time. The main idea of this algorithm is to continuously update model parameters as the observation vectors are given and processed. HMM parameters then are updated according to each new observation without storing the previous observations. The recursive EM algorithm uses the Expectation-Maximization algorithm and maximum likelihood estimator to learn HMM parameters sequentially in real time. We should note that the EM algorithm is often used to learn HMM parameters with the observation sequence and the set of possible states in HMM [32, 33]. The MLE for HMM has proved to be a consistent and asymptotically normal estimator that converges on a stationary point of the sample likelihood.

HMM for the recursive EM algorithm is specified by the following components [34, 35]:

- length (T) of the observation sequence,
- number of states (N) in HMM,
- the state transition probability matrix (A)

$$A = \begin{bmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \dots & \vdots \\ a_{N1} & \dots & a_{NN} \end{bmatrix}, \tag{1}$$

- the initial state distribution vector (π)

$$\pi = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_N \end{bmatrix}^T, \tag{2}$$

- the probability density function B at state s (which expresses the probability of an observation o being generated from state s):

$$B(o, \mu_s, \sigma_s) = \frac{1}{\sqrt{(2\pi)^n |\sigma_s|}} e^{-\frac{1}{2}(o-\mu_s)^T \sigma_s^{-1} (o-\mu_s)}. \tag{3}$$

Observations are defined by a normal distribution with M-dimensional mean μ_s and covariance σ_s, $1 \leq s \leq N$:

$$\mu_s = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_M \end{bmatrix}, \sigma_s = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1M} \\ \vdots & \dots & \vdots \\ \sigma_{M1} & \dots & \sigma_{MM} \end{bmatrix}.$$

Then, the logarithmic likelihood function describes the observation in a state as:

$$l(o, \mu, \sigma) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\sigma|) - \frac{1}{2}(o-\mu)^T\sigma^{-1}(o-\mu). \quad (4)$$

The maximum-likelihood estimation problem is to find

$$\Theta_{ML} = argmax_{\Theta\in\Omega}(l(\Theta))$$

where

- Θ is a vector of parameters. It contains three parameters: π, A, $B(o, \mu_s, \sigma_s)$.
- Ω is a parameter space specifying the set of allowable parameter settings. In the HMM, Ω would enforce the restrictions that all parameter values were ≥ 0:
  - $\sum_{i=1}^{N} \pi_i = 1$;
  - for all $i = 1 \ldots (N-1)$, $\sum_{k=1}^{N} a_{i,k} = 1$;
  - for all $i = 1 \ldots (N-1)$, $\sum_{o\in\Sigma} B(o, \mu_i, \sigma_i) = 1$.

The log-likelihood function (4) in this case gives us a formal measure of how well a particular parameter setting Θ fits the observed sample.

EM algorithm then should consist of these steps:

- Choose the starting values to the parameters to be estimated.
- E-step: Compute the conditional expectations of those functions of the missing data appear in the full log-likelihood.
- M-step: Maximization of the log-likelihood with respect to the set of parameters to be estimated (the missing data are substituted by their conditional expectation).
- Assess convergence (with respect to some criterion) and repeat the E and M-steps until convergence is reached.

Since the full likelihood of each observation sequence is based on the summation of all possible state sequences, each observation is assigned to every state in proportion to the probability of the model being in that state when the vector was observed. Thus, the probability density function (3) parameters of the HMM can be re-estimated through recursive summations of these weighted averages.

Thus, state transition probabilities are defined as

$$\rho_t^i = \frac{1}{t}\sum_{i=1}^{t} \phi_t^i$$

where coefficient

$$\phi_t^i = \frac{e^{-l(o_t, \widehat{\mu_i}, \widehat{\sigma_i}) + \ln(\pi_i)}}{\theta_t},$$

and

$$\theta_t = \sum_{i=1}^{N} e^{-l(o_t,\widehat{\mu}_i,\widehat{\sigma}_i)+\ln(\pi_i)}.$$

Then, the re-estimation of the mean vector and covariance matrix is performed with recursive formulas:

$$\mu_t^i = \mu_{t-1}^i + \frac{(o_t - \mu_{t-1}^i)}{t} \cdot \frac{\phi_t^i}{\rho_t}, \tag{5}$$

$$\sigma_t^i = \left(\frac{\rho_{t-1}^i \cdot (t-1)}{\rho_t^i \cdot t}\right)\left(\sigma_{t-1}^i + \frac{(o_t - \mu_{t-1}^i)(o_t - \mu_{t-1}^i)^T}{t} \cdot \frac{\phi_t^i}{\rho_t}\right), \tag{6}$$

$$\rho_t^i = \rho_{t-1}^i + \frac{1}{t}(\phi_t^i - \rho_{t-1}^i). \tag{7}$$

Usually, the Forward-Backward procedure is applied in classical HMM parameter estimation methods to calculate transition probabilities [35]. The goal of the Forward-Backward procedure is to find the conditional distribution over hidden states given the data.

The Forward-Backward procedure (see Figure 4) is an algorithm for HMM which computes the posterior marginals of all hidden state variables given a sequence of observations $o_1, \ldots, o_T$, i.e. it computes, for all hidden state variables $S_t \in \{S_1, \ldots, S_T\}$, the distribution $P(S_t \mid o_{1:T})$. The algorithm uses the principle of dynamic programming to efficiently compute the values that are required to obtain the posterior marginal distributions in two passes. The first pass goes forward in time while the second pass goes backward in time.

The Forward pass is a recursive algorithm for calculating $\alpha_t(i)$ for the observation sequence of increasing length t. First, the probabilities for the single-symbol sequence are calculated as a product of initial i^{th} state probability and emission probability of the given symbol o_1 in the i^{th} state. Then the recursive formula is applied. Assume we have calculated $\alpha_t(i)$ for some t. To calculate $\alpha_{t+1}(j)$, we multiply every $\alpha_t(i)$ by the corresponding transition probability from the i^{th} state to the j^{th} state, sum the products over all states, and then multiply the result by the emission probability of the symbol o_{t+1}. Iterating the process, we can eventually calculate $\alpha_T(i)$, and then summing them over all states, we can obtain the required probability.

In a similar manner, there is a symmetrical backward variable $\beta_t(i)$ as the conditional probability of the partial observation sequence from o_{t+1} to the end to be produced by all state sequences that start at i^{th} state. The Backward pass calculates recursively backward variables going backward along the observation sequence.

However, Forward-Backward procedure is not fully implemented in recursive algorithms as the Backward part of this procedure is often skipped because of the complexity to implement it in real-time systems. We calculate the transition probabilities by adapting the Chapman-Kolmogorov equation into the recursive EM al-

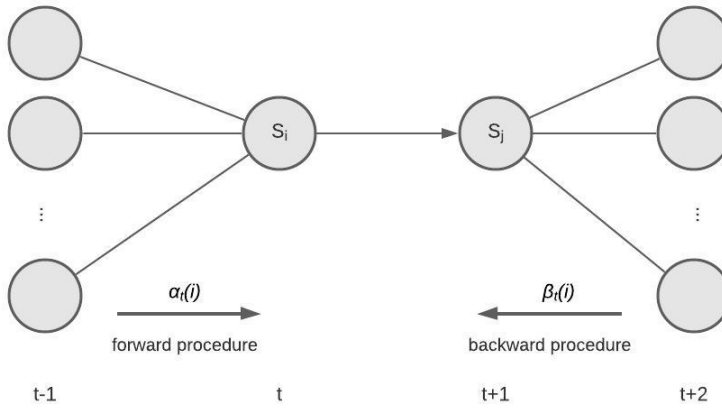


Figure 4. The Forward-Backward procedure in HMM parameter learning

gorithm. It calculates the transition probability to be in a state at time t if at the time moment $t-1$ it was in state i [36]:

$$\pi_t = A \cdot \pi_{t-1}. \tag{8}$$

The significance of this proposition is explored in Section 5.

Our proposed recursive EM algorithm for HMM parameter estimation (Algorithm 1) consists of two parts.

- The first part is for initial parameter estimation given the small fixed-size observation set using formulas (5)–(7). At the initial estimation phase, $\widehat{\mu}$ and $\widehat{\sigma}$ denotes fixed parameter values during estimation. Initial parameter estimation ensures the stability of the algorithm because without it, the algorithm might converge to distorted local extremes of the likelihood function. However, the initial dataset can become the main drawback of the recursive algorithm, so it is very important to have a dataset with a sufficient size to initialize parameter values that would allow us to identify and correctly classify the observations.
- The second part is for parameter re-estimation according to identified observations using formulas (5)–(7)). In the re-estimation phase, $\widehat{\mu}$ and $\widehat{\sigma}$ denotes values of the previous steps μ_{t-1}^i and σ_{t-1}^i. Classification of the observation is performed with a Bayes classifier. The likelihood of each HMM generating the word is calculated and the most likely model identifies the word. However, when dealing with speech recognition, this classifier can be replaced with *Viterbi* or another classification procedure.

Algorithm 1: Recursive EM algorithm for HMM parameter estimation consisting of two parts: a) initial HMM parameter estimation with fixed observation dataset and b) recursive HMM parameter re-estimation when the observation data is given sequentially in real time.

Initial HMM parameter approximation;
Set: $t = 0$;
Initialize: μ_t, σ_t, ρ_t, ϵ, π_1, A;
while *Input observation* O_t, $1 \leq t \leq T_1$ **do**
    Calculate π_t, θ_t;
    Calculate ϕ_t^i, $1 \leq i \leq N$;
    Calculate μ_t^i, σ_t^i, ρ_t^i, $1 \leq i \leq N$;
    if $|\mu_t - \mu_{t-1}| \leq \epsilon$ *AND* $|\sigma_t - \sigma_{t-1}| \leq \epsilon$ **then**
        Result: Output: μ_t, σ_t, ρ_t;
    else
    end
end
HMM parameter re-estimation;
while *Input observation* O_t, $1 \leq t \leq T_1$ **do**
    Input: μ_t, σ_t, ρ_t;
    Calculate π_t, θ_t;
    Calculate ϕ_t^i, $1 \leq i \leq N$;
    Bayes classification of observation O_t*: argmax value of* $e^{l(O_t,\widehat{\mu_i},\widehat{\sigma_i})+\ln(\pi_i)}$;
    Calculate μ_t^i, σ_t^i, ρ_t^i, $1 \leq i \leq N$;
    Result: μ_t, σ_t, ρ_t
end

5 ADAPTATION OF THE RECURSIVE EM ALGORITHM TO ISOLATED WORD RECOGNITION

To apply the recursive EM algorithm to IWR, we must consider the data processing procedure. Isolated words can be processed in two ways – at the symbol/phoneme level or in blocks of information. To adapt the recursive EM algorithm for IWR, the data will be processed in blocks/words.

The first part of the recursive EM algorithm performs an initial approximation to HMM parameters using training data.

In the second part of the algorithm, a recognition procedure was implemented to identify observations. The identification can be performed with a *Viterbi* algorithm that forms a trellis for computing the best hidden state sequence for the observation sequence [1]. Given an observation sequence and HMM, the algorithm returns the state path through the HMM that assigns the maximum likelihood to the observation sequence. HMM parameters are then updated according to the identified word. The scheme for the adapted recursive EM algorithm is presented in Figure 5.

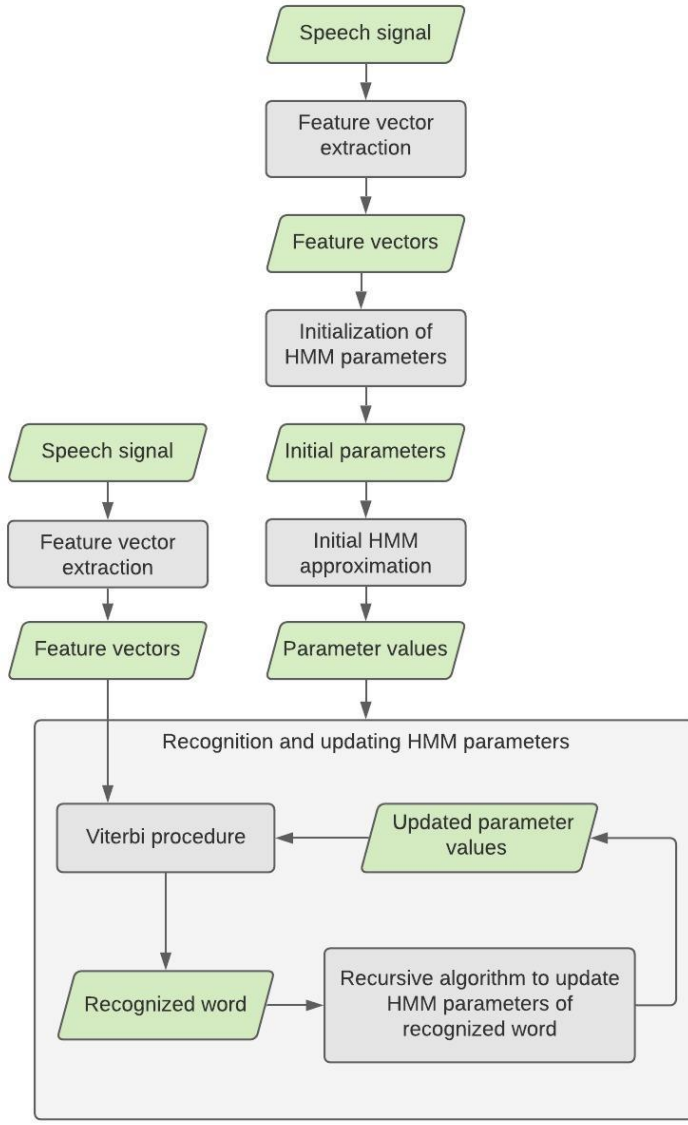


Figure 5. The concept model of an implemented algorithm for isolated word recognition

5.1 Results for Isolated Word Recognition

5.1.1 TIDIGITS Dataset

The recursive EM algorithm was adapted to perform recognition and parameter estimation for isolated speech data. Training and testing were performed with a subset of the *TIDIGITS* dataset [37]. The *TIDIGITS* corpus is used to train the algorithms for speaker-independent recognition of connected digit sequences. The subset consists of 208 speakers (94 men, 114 women) each pronouncing 22 digit sequences (from zero to nine). Each speaker group is partitioned into test and training subsets.

The feature vector consists of 39 features in MFCC format. Each word (digit) was modeled as a ten-state HMM. Each state is modeled with a 39-dimensional mean vector and covariance matrix.

The experiments were conducted in the following manner. First, fixed initial training datasets of various sizes ($100 \leq t \leq 2\,000$ words) were chosen to perform the calculations. Second, further training and recognition were performed with 1 500 word dataset. The word recognition rate (WRR, recognition accuracy) was calculated during the second part of the algorithm.

Word recognition rate can be computed as:

$$WRR = \frac{N - S - D - I}{N}$$

where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of correct words,
- N is the number of words in the reference ($N = S + D + C$).

Values of the state transition probability matrix and the initial state distribution vector were chosen according to [38]. The state transition probability matrix was set to:

$$\begin{bmatrix}
0 & 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.6 & 0.3 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.6 & 0.3 & 0.1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.6 & 0.3 & 0.1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.6 & 0.3 & 0.1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.6 & 0.3 & 0.1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.3 & 0.1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.3 & 0.1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.67 & 0.33 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

The initial state distribution vector was set to: $\begin{bmatrix}0 & 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T$.

The value of the stopping criterion was set to $\epsilon = 0.01$.

The main focus of the experiment was to explore the influence of initial training (approximation) dataset size on the word recognition rate. The results of experiment are presented in Table 1. The first column of the recognition rate has the results of recursive EM algorithm, and the second one has the recognition of traditional Rabiner's isolated word algorithm [8] based on HMM parameter estimation where different sizes of initial training dataset were used. The traditional algorithm was trained with the same parameters as the recursive EM algorithm. The results of both algorithms reveal a similar trend of recognition rate – the word recognition rate increases as the initial training dataset size gets bigger. For an initial dataset size of 100 words, the word recognition rate of the recursive EM algorithm was 92.53 %, and that of the traditional algorithm was 89.15 %. Likewise, for an initial dataset size of 2 000 words, the word recognition rate of the recursive EM algorithm was 97.27 %, and that of the traditional algorithm was 96.03 %. These results show the recursive EM algorithm performs better than the traditional one. This is due to the fact that during the words recognition phase, the recursive EM algorithm updates its model parameters according to the newly received data characteristics.

		Recognition Rate (%)	
		Recursive EM	**Traditional Algorithm**
Size (in words) of the initial training dataset	**100**	92.53	89.15
	500	94.33	93.45
	1 000	95.87	96.73
	1 500	97.60	96.96
	2 000	97.27	96.03

Table 1. Word recognition rate of recursive EM algorithm

It is equally significant to determine an appropriate size for the initial parameter estimation dataset. The experiments show that the increase of initial training dataset results in the higher recognition rate. We see this in both recursive EM and traditional algorithms. The size of the dataset you choose depends on the recognition accuracy you want to achieve. However, in this case of isolated word recognition, it should not be less than a hundred words because a larger training dataset typically prevents the algorithm from converging to a local extreme of the objective function.

5.1.2 Spoken Arabic Digits Dataset

Additional experiments were performed with the *Spoken Arabic Digits* dataset [39], which consists of two parts: training and testing. The training dataset consists of 8 143 observations, which were used for initial HMM parameter learning. The testing dataset consists of 2 665 observations, which were used for re-estimation and recognition in real-time.

For modeling, we used a multivariate Gaussian HMM with multivariate parameters. The dataset consists of a feature vector with 12 features. Thus, HMM states are modeled with a 12-dimensional mean vector and covariance matrix. Each word (digit) was modeled as a 10-state HMM.

The state transition probability matrix was set to:

$$\begin{bmatrix} 0 & 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.6 & 0.3 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0.3 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0.3 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.6 & 0.3 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.6 & 0.3 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.3 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.3 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.67 & 0.33 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The initial state distribution vector was set to: $\begin{bmatrix} 0 & 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$.

The value of the stopping criterion was set to $\epsilon = 0.01$.

The recognition rate was calculated during the re-estimation phase.

The results show that the recognition rate of isolated words recognition was 91.86 %. Out of 2 200 words, the algorithm correctly classified 2 021 words.

5.2 Results of Experiments with Synthetic Data

The following experiment was performed to explore the convergence of the HMM parameters to the original parameter values. The implemented recursive EM algorithm for estimating HMM parameters was compared to the algorithm described in [24]. The main focus of this experiment was to show the impact of a transition probability calculation incorporating the Chapman-Kolmogorov equation. The algorithm from [24] was chosen for comparison because it implements the classic forward-backward procedure skipping the backward part.

To examine the convergence property of the implemented recursive EM algorithm, we calculated the standard error of the estimated HMM model parameters as the difference between the parameter values used to generate the dataset and the estimated model parameters.

The experiments were performed as a simulation of the signal of a single isolated word. Three datasets for 800 multivariate feature vectors consisting of three, five, and twelve features were generated (see Algorithm 2). For data generation each HMM was defined by transition matrix, initial state distribution vector, and mean vector and covariance matrix of each state. Three, five, and twelve dimensional mean vectors and covariance matrices of each state were taken as the excerpts from HMM pre-trained with *TIDIGITS* dataset.

Algorithm 2: Generation of random observation sequences from a Hidden Markov Model

Set length T of the observation sequence;
Set HMM parameters: μ, σ, A and π;
Set state s according to initial state distribution vector;
Set $t = 1$;
while $t \leq T$ **do**
    Generate o_t random numbers according to probability density function with mean μ_s and covariance σ_s at state s;
    Transition to a new state s according to transition probability matrix A;
    $t = t + 1$;
end

For training and recognition each HMM state was modeled with three, five, and twelve dimensional mean vector and a covariance matrix. Each word was modeled as a 5-state HMM in which the state transition probability matrix was set to:

$$\begin{bmatrix} 0 & 0.8 & 0.2 & 0 & 0 \\ 0 & 0.6 & 0.3 & 0.1 & 0 \\ 0 & 0 & 0.6 & 0.3 & 0.1 \\ 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The initial state distribution vector was set to: $\begin{bmatrix} 0 & 0.8 & 0.2 & 0 & 0 \end{bmatrix}^T$.

The value of the stopping criterion was set to $\epsilon = 0.01$.

All experiments were repeated one hundred times.

The results are presented in Table 2. They show that the difference between the estimated parameter values and original parameter values does not increase significantly when the number of dimensions increases in cases of both recursive EM and the algorithm from [24]. The average standard error of the recursive EM algorithm is smaller than the algorithm from [24] for all three simulated datasets (see Figure 6).

This experiment shows the importance of the proposed state transition probability calculation when estimating HMM parameters in a recursive way. The state transition probability calculation with the Chapman-Kolmogorov equation improves the overall parameter estimation compared to an algorithm with only the forward procedure.

The results in Table 2 show that the standard error of the recursive EM algorithm is significantly small. Thus, we can assert that the recursive EM algorithm converges to the original parameter values of the HMM.

Algorithm	**Parameters**	$N = 3$	$N = 5$	$N = 12$
Recursive EM	μ	0.008333	0.007392	0.007792
	σ	0.016833	0.011575	0.004817
Stenger [24]	μ	0.198033	0.171083	0.721975
	σ	0.281967	0.100667	0.109167

Table 2. Standard error of mean and covariance matrices

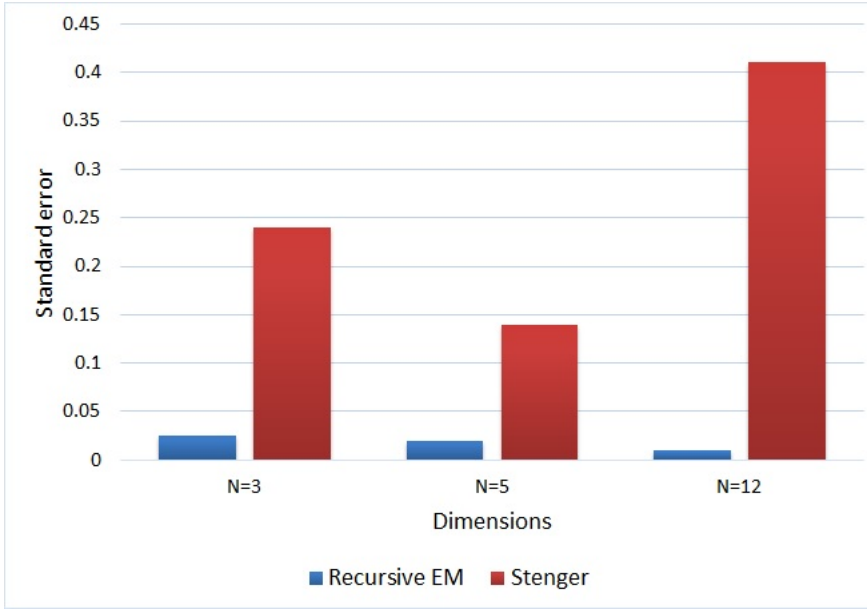


Figure 6. The average standard error of mean vector μ and covariance matrix σ for Recursive EM and Stenger [24] algorithms when the number of dimensions for feature vectors are set to $N = 3$, $N = 5$, and $N = 12$

6 CONCLUSIONS

This paper describes a recursive hidden Markov model multivariate parameter estimation algorithm and its application to on-line isolated word recognition. The recursive EM algorithm presents a novel approach to solving this problem compared to other on-line algorithms. In contrast to the recursive methods where state transition probabilities are obtained with a modified classical Forward-Backward procedure, we calculate the state transition probability by incorporating the Chapman-Kolmogorov equation into the algorithm. The significance of this proposition is shown by a computer simulation comparing it to another recursive algorithm. The results of the experiments showed that our proposed method leads to more accurate parameter estimates. The recursive EM algorithm was also used in three different multivariate datasets, which demonstrated its classification capabilities. The influ-

ence of the initial training dataset size on recognition rate was also explored. The experimental results showed that having a sufficient dataset for initial HMM parameter estimation leads to a word recognition rate higher than 90 %. According to the experiments performed, we can conclude that the recursive EM algorithm can be efficiently applied to real-time speech recognition tasks based on a multivariate HMM model.

REFERENCES

[1] Gruhn, E. R.—Minker, W.—Nakamura, S.: Automatic Speech Recognition. Statistical Pronunciation Modeling for Non-Native Speech Processing, Springer, Berlin, Heidelberg, 2011, pp. 5–17, doi: 10.1007/978-3-642-19586-0_2.

[2] Ultes, S.—Rojas-Barahona, L. M.—Su, P. H.—Vandyke, D.—Casanueva, I.—Budzianowski, P.—Mrkšić, N.—Wen, T. H.—Gašić, M.—Young, S.: PyDial: A Multi-Domain Statistical Dialogue System Toolkit. Proceedings of ACL 2017, System Demonstrations, Association for Computational Linguistics, 2017, pp. 73–78, doi: 10.18653/v1/P17-4013.

[3] McGraw, I.—Prabhavalkar, R.—Alvarez, R.—Arenas, M. G.—Rao, K.—Rybach, D.—Alsharif, O.—Sak, H.—Gruenstein, A.—Beaufays, F.—Parada, C.: Personalized Speech Recognition on Mobile Devices. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5955–5959, doi: 10.1109/ICASSP.2016.7472820.

[4] Kiran, R.—Nivedha, K.—Pavithra Devi, S.—Subha, T.: Voice and Speech Recognition in Tamil Language. 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), 2017, pp. 288–292, doi: 10.1109/ICCCT2.2017.7972293.

[5] Chelba, C.—Schalkwyk, J.—Harb, B.—Parada, C.—Allauzen, C.—Johnson, L.—Riley, M.—Xu, P.—Jyothi, P.—Brants, T.—Ha, V.—Neveitt, W.: Language Modeling for Automatic Speech Recognition Meets the Web. Google Search by Voice, 2012, https://storage.googleapis.com/pub-tools-public-publication-data/pdf/40380.pdf.

[6] Ghose, R.—Dasgupta, T.—Basu, A.: Architecture of a Web Browser for Visually Handicapped People. 2010 IEEE Students Technology Symposium (TechSym), 2010, pp. 325–329, doi: 10.1109/TECHSYM.2010.5469172.

[7] Sun, X.—Miyanaga, Y.—Sai, B.: Dynamic Time Warping for Speech Recognition with Training Part to Reduce the Computation. Journal of Signal Processing, Vol. 18, 2014, No. 2, pp. 89–96, doi: 10.2299/jsp.18.89.

[8] Rabiner, L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, Vol. 77, 1989, No. 2, pp. 257–286, doi: 10.1109/5.18626.

[9] Boruah, S.—Basishtha, S.: A Study on HMM Based Speech Recognition System. 2013 IEEE International Conference on Computational Intelligence and Computing Research, 2013, pp. 1–5, doi: 10.1109/ICCIC.2013.6724147.

[10] Verstraeten, D.—Schrauwen, B.—Stroobandt, D.—Van Campenhout, J.: Isolated Word Recognition with the Liquid State Machine: A Case Study. Information Processing Letters, Vol. 95, 2005, No. 6, pp. 521–528, doi: 10.1016/j.ipl.2005.05.019.

[11] Fohr, D.—Mella, O.—Illina, I.: New Paradigm in Speech Recognition: Deep Neural Networks. IEEE International Conference on Information Systems and Economic Intelligence, 2017, https://hal.archives-ouvertes.fr/hal-01484447.

[12] Khreich, W.—Granger, E.—Miri, A.—Sabourin, R.: A Survey of Techniques for Incremental Learning of HMM Parameters. Information Sciences, Vol. 197, 2012, pp. 105–130, doi: 10.1016/j.ins.2012.02.017.

[13] Khreich, W.—Granger, E.—Miri, A.—Sabourin, R.: On The Memory Complexity of the Forward-Backward Algorithm. Pattern Recognition Letters, Vol. 31, 2010, No. 2, pp. 91–99, doi: 10.1016/j.patrec.2009.09.023.

[14] Cappé, O.—Moulines, E.: On-Line Expectation-Maximization Algorithm for Latent Data Models. Journal of the Royal Statistical Society, Vol. 71, 2009, No. 3, pp. 593–613, doi: 10.1111/j.1467-9868.2009.00698.x.

[15] He, J.— Mao, R.—Shao, Z.—Zhu, F.: Incremental Learning in Online Scenario. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13923–13932, doi: 10.1109/CVPR42600.2020.01394.

[16] Castro, F. M.—Marín-Jiménez, M. J.—Guil, N.—Schmid, C.—Alahari, K.: End-to-End Incremental Learning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11216, 2018, pp. 241–257, doi: 10.1007/978-3-030-01258-8_15.

[17] Losing, V.—Hammer, B.—Wersing, H.: Incremental On-Line Learning: A Review and Comparison of State of the Art Algorithms. Neurocomputing, Vol. 275, 2018, pp. 1261–1274, doi: 10.1016/j.neucom.2017.06.084.

[18] Royer, A.—Lampert, C. H.: Classifier Adaptation at Prediction Time. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1401–1409, doi: 10.1109/CVPR.2015.7298746.

[19] Wu, Y.—Chen, Y.—Wang, L.—Ye, Y.—Liu, Z.—Guo, Y.—Fu, Y.: Large Scale Incremental Learning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 374–382, doi: 10.1109/CVPR.2019.00046.

[20] Rebuffi, S.-A.—Kolesnikov, A.—Sperl, G.—Lampert, C. H.: iCaRL: Incremental Classifier and Representation Learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5533–5542, doi: 10.1109/CVPR.2017.587.

[21] Huo, Q.—Lee, C. H.: On-Line Adaptive Learning of the Correlated Continuous Density Hidden Markov Models for Speech Recognition. Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP '96), Vol. 2, 1996, pp. 985–988, doi: 10.1109/ICSLP.1996.607768.

[22] Ephraim, Y.—Merhav, N.: Hidden Markov Processes. IEEE Transactions on Information Theory, Vol. 48, 2002, No. 6, pp. 1518–1569, doi: 10.1109/TIT.2002.1003838.

[23] Holst, U.—Lindgren, G.: Recursive Estimation in Mixture Models with Markov Regime. IEEE Transactions on Information Theory, Vol. 37, 1991, No. 6, pp. 1683–1690, doi: 10.1109/18.104334.

[24] Stenger, B.—Ramesh, V.—Paragios, N.—Coetzee, F.—Buhmann, J. M.: Topology Free Hidden Markov Models: Application to Background Modeling. Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001), Vol. 1, 2001, pp. 294–301, doi: 10.1109/ICCV.2001.937532.

[25] Krishnamurthy, V.—Moore, J. B.: On-Line Estimation of Hidden Markov Model Parameters Based on the Kullback-Leibler Information Measure. IEEE Transactions on Signal Processing, Vol. 41, 1993, No. 8, pp. 2557–2573, doi: 10.1109/78.229888.

[26] Mongillo, G.—Deneve, S.: Online Learning with Hidden Markov Models. Neural Computation, Vol. 20, 2008, No. 7, pp. 1706–1716, doi: 10.1162/neco.2008.10-06-351.

[27] Cappé, O.: Online EM Algorithm for Hidden Markov Models. Journal of Computational and Graphical Statistics, Vol. 20, 2011, No. 3, pp. 728–749, doi: 10.1198/jcgs.2011.09109.

[28] LeGland, F.—Mevel, L.: Recursive Estimation in Hidden Markov Models. Proceedings of the 36th IEEE Conference on Decision and Control, Vol. 4, 2002, pp. 3468–3473, doi: 10.1109/CDC.1997.652384.

[29] Tadić, V. B.: Analyticity, Convergence, and Convergence Rate of Recursive Maximum-Likelihood Estimation in Hidden Markov Models. IEEE Transactions on Information Theory, Vol. 56, 2010, No. 12, pp. 6406–6432, doi: 10.1109/TIT.2010.2081110.

[30] Slívová, M.—Partila, P.—Továrek, J.—Vozňák, M.: Isolated Word Automatic Speech Recognition System. In: Dziech, A., Mees, W., Czyżewski, A. (Eds.): Multimedia Communications, Services and Security (MCSS 2020). Springer, Cham, Communications in Computer and Information Science, Vol. 1284, 2020, pp. 252–264, doi: 10.1007/978-3-030-59000-0_19.

[31] Vaseghi, S. V.: Hidden Markov Models. Chapter 5. Advanced Digital Signal Processing and Noise Reduction. Fourth Edition. John Wiley & Sons, 2009, pp. 147–172, doi: 10.1002/9780470740156.ch5.

[32] Neal, R. M.—Hinton, G. E.: A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants. In: Jordan, M. I. (Ed.): Learning in Graphical Models. Springer, Dordrecht, NATO ASI Series (Series D: Behavioural and Social Sciences), Vol. 89, 1998, pp. 355–368, doi: 10.1007/978-94-011-5014-9_12.

[33] Ghahramani, Z.: An Introduction to Hidden Markov Models and Bayesian Networks. In: Bunke, E., Caelli, T. (Eds.): Hidden Markov Models: Applications in Computer Vision. World Scientific, Series in Machine Perception and Artificial Intelligence, Vol. 45, 2001, pp. 9–41, doi: 10.1142/9789812797605_0002.

[34] Nánási, M.—Vinař, T.—Brejová, B.: Sequence Annotation with HMMs: New Problems and Their Complexity. Information Processing Letters, Vol. 115, 2015, No. 6-8, pp. 635–639, doi: 10.1016/j.ipl.2015.03.002.

[35] Stamp, M.: Introduction to Machine Learning with Applications in Information Security. Chapman and Hall/CRC, 2017, pp. 7–36, doi: 10.1201/9781315213262.

[36] Durrett, R.: Probability: Theory and Examples. Cambridge University Press, 2019, pp. 232–285, doi: 10.1017/9781108591034.006.

[37] Leonard, R. G.—Doddington, G. R.: TIDIGITS LDC93S10: Philadelphia: Linguistic Data Consortium, 1993, https://catalog.ldc.upenn.edu/LDC93S10.

[38] Young, S.—Kershaw, D.—Odell, J.—Ollason, D.—Valtchev, V.—Woodland, P.: The HTK Book, Microsoft Corporation, 2000, https://htk.eng.cam.ac.uk/docs/docs.shtml.

[39] Dua, D.—Graff, C.: Spoken Arabic Digits in UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2017, http://archive.ics.uci.edu/ml.

Jūratė Vaičiulytė received her Bachelor and Master degrees in informatics, computer science from Šiauliai University in 2013 and 2015, respectively. She then received Ph.D. degree in informatics from Vilnius University, Lithuania in 2020. Her main scientific interests are focused on computer modelling, classification, and pattern recognition.

Leonidas Sakalauskas is active in science since the 1980s and has published more than 250 publications in reviewed scientific journals in the areas of stochastic optimization, data mining, and operation research. He developed the concept of implementable stochastic numerical methods, created an approach for stochastic nonlinear programming by Monte-Carlo estimators with admissible accuracy, etc. His main scientific interests are focused on stochastic optimization, queuing theory, and other fields of operation research.

Computing and Informatics, Vol. 40, 2021, 298–317, doi: 10.31577/cai_2021_2_298

LOGISTIC REGRESSION BASED ON STATISTICAL LEARNING MODEL WITH LINEARIZED KERNEL FOR CLASSIFICATION

Xiaochun Guan, Jianhua Zhang, Shengyong Chen

School of Computer Science and Technology
Zhejiang University of Technology
Liuhe Road No. 288
310 023 Hangzhou, China
e-mail: guanxc@wzu.edu.cn, zjh@zjut.edu.cn, csy@tjut.edu.cn

Abstract. In this paper, we propose a logistic regression classification method based on the integration of a statistical learning model with linearized kernel pre-processing. The single Gaussian kernel and fusion of Gaussian and cosine kernels are adopted for linearized kernel pre-processing respectively. The adopted statistical learning models are the generalized linear model and the generalized additive model. Using a generalized linear model, the elastic net regularization is adopted to explore the grouping effect of the linearized kernel feature space. Using a generalized additive model, an overlap group-lasso penalty is used to fit the sparse generalized additive functions within the linearized kernel feature space. Experiment results on the Extended Yale-B face database and AR face database demonstrate the effectiveness of the proposed method. The improved solution is also efficiently obtained using our method on the classification of spectra data.

Keywords: Elastic net, generalized additive model, kernel, lasso regression, spectra data

Mathematics Subject Classification 2010: 68U10

1 INTRODUCTION

In statistical data modeling, regression is a popular method to explore low-dimensional structures in Statistics and Computation. It uses the samples to estimate

the parameters in the formula [1, 2]. It is a widely used statistical analysis method for data modeling. There are two standards to evaluate the regression. One is the prediction accuracy, the other is a better interpretation. Prediction accuracy means the model's prediction accuracy on future testing data. The interpretation of the model refers to a more parsimonious model. Parsimony plays an important role in inference. The ordinary least squares (OLS) regression is obtained by minimizing the residual squared error, ridge regression plus the square sum of the regression coefficients as a penalty function on the residual squared error. The OLS regression tends to obtain a lower prediction accuracy compared with ridge regression. Ridge regression shrinks coefficients continually and hence is more practical and reliable. Its prediction accuracy is better than OLS regression, but it does not set any coefficients to 0 and hence it does not improve the model's interpretation. Subset selection can provide interpretable models because regressors are either retained or dropped from the model by subset selection, but its prediction accuracy tends to be very unstable because of its inherent discreteness. An influential regularization technique called least absolute shrinkage and selection operator (lasso) was proposed by Tibshirani [3]. Lasso is a penalized least squares method that imposes an L1 penalty on the regression coefficients. Owing to the sparse nature of the L1 penalty, the lasso can compress some coefficients and simultaneously set some coefficients to zero, thus it can produce a sparse representation of the model. It also has been proved that the L1 penalty can discover the "right" sparse representation of the model under certain conditions [4, 5, 6]. The success of the lasso is accomplished by the L1 penalty applied to the coefficients. This L1 penalty approach is also called basis pursuit in the field of signal processing [7]. In 2004, Efron et al. proposed the least angle regression algorithm (LARS) to solve the entire lasso solution path efficiently [8]. The LARS makes lasso widely used in feature selection and parameter estimation. Lasso also has been supported by much theoretical work in sparse representation and compressed sensing. Especially since 2006, Donoho and Tao et al. have put forward a theoretical basis for compressed sensing, which successfully constructed theory and practical methods in the field [4, 9, 10, 11, 12, 13].

However, for high dimension and small sample data, such as the gene selection problem in microarray data analysis, the lasso can not select the grouping information in situations consisting of grouped variables [14]. Zou et al. proposed a new feature selection algorithm called elastic net. The elastic net can not only simultaneously do automatic variable selection and continuous shrinkage, but also select groups of closely correlated variables, i.e. either group selection or omission of the correlated variables. Also in 2015, Chouldechova and Hastie introduced an extension of the lasso to the additive model setting, the method is called Generalized Additive Model Selection (GAMSEL), this method can select among zero, linear and non-linear fits as component functions in a generalized additive model framework by an overlap group-lasso penalty [15]. It also incorporates a penalized likelihood procedure for fitting sparse generalized additive models.

In this paper, we consider the classification as a multinomial/binomial logistic regression problem. There is research on the regression coefficients matrix of multinomial regression [16]. Due to the important and remarkable applications of face recognition technology, there have been many successful algorithms for face recognition, such as sparse representation classification, linear regression, elastic net et al. [11, 17, 18, 19, 20, 21]. In recent years, significant progress has been made on face recognition systems [22, 23]. Especially in [24], a semi-supervised sparse representation based classification method is proposed to address the problem of face recognition when the labeled samples are insufficient. Face recognition can be considered as a multi-class classification problem, it can also be regarded as a multinomial logistic regression problem. We use the elastic net's grouping effect to find the grouping features in the linearized kernel (LK) feature space of the samples based on the statistical learning model-the generalized linear model. In the model, an elastic net penalized negative log-likelihood function method was adopted to perform variables selection for classification. The elastic net regularization can also properly adapt to the situation where the number of samples is much smaller than the predicted variables. Thus the algorithm can perform well for some face databases which do not contain enough samples in each sample space. Simulation and experiments on publicly available face data and Raman spectra data are used to demonstrate the feasibility of our proposed method. The experiment results show that our method improved the classification accuracy by up to 0.83 % and 3.7 % on the Extended Yale-B face database and AR face database respectively compared with the best result in [25]. We also show the classification results of the GAMSEL model with or without LK pre-processing on spectra data, our method with LK pre-processing can improve the performance by 10 %. We apply the GAMSEL on a subset of Raman spectra data with or without LK pre-processing for the binomial logistic regression problem, the GAMSEL can fit the nonlinear functions within the linearized kernel feature space on the subset of Raman spectra data. It shows that the binomial classification accuracy of the subset of Raman spectra data can be improved with LK pre-processing based on GAMSEL.

The main contributions of this paper can be summarized in two aspects. First, it proposes a novel method that integrates linearized kernel pre-processing into a statistical learning model for multiclass classification. It provides us a perspective to explore the low-dimensional space embedded in the high dimension data. Second, it adopts the fusion of Gaussian and cosine kernels for linearized kernel pre-processing with improved accuracy compared with a single Gaussian kernel. The rest of this paper is organized as follows: Section 2 briefly introduces sparsity and statistical learning. Section 3 depicts information on the kernel and linearized kernel pre-processing. Section 4 describes the logistic regression classification method combining the statistical learning model with linearized kernel pre-processing. Section 5 elaborates extensive experiments. Section 6 includes the analysis and conclusion remarks.

2 SPARSITY AND STATISTICAL LEARNING

Research of sparse representation had started in the 1990s [26], it has been flourishing since the beginning of this century. Sparse representation of signal has attracted many concerns from researchers, for example, the typical image compression algorithm JPEG utilizes image's sparsity in the DCT domain to achieve image compression. The core model in the sparse domain is the linear equations to describe an underdetermined system that has infinitely many solutions. The sparsest solution, which has the least nonzero terms, is the most interesting. The L0 norm can find a sparse solution, and the L0 norm represents the total number of non-zero elements in a vector $\mathbf{x}$. The L0 norm can be defined as in Equation (1), x_i is the elements of the vector $\mathbf{x}$.

$$\|\mathbf{x}\|_0 = \#\left(i \mid x_i \neq 0\right). \tag{1}$$

However, the optimization problem of the L0 norm is an NP-hard problem. It is proved theoretically that the L1 norm is the optimal convex approximation of the L0 norm, so the L1 norm is usually used instead of the L0 norm. The L1 norm represents the sum of the absolute values of each element in a vector. The L1 norm can be defined as in Equation (2).

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i|. \tag{2}$$

The solution of the L1 norm is usually sparse and tends to select a very small number of very large values or a small number of insignificant values. L1 norm regularization adds the L1 norm to the cost function, which makes the learning result satisfy the sparsity, so the main features can be extracted. L1 norm has become a popular tool in many research fields [27]. Lasso is one typical example of L1 norm regularization. Given the predictors, $x_1, \ldots, x_p$, the usual linear regression model with response y can be predicted by Equation (3).

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \ldots + x_p\hat{\beta}_p. \tag{3}$$

The vector of coefficients $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ should be fitted by the model. We can assume without loss of generality that the mean of y is 0 and hence omit β_0. Lasso regularization can be defined as in Equation (4).

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\| y - \sum_{j=1}^{p} x_j\beta_j \right\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq t \tag{4}$$

where t is a nonnegative tuning parameter. It can control the amount of shrinkage that is applied to the estimates. If t is sufficiently small, the lasso can cause continuous shrinkage of the coefficients to 0 as t decreases, and some coefficients can be exactly shrunk to zero. The bias and variance trade-off introduced by the L1 norm

penalty can lead to the coefficients continuous shrinkage and variable selection and thus improve the prediction accuracy [28]. Lasso can produce a parsimonious model. However, lasso can not also reveal the grouping information in situations consisting of grouped variables. To overcome its limitations, Zou et al. proposed a new feature selection algorithm called elastic net [14]. Elastic net combines the L1 norm and L2 norm together as the penalty function on the regression coefficients. The elastic net estimates are defined as in Equation (5).

$$\hat{\boldsymbol{\beta}}_{\text{Enet}} = \arg\min_{\boldsymbol{\beta}} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 \quad \text{s.t. } (1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2 \leq t. \tag{5}$$

$\|\boldsymbol{\beta}\|_1$and $\|\boldsymbol{\beta}\|_2$ represent the vector β's L1 norm and L2 norm, respectively. L1 norm refers to the sum of absolute values of each element in a vector. L2 norm is the square root of the sum of squares of each element in a vector. Similar to lasso, the algorithm LARS-EN is proposed to compute the entire elastic net regularization paths efficiently [14], just like algorithm LARS for the lasso. However, both for the lasso and elastic net, it retains the linear fit for all the predictors, Chouldechova and Hastie introduced an extension of the lasso to the additive model setting in 2015. The method is called Generalized Additive Model Selection (GAMSEL), this method selects between zero, linear and non-linear fits for predictor functions in a generalized additive model framework by an overlap group-lasso penalty. It incorporates a penalized likelihood procedure for fitting sparse generalized additive models [15]. For data $(\mathbf{x}, y)$, a simple linear fit is of the form in Equation (6).

$$\eta(\mathbf{x}) = \sum_{j=1}^{p} \beta_j x_j. \tag{6}$$

For more generative, the generalized additive model was defined as in Equation (7).

$$\eta(\mathbf{x}) = \sum_{j=1}^{p} f_j\left(x_j\right) \tag{7}$$

where the f_j are unknown functions that should be estimated, which can be zero, linear or nonlinear. The GAMSEL is to optimize a penalized negative log-likelihood criterion of the form defined in Equation (8).

$$\hat{f}_1, \ldots, \hat{f}_p = \arg\min_{f_1, \ldots f_p \in \mathcal{F}} l\left(y; f_1, \ldots, f_p\right) + \sum_{j=1}^{p} J\left(f_j\right). \tag{8}$$

It can fit each f_j as zero, linear or nonlinear, as determined by the data. It can capture non-linear relationships among the data. In this paper, we also employ it to explore the nonlinearity of the linearized kernel feature space. The last term in Equation (8) represents the sum of the penalty term of each component f_j. For more detail, please refer to [15].

3 KERNEL AND LINEARIZED KERNEL PRE-PROCESSING

In recent years, with the development of machine learning, a series of kernel function learning methods have been developed. The kernel method is a powerful nonparametric modeling tool. In some cases, it can make problems such as classification and regression easier to solve. It is the Reproducing Kernel Hilbert Space (RKHS) underlying the kernel method that provides linearity, convexity, and general approximation capability, the research of RKHS technology began in the 1940s. The theory of kernel function can be traced back to 1909. The main idea of the kernel method is to transform low-dimensional linear inseparable data into high-dimensional linear separable data. It transforms low-dimensional data into a high-dimensional feature space by a kernel function, which can be equivalent to the inner product of corresponding high-dimensional feature vectors. Then the high-dimensional feature data can be processed by the appropriate linear method as long as the algorithm of that linear method can be expressed by the inner product of the high-dimensional feature vectors of the samples. It is not necessary to know what is the specific high-dimensional features. This is called the "kernel trick". In 1992, Vapnik et al. successfully used this technology to extend linear support vector machine (SVM) to nonlinear SVM [29], its potential was fully realized by researchers. The popular kernel functions mainly include the Gaussian kernel, the polynomial kernel, and the sigmoid kernel. In [30], the author proposed a fusion kernel that fuses the Euclidean and cosine distance measures. The fusion kernel can also be applied to our problem achieving better performance.

Recently kernel method has been widely used in the machine learning field. The kernel method has played an important role in system identification, machine learning, and function estimation [31]. Kernel method has been integrated with sparse representation-based classifier (KSRC) for face recognition with good representation and classification performance [32]. Recently, a new video semantic analysis method with kernel discriminative sparse representation was adopted to efficiently detect the event and concept in video surveillance [33]. Kernel-based machine learning method has been also applied for Chinese license plate recognition [34]. Kernel method is a common way of extending a specific algorithm to deal with a higher dimension "feature space", it has been also incorporated into dictionary learning (DL) [25, 35, 36, 37, 38, 39, 40, 41, 42]. One typical application is its incorporation with dictionary learning in sparse land. There are many successful image processing applications based on DL [43, 44, 45, 46, 47]. Popular algorithms for DL are Method of Optimal Directions (MOD) [48] and other algorithms based on K-means clustering via singular value decomposition (K-SVD), such as label consistent K-SVD1 (LC-KSVD1), label consistent K-SVD2 (LC-KSVD2) and the kernel K-SVD algorithm (KKSVD) [36, 49, 50]. And in [25], Golts et al. give out a new method of incorporating linearized kernel pre-processing into these dictionary learning algorithms, termed "Linearized Kernel Dictionary Learning" (LKDL), which typically gets a relatively good experiment result compared with LC-KSVD1 and LC-KSVD2.

In this paper, we explore the linearized kernel feature space using a statistical learning model. Our method is inspired by the kernel DL method termed "Linearized Kernel Dictionary Learning" (LKDL) by Golts and Elad [25]. They introduce a pre-processing stage based on the kernel method for the dictionary learning algorithm. The idea of Linearized Kernel (LK) pre-processing using the Nyström method to obtain a good approximation of the regular kernel matrix. Please refer to [25] for the details of LKDL. The LK pre-processing method can address the problems of high computational cost and the large storage space for a very large kernel matrix when the kernel trick is used. Without too much effort the LK pre-processing can be incorporated into the algorithms as a kernel layer application. This paper proposed a logistic regression classification method based on the fusion of statistical learning models and linearized kernel pre-processing. The adopted two statistical learning models are the generalized linear model and the generalized additive model. With the generalized linear model, the elastic net regularization is adopted to explore the grouping effect of the linearized kernel feature space. With the generalized additive model, an overlap group-lasso penalty is used to fit the sparse generalized additive functions within the linearized kernel feature space. It can explore the nonlinearity of the linearized kernel feature space.

4 LOGISTIC REGRESSION AND STATISTICAL LEARNING MODEL

Logistic regression is a widely-used method for classification. The logistic regression method is mainly applied to the study of the occurrence probability of certain events. When there are more than two possible outcomes in a problem, multinomial logistic regression can be adopted. For logistic regression, when facing a regression or classification problem, firstly it establishes a cost function, and then iteratively solves the optimal model parameters through a specific optimization method on a training set, and then to verify the quality of the logistic regression model on the testing set. In this study, we adopt two statistical learning models to do logistic regression. They are the generalized linear model and the generalized additive model. Supposing the response variable has K classes $G = (1, 2, \ldots, K)$, for the multinomial logistic regression model, the model can be defined as follows:

$$\Pr(G = k \mid X = \mathbf{x}) = \frac{e^{\beta_{0k} + \beta_k^T \mathbf{x}}}{\sum_{l=1}^{K} e^{\beta_{0l} + \beta_l^T \mathbf{x}}}. \tag{9}$$

For the multinomial logistic regression model, we adopt the Glmnet R package. The Glmnet can fit the generalized linear model via penalized maximum likelihood. Its regularization path can be computed for the elastic net penalty at different regularization parameter lambda. The Glmnet's elastic-net penalized negative log-likelihood function is defined as Equation (10), which can realize the grouping effect

of variables [51].

$$l\left(\{\beta_{0k},\beta_k\}_1^K\right) = -\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{il}\left(\beta_{0k}+\mathbf{x}_i^T\beta_k\right)-\log\left(\sum_{k=1}^{K} e^{\beta_{0k}+x_i^T\beta_k}\right)\right)\right] + \lambda\left[(1-\alpha)\|\beta\|_F^2/2+\alpha\sum_{j=1}^{p}\|\beta_j\|_q\right]. \tag{10}$$

Here Y to be the $N\times K$ indicator response matrix, with elements $y_{il}=\mathrm{I}(\mathrm{g_i}=1)$, I() is the indication function. β is a $p\times K$ matrix of coefficients. β_k refers to the k^{th} column of outcome class k, and β_j refers to the j^{th} row vector of K coefficients for variable j. For the last penalty term $\|\beta_j\|_q$, if $q=2$, it is a grouped-lasso penalty on all the K coefficients for the particular variables. The tuning parameter λ controls the overall strength of the penalty.

The algorithm flow of multinomial logistic regression with elastic net (MLR-elastic net) is shown in Algorithm 1. It is based on the generalized linear statistical learning model with LK pre-processing.

Algorithm 1. Multinomial logistic regression with elastic net based on linearized kernel pre-processing

1: Input: $\mathrm{X_{train}}=[\mathrm{X}_l,\ldots,\mathrm{X}_L]$, $\mathrm{X_{test}}$, the kernel κ, sampling-method, c, k
2: $\mathrm{X}_R=\mathrm{sub\_sample}\ (\mathrm{X_{train}}\ ,\ \text{sampling-method},c)$
3: Compute $\mathrm{C_{train}}=\mathrm{K}\left(\mathrm{X_{train}},\mathrm{X}_R\right)$
4: Compute $\mathrm{W}=\mathrm{K}\left(\mathrm{X}_R,\mathrm{X}_R\right)$
5: Approximate W_k using k largest eigenvalues and eigenvectors $\mathrm{W}_k=\mathrm{V}_k\Sigma_k\mathrm{V}_k^T$
6: Compute virtual train set $\mathrm{F_{train}}=\left(\Sigma_k^{\dagger}\right)^{1/2}\mathbf{V}_k^T\mathbf{C}_{\text{train}}^T$
7: Compute $\mathrm{C_{test}}=\mathrm{K}\left(\mathrm{X_{test}},\mathrm{X}_R\right)$
8: Compute virtual test set $\mathrm{F_{test}}=\left(\Sigma_k^{\dagger}\right)^{1/2}\mathbf{V}_k^T\mathbf{C}_{\text{test}}^T$
9: Using $\mathrm{F_{train}}$ to obtain the model parameters by multinomial logistic regression with elastic net based on generalized linear model
10: Carry out classification of $\mathrm{F_{test}}$ using the model obtained above
11: Output: classification result of $\mathrm{F_{test}}$

For the binomial logistic regression problem, the generalized additive model is adopted. The GAMSEL R package is used. The generalized additive model uses overlap grouped-lasso penalties, it can select whether a term in a general additive model is zero, linear, or a non-linear spline for Gaussian or binomial applications [15]. We adopt LK pre-processing for binomial logistic regression based on the GAMSEL. The algorithm flow is the same as in Algorithm 1. The difference only lies in feeding the virtual samples to the GAMSEL model. Figure 1 shows the block diagram of our proposed method.

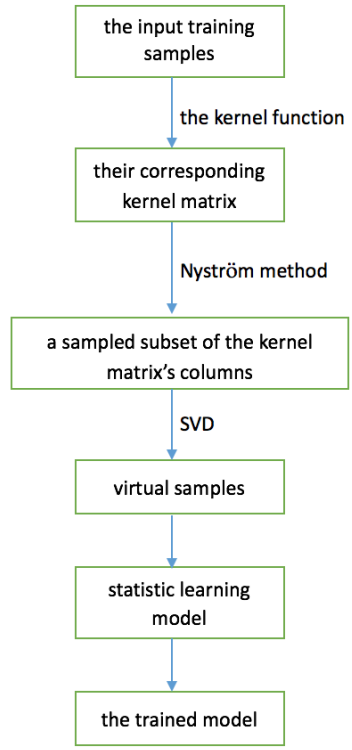


Figure 1. Block diagram of the proposed method

5 EXPERIMENT AND SIMULATION

In this section, we evaluate the performance of multinomial/binomial logistic regression based on a statistical learning model with LK pre-processing on face recognition databases and spectra dataset.

5.1 Evaluation on Face Recognition Database

The adopted face recognition databases are the Extended YaleB face database and AR-face database [52, 53]. The "Extended YaleB" face database has 38 classes with 2 414 frontal face images taken under varying lighting conditions. Each class nearly has 64 images. The AR Face database possesses 126 classes with 4 000 color images, which are with different lighting conditions, facial variations, and facial disguises for each class. For the sake of fairness and convenience for contrast experiments, our experimental details are configured with the same settings as [25]. For LK pre-

processing, the Gaussian kernel and the fusion kernel [30] are used. The specific parameters of the Gaussian kernel are configured with the same configuration as in [25]. For a fair comparison, in the case of LC-KSVD1 and LC-KSVD2, the parameters are chosen as identical to the best classification result in [25]. The classification results were shown in Table 1 with our method comparing with the best results obtained in [25]. The LK pre-processing with the fusion kernel is already stated in Table 1. Other LK pre-processing is with the Gaussian kernel. It can be seen that the addition of LK pre-processing and the elastic net's grouping effect can increase the prediction accuracy. The experiment result also shows that the LK pre-processing with the manual fusion kernel outperforms that with the Gaussian kernel. The fusion kernel can exploit the reciprocating properties of the Euclidean and cosine distance measures. We also used the support vector machine(SVM) toolbox LIBSVM [54] in our experiments. We use a coarse grid search for the SVM parameters, we use the Gaussian kernel. We use the grid search strategy to look for suitable parameters. We chose $g \in [0.001\ 0.001\ 0.01\ 0.1\ 0.5\ 1\ 10\ 100\ 500\ 1\,000]$ and $c \in [0.000006\ 0.000008\ 0.000009\ 0.0000092\ 0.0000093\ 0.0000095\ 0.0000096\ 0.0000097\ 0.0000098\ 0.00001]$ and run the search for 100 times. And choosing the parameters for the best accuracy. Experiment result with SVM using LIBSVM toolbox shows inferior accuracy compared to that of the proposed method. For method comparisons, we use the same random training and testing samples for each algorithm. Our method improves the classification results by up to 2.42 % and 4.8 %, when compared to LC-KSVD2 results [25] on the Extended YaleB face dataset and AR-face dataset, respectively. Due to the grouping effect of the elastic net, a group of related or correlated variables can be detected, when the LK pre-processing is adopted, more grouped features from the high dimension kernel space can be incorporated, so it can obtain the grouped features in the high dimension, which can lead to increased performance. Figure 2 shows the solution path of multinomial regression with the elastic net for the 38^{th} subject on Yale-B face database based on the Glmnet. Each curve corresponds to a variable. Each curve shows the solution path of its coefficient against the L2 norm of the whole coefficient vector as $\log(\lambda)$ varies. The axis above indicates the number of nonzero coefficients at each corresponding $\log(\lambda)$. There are about 332 variables selected for this algorithm when $\log(\lambda)$ is equal to -8. It shows that more variables will be shrunk to be zero eventually as $\log(\lambda)$ increases. The λ represents the tuning parameter in Equation (10) for controlling the overall strength of the penalty.

To further justify the performance of our method, on the AR face database, we perform simulations by randomly selecting training samples, and the remaining samples are used as testing samples. Figure 3 shows the classification accuracy of a total of 20 experiments by randomly selecting 20 training samples from the input samples. The Y-axis is the testing accuracy and the horizontal axis is the experiment number(ID). We adopted different kernel functions for LK preprocessing, the solid line in Figure 3 uses the Gaussian kernel, the dashed line in Figure 3 uses the fusion kernel. The result shows that the proposed method is effective.

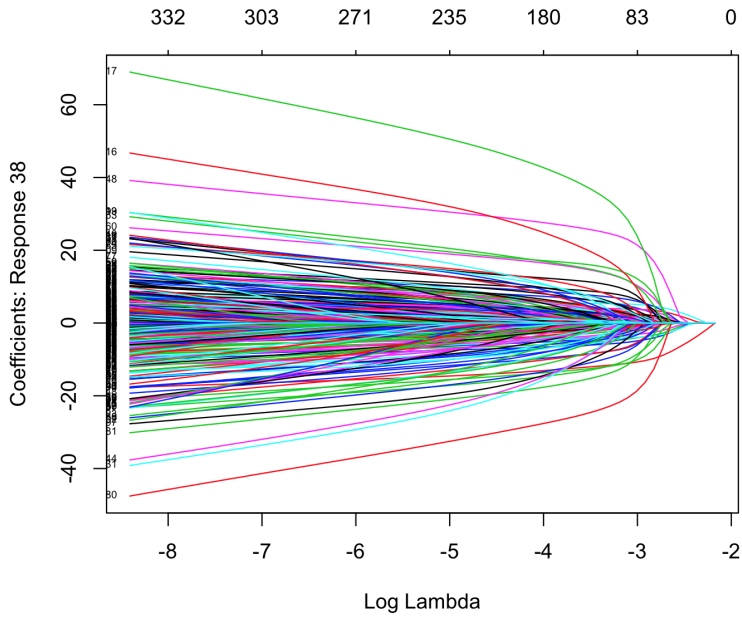


Figure 2. Solution path of multinomial logistic regression with elastic net for the 38^{th} subject on Extended Yale-B face database

We also use 10 times 5-fold cross validation to further show the effectiveness of our method on the Extended Yale-B face database and AR face database. The experiment was shown in Table 2.

For the LK pre-processing, the time complexity is $\mathrm{O}\left(Nck + c^2k\right)$, $\mathrm{O}(Nck)$ represents the complexity of getting the virtual samples, $\mathrm{O}\left(c^2k\right)$ stands for the eigenvalue decomposition of the sampled kernel matrix. Although the process of computing the

Algorithm	Yale-B	AR-Face
LC-KSVD1	94.49	92.5
LC-KSVD1 + LK	96.08	94.8
LC-KSVD2	94.99	93.7
LC-KSVD2 + LK	96.58	94.8
SVM	91.32	95.5
SVM + LK	95.41	94.7
MLR-elastic net	93.90	97.3
MLR-elastic net+ LK	97.16	98.3
MLR-elastic net+ LK (fusion kernel)	**97.41**	**98.5**

Table 1. Classification accuracy of LC-KSVD1, LC-KSVD2, SVM and our method on Extended Yale-B and AR face database, with or without LK pre-processing

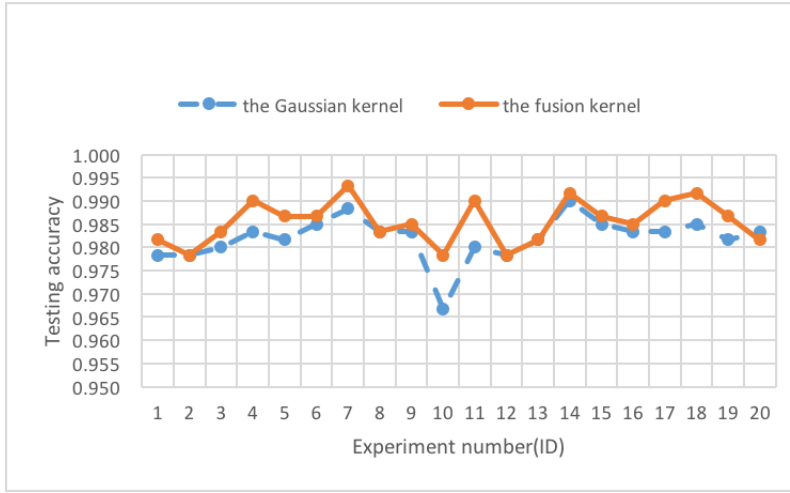


Figure 3. Classification accuracy by randomly selecting training samples with different kernel function on AR face database

Algorithm	Yale-B	AR-Face
SVM + LK	95.41	94.7
MLR-elastic net+ LK (fusion kernel)	**98.73**	**98.62**

Table 2. Average classification accuracy of our method on Extended Yale-B and AR face database with LK pre-processing by 10 rounds 5-fold cross validation

virtual samples may seem inefficient, it is only performed once, after which the complexity is dictated by the chosen model. The total training time and test time required to classify one sample with or without LK pre-processing on the AR face are shown in Table 3. The experiment is carried out on MacBook Pro with a 2.6 GHz Intel Core i5 processor and 8 GB 1 600 MHz DDR3 memory. Its operation system is OS X Yosemite 10.10.2.

Algorithm	Total training time	Test time for one sample
MLR-elastic net	579.503 s	0.006 s
MLR-elastic net + LK (fusion kernel)	**323.125 s**	**0.004 s**

Table 3. Total training time and testing time required to classify one sample with or without LK pre-processing on AR face database

It shows that the training time and test time on the statistical model is decreased greatly with LK pre-processing.

5.2 Evaluation on Spectra Data

We also evaluated our method on the spectral dataset. A publicly available near infrared (NIR) transmittance dataset and a Raman transmittance spectroscopy dataset are adopted. for evaluating the method. This dataset is about Escitalopram tablets from the pharmaceutical company H. Lundbeck A/S. The tablets are qualitatively divided into four categories according to dosage values of this pharmaceutical drug, which are 5, 10, 15, and 20 mg tablets, respectively. Classification is carried out on the four types of tablets. The instrument for collecting the Raman spectra data of each tablet is a Perkin–Elmer System 2000 NIR FT-Raman spectrometer equipped with a diode pumped Nd:YAG laser emitting 400 mW at $v_0 = 9\,394.69\,\mathrm{cm}^{-1}$ and an InGaAs detector. Its Raman wavenumber shifts range is 200–3 600 cm^{-1} with the interval of 1 cm^{-1} and the resolution of 8 cm^{-1} [55]. Please refer to the following website for further detail about the dataset: `http://www.models.life.ku.dk/Tablets`. The NIR transmittance spectra data include 310 samples with 404 variables. Approximately 80 samples belong to each class. In the experiment, nine–tenths samples of each class were randomly selected for training, and the remaining samples were used for testing. We repeated the experiment 10 times and compared the classification results with and without LK pre-processing, as shown in Table 4. Our method improves the testing accuracy by 6.78 % on this NIR transmittance spectra data.

Algorithm	Testing accuracy on NIR transmittance spectra data
MLR-elastic net	89.67
MLR-elastic net + LK (fusion kernel)	**96.45**

Table 4. Classification accuracy of the method on near infrared transmittance dataset with and without LK pre-processing

The tablets' Raman spectra data include 120 samples. Approximately 30 samples belong to each class. We conducted a binomial logistic regression experiment on a subset of the Raman spectra data. The first class and the fourth class were selected as the subset of the Raman spectra data, corresponding to the tablets with a dosage of 5 and 20 mg, respectively. Each of these two classes has 30 samples. We used 10 rounds of 5-fold cross-validation to further illustrate the effectiveness of our method. In the experiment, for each class, 25 samples were chosen for training, and the remaining 5 samples were selected for evaluation. Binomial logistic regression based on the generalized additive model was adopted for fitting the regularization path of the data. We compared the classification results based on the statistical learning model GAMSEL and SVM with or without LK preprocessing, as shown in Table 5. The classification accuracy was improved by GAMSEL; therefore, exploring the LK feature space with the generalized additive model is more effective than using SVM. The experiment shows that exploration of the LK feature space based on a statistical learning model is effective.

Algorithm	Testing accuracy on one subset of Raman spectra data
SVM	96
SVM + LK	94.83
GAMSEL	**100**
GAMSEL + LK	**100**

Table 5. 10 rounds of 5-fold cross-validation classification accuracy of SVM and GAMSEL on one subset of Raman spectra data with LK pre-processing

6 ANALYSIS AND CONCLUSION

This paper proposes a novel method that integrates LK pre-processing into a statistical learning model for classification. The Gaussian kernel and the fusion of Gaussian and cosine kernels are adopted for linearized kernel pre-processing. For multinomial logistic regression, we use the elastic net's grouping effect to find the grouping features in the high dimension features space. Experimental results on the Extended Yale-B database and AR-Face database demonstrated the good performance of the multinomial regression with elastic net methods based on the statistical learning model. This method can overcome the restriction of a small number of samples. The elastic net penalty can guarantee the robustness of the least square solution and strengthen the sparseness of the solution vector so that the model is more parsimonious and its accuracy is greatly improved. The elastic net can also deal with high dimensional and small sample data effectively, and the model can obtain a good trade-off between sparsity and prediction accuracy. A relatively high accuracy model can be established with the combination of LK pre-processing and elastic net based on statistical learning.

We also examined the classification of the different dosages of the active substance in Escitalopramtablets using Raman transmittance spectroscopy. This method also makes progress in classification accuracy with LK pre-processing on spectral data. In this study, the experiment was further carried out on a subset of the Raman transmittance spectroscopy dataset as a binomial logistic regression problem. We adopt the GAMSEL R package to capture the generalized additive model within the LK pre-processing feature spaces. The GAMSEL still can be used to fit the nonlinearity on the linearized kernel feature space. But the relatively small number of samples leads to kind of over fitting in it. The experiment shows that it is effective to explore the linearized kernel feature space based on the statistical learning model. We found that linearized kernel pre-processing is an effective descending dimension method and the fusion of Gaussian and cosine kernels for linearized kernel pre-processing with improved accuracy compared with a single Gaussian kernel. It provides us a new perspective to explore the low-dimensional space embedded with LK pre-processing in the high dimension data. Nowadays deep convolution neural network has achieved satisfactory performance in many fields. One of our future directions is to use the proposed model to explore the deep CNN feature space.

We will also focus on combining our method with Chemometric for spectral data analysis.

Acknowledgements

This work was supported by the Youth Science Foundation Project of Zhejiang Natural Science Foundation: Study on Grouping Characteristics of High Dimensional Data in Spectral Data Analysis (LQ19F020006).

REFERENCES

[1] Hastie, T.—Tibshirani, R.—Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, USA, 2001.

[2] James, G.—Witten, D.—Hastie, T.—Tibshirani, R.: An Introduction to Statistical Learning: With Applications in R. Springer, New York, USA, 2013.

[3] Tibshirani, R.: Regression Shrinkage Selection via the LASSO. Journal of the Royal Statistical Society, Series B (Methodological), Vol. 58, 1996, No. 1, pp. 267–288.

[4] Donoho, D. L.: For Most Large Underdetermined Systems of Equations, the Minimal $l(1)$-Norm Near-Solution Approximates the Sparsest Near-Solution. Communications on Pure and Applied Mathematics, Vol. 59, 2006, No. 7, pp. 907–934, doi: 10.1002/cpa.20131.

[5] Donoho, D. L.—Elad, M.: Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via $\ell 1$ Minimization. Proceedings of the National Academy of Sciences of the United States of America (PNAS), Vol. 100, 2003, No. 5, pp. 2197–2202, doi: 10.1073/pnas.0437847100.

[6] Donoho, D. L.—Huo, X.: Uncertainty Principles and Ideal Atomic Decomposition. IEEE Transactions on Information Theory, Vol. 47, 2001, No. 7, pp. 2845–2862, doi: 10.1109/18.959265.

[7] Chen, S. S.—Donoho, D. L.—Saunders, M. A.: Atomic Decomposition by Basis Pursuit. SIAM Review, Vol. 43, 2001, No. 1, pp. 129–159, doi: 10.1137/S003614450037906X.

[8] Efron, B.—Hastie, T.—Johnstone, I.—Tibshirani, R.: Least Angle Regression. The Annals of Statistics, Vol. 32, 2004, No. 2, pp. 407–499, doi: 10.1214/009053604000000067.

[9] Candes, E. J.—Romberg, J.—Tao, T.: Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information. IEEE Transactions on Information Theory, Vol. 52, 2006, No. 2, pp. 489–509, doi: 10.1109/TIT.2005.862083.

[10] Candes, E. J.—Tao, T.: Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies? IEEE Transactions on Information Theory, Vol. 52, 2006, No. 12, pp. 5406–5425, doi: 10.1109/TIT.2006.885507.

[11] Donoho, D. L.: Compressed Sensing. IEEE Transactions on Information Theory, Vol. 52, 2006, No. 4, pp. 1289–1306, doi: 10.1109/TIT.2006.871582.

[12] Donoho, D. L.—Elad, M.—Temlyakov, V. N.: Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. IEEE Transactions on Information Theory, Vol. 52, 2006, No. 1, pp. 6–18, doi: 10.1109/TIT.2005.860430.

[13] Tsaig, Y.—Donoho, D. L.: Extensions of Compressed Sensing. Signal Processing, Vol. 86, 2006, No. 3, pp. 549–571, doi: 10.1016/j.sigpro.2005.05.029.

[14] Zou, H.—Hastie, T.: Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B (Statistical Methodology), Vol. 67, 2005, No. 2, pp. 301–320, doi: 10.1111/j.1467-9868.2005.00503.x.

[15] Chouldechova, A.—Hastie, T.: Generalized Additive Model Selection. Statistics, 2015, arXiv: 1506.03850v2.

[16] Powers, S.—Hastie, T.—Tibshirani, R.: Nuclear Penalized Multinomial Regression with an Application to Predicting at Bat Outcomes in Baseball. Statistical Modelling, Vol. 18, 2018, No. 5-6, pp. 388–410, doi: 10.1177/1471082X18777669.

[17] Zhang, Z.—Lai, Z.—Xu, Y.—Shao, L.—Wu, J.—Xie, G. S.: Discriminative Elastic-Net Regularized Linear Regression. IEEE Transactions on Image Processing, Vol. 26, 2017, No. 3, pp. 1466–1481, doi: 10.1109/TIP.2017.2651396.

[18] Li, G. Z.—Wang, S. T.: Face Recognition Based on Sparse Representation and Elastic Network. Journal of Computer Applications, Vol. 37, 2017, No. 3, pp. 901–905, doi: 10.11772/j.issn.1001-9081.2017.03.901.

[19] Jiang, Z. L.—Zhe, L.—Davis, L. S.: Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, 2013, No. 11, pp. 2651–2664, doi: 10.1109/TPAMI.2013.88.

[20] Naseem, I.—Togneri, R.—Bennamoun, M.: Linear Regression for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, 2010, No. 11, pp. 2106–2112, doi: 10.1109/TPAMI.2010.128.

[21] Wright, J.—Yang, A. Y.—Ganesh, A.—Sastry, S. S.—Ma, Y.: Robust Face Recognition via Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, 2009, No. 2, pp. 210–227, doi: 10.1109/TPAMI.2008.79.

[22] Hassaballah, M.—Aly, S.: Face Recognition: Challenges, Achievements, and Future Directions. IET Computer Vision, Vol. 9, 2015, No. 4, pp. 614–626, doi: 10.1049/iet-cvi.2014.0084.

[23] Parkhi, O. M.—Vedaldi, A.—Zisserman, A.: Deep Face Recognition. Proceedings of the British Machine Vision Conference (BMVC), 2015, doi: 10.5244/c.29.41.

[24] Gao, Y.—Ma, J.—Yuille. A. L.: Semi-Supervised Sparse Representation Based Classification for Face Recognition with Insufficient Labeled Samples. IEEE Transactions on Image Processing, Vol. 26, 2017, No. 5, pp. 2545–2560, doi: 10.1109/TIP.2017.2675341.

[25] Golts, A.—Elad, M.: Linearized Kernel Dictionary Learning. IEEE Journal of Selected Topics in Signal Processing, Vol. 10, 2016, No. 4, pp. 726–739, doi: 10.1109/JSTSP.2016.2555241.

[26] Levy, S.—Fullagar, P. K.: Reconstruction of a Sparse Spike Train from a Portion of Its Spectrum and Application to High-Resolution Deconvolution. Geophysics, Vol. 46, 1981, No. 9, pp. 1235–1243, doi: 10.1190/1.1441261.

[27] Kim, S. J.—Koh, K.—Lustig, M.—Boyd, S.—Gorinevsky, D.: An Interior-Point Method for Large-Scale $l(1)$-Regularized Least Squares. IEEE Journal of Selected Topics in Signal Processing, Vol. 1, 2007, No. 4, pp. 606–617, doi: 10.1109/JSTSP.2007.910971.

[28] Zou, H.: The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association, Vol. 101, 2006, No. 476, pp. 1418–1429, doi: 10.1198/016214506000000735.

[29] Boser, B. E.—Guyon, I. M.—Vapnik, V. N.: A Training Algorithm for Optimal Margin Classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT '92), 1992, pp. 144–152, doi: 10.1145/130385.130401.

[30] Khan, S.—Naseem, I.—Togneri, R.—Bennamoun, M.: A Novel Adaptive Kernel for the RBF Neural Networks. Circuits, Systems, and Signal Processing, Vol. 36, 2017, No. 4, pp. 1639–1653, doi: 10.1007/s00034-016-0375-7.

[31] Pillonetto, G.—Dinuzzo, F.—Chen, T.—De Nicolao, G.—Ljung, L.: Kernel Methods in System Identification, Machine Learning and Function Estimation: A Survey. Automatica, Vol. 50, 2014, No. 3, pp. 657–682, doi: 10.1016/j.automatica.2014.01.001.

[32] Zhang, L.—Zhou, W. D.—Chang, P. C.—Liu, J.—Yan, Z.—Wang, T.—Li, F. Z.: Kernel Sparse Representation-Based Classifier. IEEE Transactions on Signal Processing, Vol. 60, 2012, No. 4, pp. 1684–1695, doi: 10.1109/TSP.2011.2179539.

[33] Zhan, Y.—Dai, S.—Mao, Q.—Liu, L.—Sheng, W.: A Video Semantic Analysis Method Based on Kernel Discriminative Sparse Representation and Weighted KNN. The Computer Journal, Vol. 58, 2015, No. 6, pp. 1360–1372, doi: 10.1093/comjnl/bxu121.

[34] Yang, Y.—Li, D.—Duan, Z.: Chinese Vehicle License Plate Recognition Using Kernel-Based Extreme Learning Machine with Deep Convolutional Features. IET Intelligent Transport Systems, Vol. 12, 2018, No. 3, pp. 213–219, doi: 10.1049/iet-its.2017.0136.

[35] Van Nguyen, H.—Patel, V. M.—Nasrabadi, N. M.—Chellappa, R.: Design of Non-Linear Kernel Dictionaries for Object Recognition. IEEE Transactions on Image Processing, Vol. 22, 2013, No. 12, pp. 5123–5135, doi: 10.1109/TIP.2013.2282078.

[36] Van Nguyen, H.—Patel, V. M.—Nasrabadi, N. M.—Chellappa, R.: Kernel Dictionary Learning. Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 2021–2024, doi: 10.1109/ICASSP.2012.6288305.

[37] Yin, J.—Liu, Z.—Jin, Z.—Yang, W.: Kernel Sparse Representation Based Classification. Neurocomputing, Vol. 77, 2012, No. 1, pp. 120–128, doi: 10.1016/j.neucom.2011.08.018.

[38] Gao, S. H.—Tsang, I. W. H.—Chia, L. T.: Kernel Sparse Representation for Image Classification and Face Recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (Eds.): Computer Vision – ECCV 2010. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 6314, 2010, pp. 1–14, doi: 10.1007/978-3-642-15561-1_1.

[39] Chen, Z. H.—Zuo, W. M.—Hu, Q. H.—Lin, L.: Kernel Sparse Representation for Time Series Classification. Information Sciences, Vol. 292, 2015, pp. 15–26, doi: 10.1016/j.ins.2014.08.066.

[40] Gangeh, M. J.—Ghodsi, A.—Kamel, M. S.: Kernelized Supervised Dictionary Learning. IEEE Transactions on Signal Processing, Vol. 61, 2013, No. 19, pp. 4753–4767, doi: 10.1109/TSP.2013.2274276.

[41] Harandi, M. T.—Sanderson, C.—Hartley, R.—Lovell, B. C.: Sparse Coding and Dictionary Learning for Symmetric Positive Definite Matrices: A Kernel Approach. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.): Computer Vision – ECCV 2012. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7573, 2012, pp. 216–229, doi: 10.1007/978-3-642-33709-3_16.

[42] Shrivastava, A.—Nguyen, H. V.—Patel, V. M.—Chellappa, R.: Design of Non-Linear Discriminative Dictionaries for Image Classification. In: Lee, K. M., Matsushita, Y., Rehg, J. M., Hu, Z. (Eds.): Computer Vision – ACCV 2012. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7724, 2012, pp. 660–674, doi: 10.1007/978-3-642-37331-2_50.

[43] Bryt, O.—Elad, M.: Compression of Facial Images Using the K-SVD Algorithm. Journal of Visual Communication and Image Representation, Vol. 19, 2008, No. 4, pp. 270–282, doi: 10.1016/j.jvcir.2008.03.001.

[44] Zepeda, J.—Guillemot, C.—Kijak, E.: Image Compression Using Sparse Representations and the Iteration-Tuned and Aligned Dictionary. IEEE Journal of Selected Topics in Signal Processing, Vol. 5, 2011, No. 5, pp. 1061–1073, doi: 10.1109/JSTSP.2011.2135332.

[45] Elad, M.—Aharon, M.: Image Denoising via Sparse and Redundant Representations over Learned Dictionaries. IEEE Transactions on Image Processing, Vol. 15, 2006, No. 12, pp. 3736–3745, doi: 10.1109/TIP.2006.881969.

[46] Fadili, M. J.—Starck, J.-L.—Murtagh, F.: Inpainting and Zooming Using Sparse Representations. The Computer Journal, Vol. 52, 2009, No. 1, pp. 64–79, doi: 10.1093/comjnl/bxm055.

[47] Mairal, J.—Elad, M.—Sapiro, G.: Sparse Representation for Color Image Restoration. IEEE Transactions on Image Processing, Vol. 17, 2008, No. 1, pp. 53–69, doi: 10.1109/TIP.2007.911828.

[48] Engan, K.—Aase, S. O.—Hakon Husoy, J.: Method of Optimal Directions for Frame Design. Proceedings of 1999 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '99), Vol. 5, 1999, pp. 2443–2446, doi: 10.1109/ICASSP.1999.760624.

[49] Aharon, M.—Elad, M.—Bruckstein, A.: K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. IEEE Transactions on Signal Processing, Vol. 54, 2006, No. 11, pp. 4311–4322, doi: 10.1109/TSP.2006.881199.

[50] Jiang, Z.—Lin, Z.—Davis, L. S.: Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1697–1704, doi: 10.1109/CVPR.2011.5995354.

[51] Friedman, J.—Hastie, T.—Tibshirani, R. et al.: Lasso and Elastic-Net Regularized Generalized Linear Models. 2017, available at: http://web.stanford.edu/~hastie/glmnet/glmnet_beta.html.

[52] Martinez, A.—Benavente, R.: The AR Face Database. CVC Technical Report No. 24, 1998, available at: http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html.

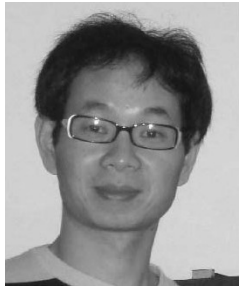
[53] Georghiades, A. S.—Belhumeur, P. N.—Kriegman, D. J.: From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, 2001, No. 6, pp. 643–660, doi: 10.1109/34.927464.

[54] Chang, C. C.—Lin, C. J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2, 2011, No. 3, Art. No. 27, pp. 1–27, doi: 10.1145/1961189.1961199.

[55] Dyrby, M.—Engelsen, S. B.—Nørgaard, L.—Bruhn, M.—Lundsberg-Nielsen, L.: Chemometric Quantitation of the Active Substance (Containing $C \equiv N$) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra. Applied Spectroscopy, Vol. 56, 2002, No. 5, pp. 579–585, doi: 10.1366/0003702021955358.

Xiaochun Guan received her B.Sc. and M.Sc. degrees in measurement technology and automation apparatus from the University of Shanghai for Science and Technology, Shanghai, in 2002 and 2005, respectively. She joined Wenzhou University in 2005 where she is currently Associate Professor in the College of Electrical and Electronic Engineering. She is now pursuing the Ph.D. degree in the School of Computer Science and Technology, Zhejiang University of Technology. Her research interests include end-side deep neural network deployment, machine learning, sparse representation and statistical learning.

Jianhua Zhang received his Ph.D. degree from the University of Hamburg, Hamburg, Germany in 2012. He is currently Professor with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China. His current research interests include SLAM, 3D vision, reinforcement learning, and machine vision.



Shengyong Chen received his Ph.D. degree in robot vision from the City University of Hong Kong, Hong Kong, in 2003. He is currently Professor with Tianjin University of Technology, China. He received a fellowship from the Alexander von Humboldt Foundation of Germany and worked with the University of Hamburg, Hamburg, Germany, from 2006 to 2007. He has authored over 100 scientific papers in international journals and is an inventor of over 100 patents. His research interests include computer vision, robotics, and image analysis. He is Fellow of IET and Senior Member of CCF. He was the recipient of the National Outstanding Youth Foundation Award of China in 2013.

Computing and Informatics, Vol. 40, 2021, 318–340, doi: 10.31577/cai_2021_2_318

CLUSTERING AND BOOTSTRAPPING BASED FRAMEWORK FOR NEWS KNOWLEDGE BASE COMPLETION

K. Srinivasa, P. Santhi Thilagam

National Institute of Technology Karnataka
Department of Computer Science and Engineering
NH 66, Srinivasnagar, Surathkal, Mangalore
Karnataka – 575 025, India
e-mail: srinivas.karur@gmail.com, santhi@nitk.edu.in

Abstract. Extracting the facts, namely entities and relations, from unstructured sources is an essential step in any knowledge base construction. At the same time, it is also necessary to ensure the completeness of the knowledge base by incrementally extracting the new facts from various sources. To date, the knowledge base completion is studied as a problem of knowledge refinement where the missing facts are inferred by reasoning about the information already present in the knowledge base. However, facts missed while extracting the information from multilingual sources are ignored. Hence, this work proposed a generic framework for knowledge base completion to enrich a knowledge base of crime-related facts extracted from online news articles in the English language, with the facts extracted from low resourced Indian language Hindi news articles. Using the framework, information from any low-resourced language news articles can be extracted without using language-specific tools like POS tags and using an appropriate machine translation tool. To achieve this, a clustering algorithm is proposed, which explores the redundancy among the bilingual collection of news articles by representing the clusters with knowledge base facts unlike the existing Bag of Words representation. From each cluster, the facts extracted from English language articles are bootstrapped to extract the facts from comparable Hindi language articles. This way of bootstrapping within the cluster helps to identify the sentences from a low-resourced language that are enriched with new information related to the facts extracted from a high-resourced language like English. The empirical result shows that the proposed clustering algorithm produced more accurate and high-quality clusters for monolingual and cross-lingual facts, respectively. Experiments also proved that the proposed framework achieves a high recall rate in extracting the new facts from Hindi news articles.

Keywords: Knowledge base completion, natural language processing, information extraction, triples, bootstrap, cluster

1 INTRODUCTION

Knowledge Bases (KBs) contain a huge collection of information in the form of entities and relations extracted from structured and unstructured sources. Such information is stored as triples in machine-readable form like $\langle e_1 - R - e_2 \rangle$ called as facts. Knowledge Base Completion (KBC) is a long-standing problem in the area of knowledge management that involves the task of identifying the missing facts from the KBs. To date, the KBC problem is studied as a Knowledge Graph (KG) refinement problem where the missing facts are inferred from the existing facts in the KB [1]. For example, $works\_for$ relation can be inferred from the fact $\langle Person\_X - CEO\_of - Comapny\_Y \rangle$ by applying the appropriate inferring techniques. The focus of such techniques is on improving the inferring accuracy so that more number of appropriate hidden facts are extracted from the KB. Hence, these techniques ensure the identification of new facts that are not explicitly stored but are hidden in the KB. However, it is also necessary to ensure the completeness of the KB by identifying the missing facts while extracting the information from multiple sources.

In the era of a multilingual environment where the information is scattered across the web in multiple languages, most of the facts are redundant but are enriched with some new facts. For instance, the online news articles from different sources with various native languages within the same window of published dates, include information related to almost similar facts. However, each source may be enriched with some new facts about an event, and failing in identifying such facts is censorious for applications like crime prevention and monitoring. For instance, "Gurgaon police arrests key Lawrence Bishnoi gang member from Hyderabad" and लॉरेंस बिश्नोई गैंग का क़ुख्यात अपराधी संपत नेहरा हैदराबाद से गिरफ्तार, कई राज्यों में था इसका आतंक shows two sample headlines from English and Hindi news articles, respectively [35, 36]. Even though both the headlines contain information about the same event, the info who was arrested is missing from the Hindi language headline. Hence, for applications that develop KB from news articles, it is not sufficient to extract the information only from English news articles to ensure the completeness of KB.

In this paper, we extended the work proposed by [2] to develop a KB called "Crime Base" which was enriched with the facts extracted from only English news articles. KB so developed was proved to be incomplete by manually cross verifying the related bilingual English-Hindi language articles. However, the task of grouping the related articles across the languages and extracting the facts from articles in Indian languages like Hindi needs language-specific tools like Parts of Speech (PoS) tagger. However, these tools are either unavailable or not accurate enough to be

used for low-resourced languages like Indian languages. Although the grouping of articles can be achieved using document clustering techniques [3], the traditional way of representing the clusters using Bag of Words is not appropriate due to its inability to represent the cluster semantically. The semantic way of representing the cluster is highly essential as the quality of clusters formed depends on how semantically a cluster is represented. Besides, it is also essential to cluster the news articles incrementally as they are published daily and are to be treated as data streams. Hence, it is most appropriate to adopt the methods used to cluster the data streams to cluster the news articles [4]. Nevertheless, these methods are to be modified to cluster the articles across the languages. Even though open information extraction (OIE) tools like ArgOE are best suited to extract the information from multiple languages, they are developed to be used with foreign languages like English, Spanish, and Portuguese [5]. A straightforward way to solve the problem which is adopted by the existing works is to translate the entire document written in a target language like Hindi to source language English [6]. Such methods are time expensive as all the texts available in a target language do not contribute to the extraction of facts. In contrast, translation of only named entities like name of the PERSON, ORGANIZATION, and LOCATION and their relationships is adequate to compare to the translation of the whole document from KB perspective. In the news domain where the corpus is a collection of multilingual news articles, the extraction of facts from such a corpus is possible using supervised machine learning techniques. However, these techniques require a large collection of sententially aligned parallel data from different language articles to train the system which is very expensive.

The news articles are the kind of comparable corpora where the multilingual articles within a window of published dates are usually redundant. Extracting information from such a corpus can be achieved by clustering the articles across the languages based on their topical similarity. Once the topically similar articles are grouped, the source language facts can be bootstrapped to extract the facts from target language articles. Such a bootstrapped way of extraction limits the translation only to named entities and their relationships and hence reduces the time required to translate the entire article from the target language to the source language. Accordingly, this work proposes a bootstrapping-based KBC framework that can be adaptable to any domain and language using the appropriate machine translation tool. Moreover, the framework also helps to extract the facts from target language articles without using language-specific tools like POS tags which are most necessary for Indian languages. The overall architecture of the proposed work is shown in Figure 1.

The primary contributions of this paper are as follows:

1. Proposed an algorithm for grouping the related news articles in a bilingual corpus.
2. Proposed a bootstrapping-based method to extract the facts from target language news articles using facts extracted from source language news articles with minimum translation efforts.

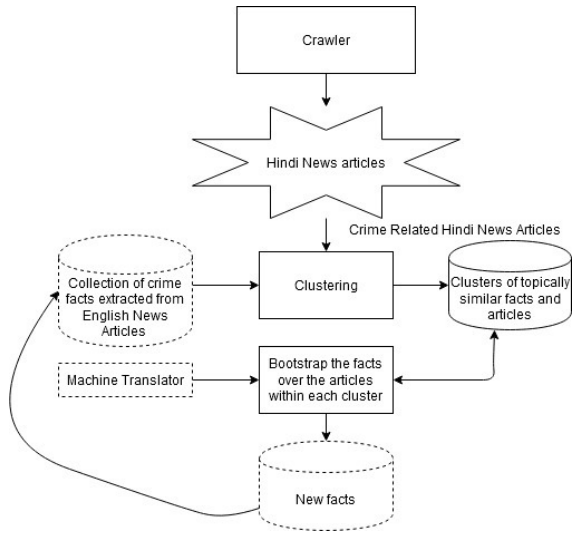


Figure 1. High level architecture of the proposed framework

The rest of the paper is organized as follows. Section 2 discusses the background and related work. Section 3 describes the problem along with the research objectives. The proposed methodology is explained in Section 4. Section 5 presents the results and analysis of the experiments conducted. Conclusion and future works are explained in Section 6.

2 BACKGROUND AND RELATED WORK

This section aims to provide an overview of the work in the domain of knowledge base construction with an emphasis on knowledge base completion. Knowledge base construction is the main vision of the semantic web to create a shared repository of KB in machine-readable form. Even though the problem of knowledge base construction is studied for decades, the knowledge base completion is studied only as a knowledge base refinement problem [1, 29, 30]. The knowledge base refinement methods try to complete the knowledge base by considering the facts internal to the knowledge base by inferring the new facts hidden inside the given knowledge base. However, these methods do not cover the external facts, i.e. facts extracted from multiple sources while constructing the knowledge base. In this perspective, the knowledge base completion can be treated as a problem of information extraction and integration, where the final knowledge base must be enriched with all the new facts extracted from multiple sources.

The existing works to extract the information from multiple sources are categorized as monolingual and multilingual based on the languages they support. Systems that extract the knowledge from sources in a single language are considered as mono-

lingual systems and most of the systems extract the knowledge only from sources in English language [7]. The authors in [8] extract the crime-related information from English news articles. [9] developed a system to construct a sports knowledge base by extracting the information from the FIFA website in the English language. [31] proposed an NLP and machine learning-based method for extraction of economic events and constructed a financial knowledge base. A T2KG system is proposed by [32] to construct a knowledge graph from unstructured text. An artist's knowledge base is constructed by [10] by extracting the information from the related websites in English. Few works also contributed to creating a knowledge base of events collected from news articles. For instance, Storybase was the knowledge base created by extracting the information from daily web news and Wikipedia current events [11]. In [12] authors extract the facts from French-language news articles. Knowledge Vault [13] is a probabilistic system that combines extractions from multiple sources like text and tabular data. PRISMATIC is a large-scale lexicalized relation resource that automatically extracts the knowledge from articles in English language [14]. Even though a lot of works are carried out to create a knowledge base by extracting the information from multiple sources, these works do not emphasize completing the knowledge base from multilingual sources.

In contrast to monolingual systems, systems that extract the knowledge from sources of different languages are considered as multilingual systems. Most of the multilingual systems consider the sources in foreign languages [15, 6, 16, 17, 18, 5] and only [19] extracts the Indian language Hindi along with foreign languages. However, these systems are based on either using language-specific processing tools or translating the entire documents into English.

Apart from the individual efforts in generating KBs like [20], few integrative projects which involve a community of users in creating KBs by extracting and updating the facts from crowd-sourced data like Wikipedia also emerged. To name few, Yago [21] is a KB created automatically from Wikipedia, WordNet and Geonames. DBpedia [22] exploits both free text as well as semi-structured data like infoboxes from Wikipedia to create the KB. BabelNet [23], the largest repository of multilingual words and senses, integrates Wikipedia and WordNet for creating the KB. Wikidata [24] is a KB enriched with facts extracted only from Wikipedia. Even though, the KB generated by these systems are well structured to support the web of linked data, the facts covered and validated by these systems are limited to Wikipedia.

There are some open-source tools like FRED and FOX [25] that generate structured knowledge graphs from unstructured texts. FRED is a powerful tool that extracts the knowledge from 48 languages. However, the capability of the tool is limited to only extraction and lack in integrating the knowledge extracted from multiple languages.

In addition to the efforts to generate the knowledge base, several studies attempted to develop knowledge base completion models using cross-lingual projection of knowledge. However, these models require the presence of a knowledge base for both the source and target language. Using the knowledge bases for both the

languages, the facts from the source language are projected with the target language for knowledge base completion. For instance, [26] and [27] developed a knowledge base completion model based on vector representation by representing the concepts in multiple languages in a unified vector space. But these models are not applicable in the absence of a knowledge base for a target language.

Table 1 shows the consolidated view of the features supported by existing works on knowledge base construction. The table lists both monolingual as well as multilingual systems. From the table, it is clear that none of the systems supports knowledge base completion by identifying the missing facts while extracting the information from multiple sources. Specific to the news domain, the existing systems considered only English news articles and ignored the facts available in other language news articles. Moreover, exploiting the redundancy that exists among the news articles for the identification of new facts is also underexplored which is observed from the last column in the table.

3 PROBLEM DESCRIPTION AND RESEARCH OBJECTIVES

Given a set of bilingual news articles from resource-rich Source Language (SL) like English and resource deficit Target Language (TL) like Hindi. This paper aimed at developing a framework for KBC to extract the facts from TL news articles so that the KB created using SL news articles is enriched with new facts available in TL news articles. This is to be achieved by exploiting the redundancies available from SL and TL news articles and without using the language-specific tools for TL news articles.

To address the problem described above, the following research objectives or tasks are set:

Task-1: To propose an algorithm for grouping the related articles from SL and TL using clustering.

Task-2: To propose a method to extract the facts from TL news articles using facts related to SL news articles as a bootstrapping data set and an appropriate machine translation tool.

4 METHODOLOGY

The detailed architecture of the proposed framework is shown in Figure 2. The proposed framework performs bootstrapping at multiple levels to extract the new facts from Hindi news articles using the triples extracted from the related English news articles. The framework consists of two main stages, namely clustering and extraction, to solve respective tasks mentioned in Section 3 and are discussed in the following sections.

System	Method of IE	Domain	Source of Extraction	Language	Integration	Support for KBC	Exploitation of Redundancy
CrimeProfiler ([8])	Stanford NER, Semi-supervised	Crime	News articles	English			
SOBA ([9])	SProUT, Rule based	Sports	FIFA website	English	✓		
Music Knowledge Base ([20])	Rule based, Un-supervised	Music	songfacts.com	English	✓		
Artequakt ([10])	Wornet, GATE, Ontology	Artists	Web	English	✓		
PRISMATIC ([14])	Frame+Un-supervised	Open	Web	English			
NELL ([15])	Semi-supervised	Open	Web	English	✓		
RdfLiveNews ([16])	Un-supervised, Supervised	Open	Web	English	✓		
Storybase ([11])	Un-supervised	Open News	Web	English			
Stern et al. ([12])	Entity based	Open News	Web	French			
Knowledge Vault ([13])	Distant Supervision	Open	Web	English	✓		
Finance KB ([31])	NLP, Machine Learning	Economic	News	English			
T2KG ([32])	Rule and similarity based	Open	Web	English			
Even centric knowledge grahs ([6])	Semantic Role Labelling	Open News	Wikinews, FIFA world cup, Global Automotive Industry, Airbus A380 airplanes.	English, Spanish, Italian and Dutch	✓		
BOA ([7])	Bootstrapping	Open	Web	English, German	✓		
New/s/leak ([17])	Using Polyglot tool	Open News	Web	40			
ZENON ([18])	GATE, Rule Based	Crime	Intelligence Reports	English, Deri	✓		
Crime base ([2])(Knowledge base enriched with facts from English news articles)	Rule based	Open News	Web	English	✓		
Proposed Work(Extension of Crime base)	Rule based	Open News	Web	English, Hindi	✓	✓	✓

Table 1. Existing vs. proposed system

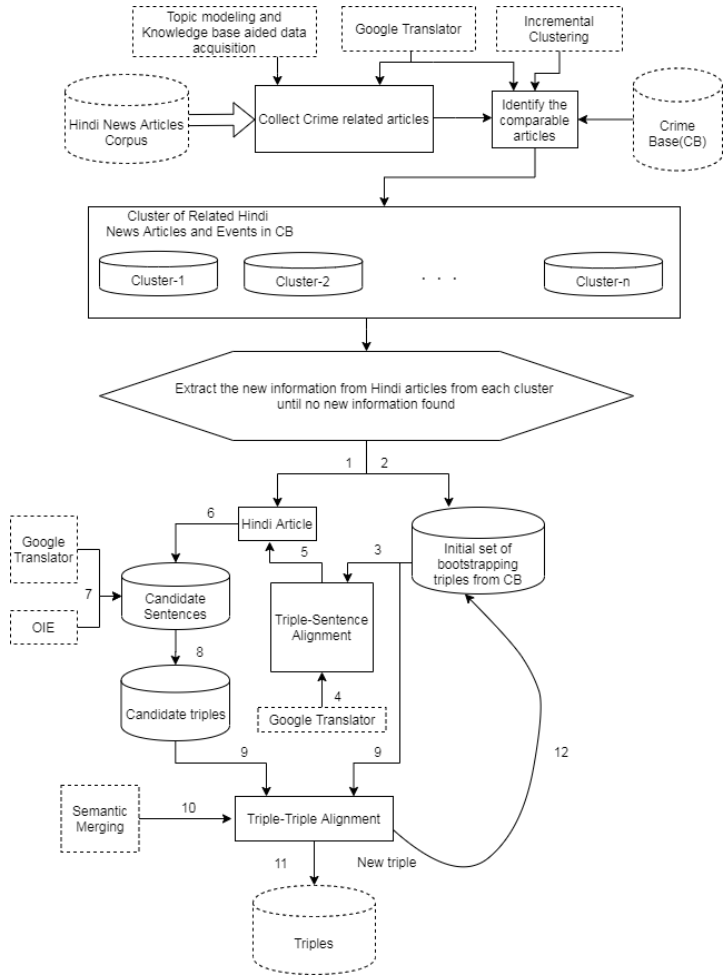


Figure 2. Detailed architecture of the proposed framework

4.1 Methodology for Task-1: Clustering

Initially, the crime-related Hindi articles are selected by applying topic modeling and knowledge base aided data acquisition method proposed by [2] over the headlines translated to English. The redundancies among the articles are exploited by identifying the comparable articles. A set of bilingual collections of articles is said to be comparable if they are related either topically or sententially. Topically related articles are contextually similar articles that discuss the same topic and are said to be semantically related. Whereas, sententially related articles are almost bilingual translations of each other and are said to be semantically similar. Hence topically re-

lated articles are enriched with more new information compare to sententially related articles. The proposed work identifies the comparable articles using the semantic merging procedure mentioned in [2]. As the news articles are published daily, the articles are considered as data streams and an incremental nearest neighborhood algorithm for clustering data streams is adopted [4]. Here, the clustering algorithm is modified to identify the sententially and topically related articles by finding sentential and topical neighbors and is named as Sentential-Topical-Nearest-Neighborhood (STNN) algorithm which is described in Algorithm 1.

The major difficulty in clustering articles is in semantically representing the articles so that clusters of better quality can be formed. Due to the availability of a large number of terms as document features, the Bag of Words way of representing the documents does not capture the semantics hidden in the sentences. To improve the semantics, the articles are represented as KB facts in the form of triples extracted over the headlines. Accordingly, the headlines from Hindi news articles are translated to English, and facts from the translated headlines are extracted using the method proposed in [2]. When a stream of facts from English and Hindi news articles comes in, we divide them into various windows based on their date of publication. Now, events in the first window are clustered using neighborhood-based clustering. The similarity between each of the elements in the first window is calculated using contextual as well as semantic similarity measures. The significance of using both the similarity measures is empirically proved and can be found in [2]. Two elements are considered to be topically neighbors if their contextual similarity is greater than a threshold value. Such neighbors are also checked for their semantic similarity. If the semantic similarity is greater than their contextual similarity score, they form a separate cluster and will be added to the set of sententially similar clusters. Otherwise, they will be added to the set of topically similar clusters. If the contextual similarity score for any two elements is less than the threshold, the elements are independent and form two separate clusters. To represent a cluster, we find the medoid of each cluster, where the medoid is an element that has the maximum similarity with all other elements in the cluster. This limits further comparison between the medoids rather than with all the elements in the cluster. A similar method is followed to find the clusters for other windows. For each new cluster, we find from the former clusters the most similar cluster to them by calculating the similarity of the medoid event of the former clusters and the medoid of the new cluster. Based on their similarity, two clusters are merged and the medoid will be updated.

4.2 Methodology for Task-2: Extraction

In this work, we propose a method to identify and extract the new facts from a target language news article like Hindi using the facts extracted from related English news articles. This is achieved by bootstrapping the triples extracted from English news articles to identify the presence of related triples from comparable Hindi news articles. The proposed extraction method constitutes two steps, namely:

Algorithm 1: STNN Algorithm

Input: $E = \{F_{E_1}, F_{E_2}, \ldots, F_{E_m}\}$:Set of m crime facts extracted from English news articles and $H_H = \{H_1, H_2, \ldots, H_n\}$:Set of n headlines extracted from Hindi language news articles

Result: Set of n_s sententially similar clusters $C_s = \{C_1, C_2, \ldots, C_{n_s}\}$, Set of n_t topically similar clusters $C_t = \{C_1, C_2, \ldots, C_{n_t}\}$

1 Translate the headlines in Hindi to English and the set of translated headlines be $H_{H_T} = \{H_{t_1}, H_{t_2}, \ldots, H_{t_n}\}$
2 Extract the facts from H_{H_T} and let $H = \{F_{H_1}, F_{H_2}, \ldots, F_{H_k}\}$ be a set of k facts related to Hindi headlines
3 Divide the events from E and H into multiple windows $W = \{w_1, w_2, \ldots\}$ where $w_i \subseteq E \cup H$ indicates facts extracted from the articles published during i^{th} date
4 Find the neighbors and hence clusters for w_1 as follows:
5 Calculate contextual similarity CS between each new couple of elements F_{E_i} and F_{H_j}.
6 If $CS >$a threshold t_c then
7 Calculate semantic similarity SS between each couple.
8 If $SS >$ a threshold t_s then
9 the elements are sententially neighbors. Each set of neighbors represent a cluster and will be added to C_s.
10 Otherwise
11 the elements are topically neighbors. Each set of neighbors represent a cluster and will be added to C_t.
12 Otherwise
13 Add the elements to C_t as new clusters.
14 Find medoid of each cluster where, medoid is the element which has the maximum similarity with all the elements in the cluster.
15 Similarly find the neighbors and hence the clusters for the subsequent windows.
16 Calculate new clusters medoids.
17 Calculate the similarity between new medoids and medoids of old clusters.
18 If found a pair of contextually or semantically similar medoids
19 Merge the clusers.
20 Update medoid.
21 Add the merged cluster to the appropriate set.
22 Otherwise
23 Retain the clusters as it is.

1. candidate sentence identification,
2. new triple generation.

Each of the steps is explained in the following subsections.

4.2.1 Candidate Sentence Identification

From each cluster, the events related to English news articles are selected as an initial set of bootstrapping triples. Each of these triples is translated to the target language using Google translator API and used to query the Hindi articles to identify a set of sentences that are enriched with new facts and are called candidate sentences. Given a set of bootstrapped triples from English articles $B_E = \{t_{E_1}, t_{E_2}, \ldots, t_{E_n}\}$, a set of candidate sentences from Hindi articles $S = \{s_1, s_2, \ldots, s_m\}$ are obtained by aligning the sentences with the triples. Formally, a sentence s_i is said to be aligned with t_{E_j}, if an element e_k belongs to t_{E_j} is a substring of s_i. Finally, a sentence that constitutes the un-aligned part in it is selected as the candidate sentence. Otherwise it is considered as similar to t_{E_j}. Figure 3 illustrates the generation of candidate sentences with an example.

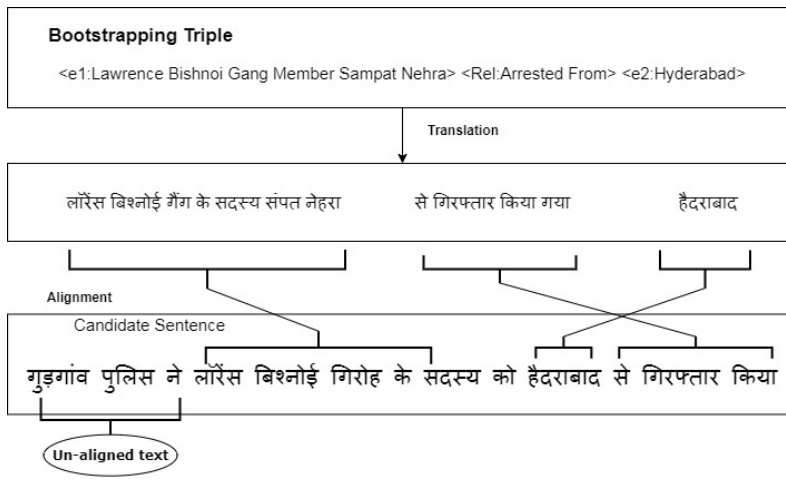


Figure 3. Candidate sentence generation

4.2.2 New Triple Generation

Once the candidate sentences are extracted, new triples are obtained in three steps, namely:

1. candidate sentence translation,
2. triple/s extraction,

3. projection of triple.

Initially, candidate sentences are translated to English language using Google API translator, and triples from each sentence are extracted using the method proposed in [2]. Triples so extracted from a candidate sentence are projected against the bootstrapped triple to identify the new triples, as shown in Figure 4 with the continuation of the example considered in Figure 3.

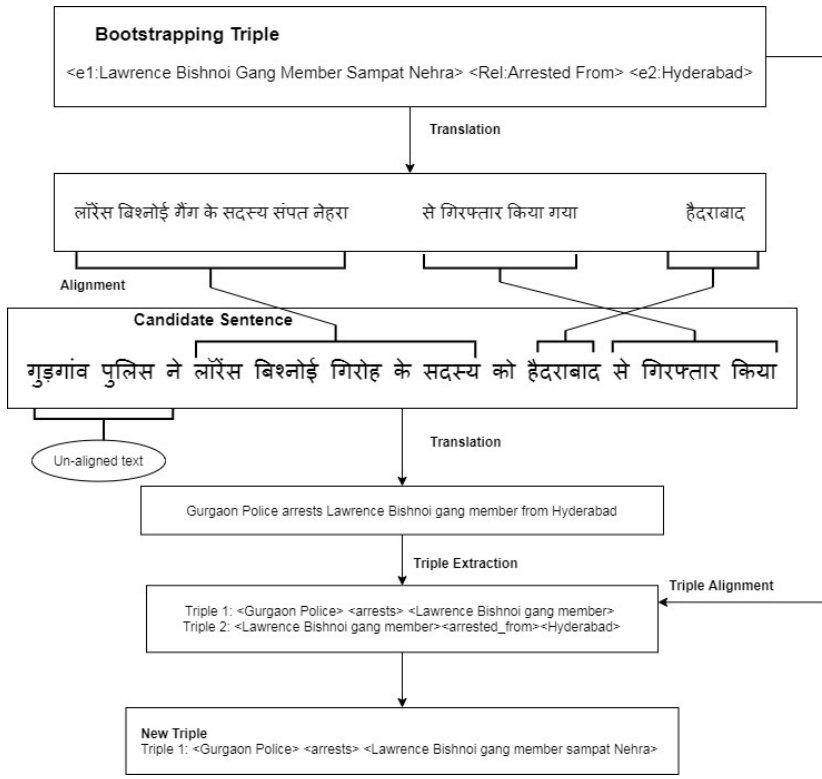


Figure 4. New triple generation

5 EXPERIMENTAL RESULTS

This work considers two prominent newspapers, namely *Indian Express* for English News articles and *Hindustan (हिंदुस्तान)*, which have articles available online. The corpus includes the data collected from *Jan* 2018 to *Jun* 2018. The following sections describe the experimental evaluation results for clustering and extraction.

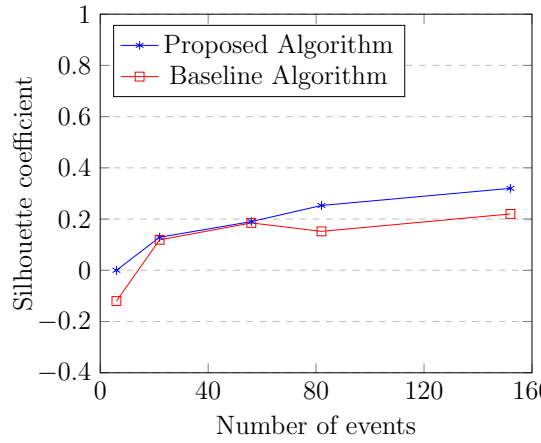


Figure 5. Bilingual evaluation of proposed clustering algorithm: Silhouette coefficient for varying number of events

5.1 Evaluation of Task-1: Clustering Algorithm

The proposed algorithm for clustering is evaluated in two phases, namely, bilingual and monolingual evaluation. In the first phase, the algorithm is evaluated for English and Hindi articles and in the second phase, the algorithm is evaluated for English articles. Due to the lack of algorithms for clustering multilingual articles, a baseline algorithm, i.e. incremental nearest neighborhood algorithm without using background KB and considering only the headlines from English and Hindi news articles, is implemented.

5.1.1 Phase-1 Evaluation: Bilingual Evaluation

The proposed algorithm is compared with the baseline algorithm in terms of the quality of clusters formed and the time taken for clustering. The clustering quality is determined using a Silhouette coefficient [28]. This is a well-known measure of internal evaluation for evaluating clusters without pre-determined labels. It measures how similar an object is to its cluster compared to other clusters. The Silhouette coefficient for i^{th} event is calculated as follows:

$$s_i = \frac{a_i - b_i}{\max(a_i, b_i)}$$

where a_i is the average similarity of the i^{th} event with all the other events in its cluster. Then for all the other clusters to which i^{th} event does not belong, we calculate the average similarity of i^{th} event to all the events in these clusters and b_i is the maximum of all these values. Figure 5 shows the silhouette coefficient obtained for proposed and baseline algorithms for varying numbers of events. We can see

that the proposed algorithm achieved a larger value of silhouette coefficient as the event size increases and hence produced a better quality of clusters. However, the Silhouette coefficient value is not very close to 1 because of many individual clusters obtained during the clustering process. These events are those which do not have similarity with any other crime events.

For instance, Figure 6 shows clustering results for 152 events extracted from 60 headlines using the proposed algorithm. There are many individual clusters and also clusters with a varying number of event elements. The medoid of each cluster can be seen highlighted in Figure 6. If we do not consider the individual clusters, then we get an average value of 0.63 and 0.45 as silhouette coefficients for proposed and baseline algorithms, respectively.

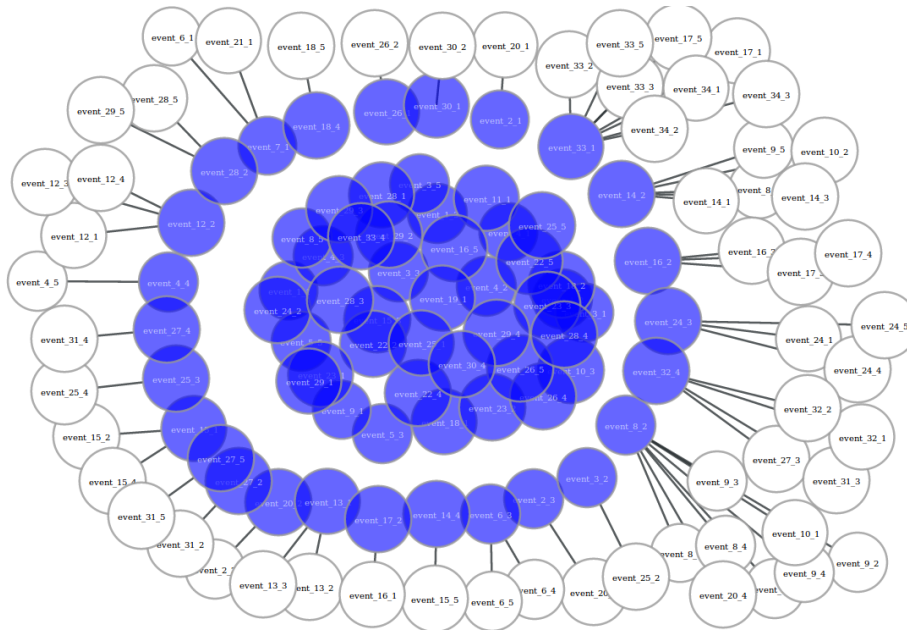


Figure 6. Visualization of cluster for 152 events

We also evaluated the quality of clusters in terms of the number of related events obtained for a given keyword. For instance, Figure 7 shows the clusters retrieved for the keyword "Navsari". It has 2 clusters associated with it with each cluster having 1 event. The keyword is directly related to both these events. The other attributes of the events are also shown. Some of the input keywords used for finding clusters of related events over a cluster in Figure 6 are shown in Table 2. From the table, it is clear that, due to the higher quality of clusters formed by the proposed algorithm, the number of related events associated with a given keyword is also significantly high.

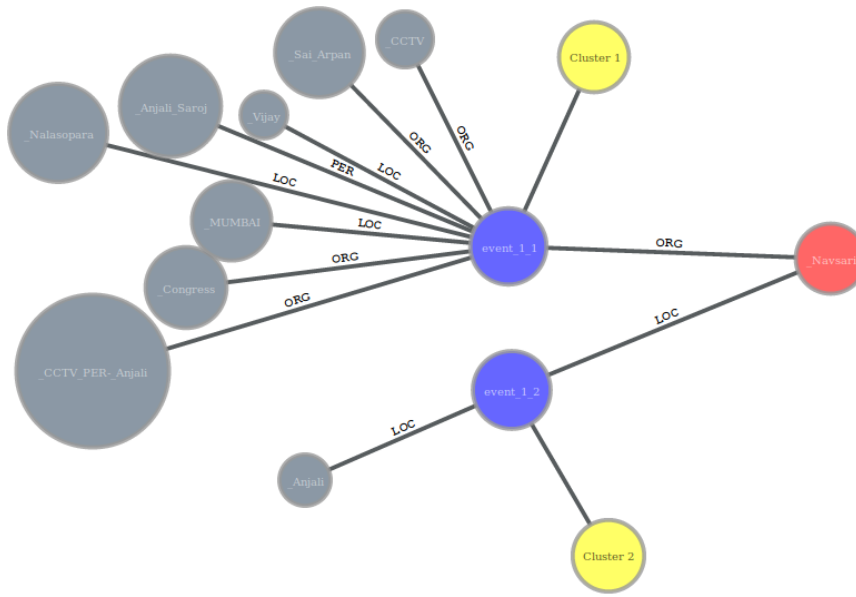


Figure 7. Cluster results for the keyword "Navsari"

Table 3 shows the clustering time taken by the proposed and baseline algorithms. From the table, it can be observed that the proposed algorithm takes more time compared to the baseline approach. This is due to the extraction of two or more triples from a single headline which is evident from the third column of the table. The time taken for machine translation and triple extraction are not considered for evaluation. However, more semantics hidden in triples compared to raw sentences

Keyword	Number of Related Events (Baseline Algorithm)	Number of Related Events (Proposed Algorithm)
Kanpur	1	3
Navsari	2	2
Venugopal	3	4
Malad	1	3
Mumbai	6	14
Bandipora	2	2
CRPF	7	9
Railway_Act	1	1
Abhijit Mukherjee	1	3
Kaluram	6	6
Congress	15	26

Table 2. Number of related events for keywords before and after clustering

produces clusters with high quality, and hence the time complexity is compromised over the cluster quality.

Features (Bag of Words in Terms of Headlines)	Clustering Time for Baseline Algorithm [s]	Features (Events in Terms of Triples)	Clustering Time for Proposed Algorithm [s]
100	92	270	98
200	194	423	222
300	298	610	343
400	372	908	402
500	536	1 022	582

Table 3. Bilingual evaluation of proposed clustering algorithm: Time taken for clustering

5.1.2 Phase-2 Evaluation: Monolingual Evaluation

Here the proposed work is evaluated by considering only the English news articles and comparing the results with two recently proposed works [33] and [34] as a baseline. Evaluation in these two works is done using Reuters and 20Newsgroup datasets. The details about the datasets can be found in [34]. [34] uses a K-means clustering algorithm with improved square root similarity measure. As an improvement to this, [33] used N-grams representation along with K-means clustering algorithm and improved square root similarity measure. The proposed algorithm is different from the baseline works by using semantically rich triples representation and a similarity measure using both contextual and semantic similarity measures proposed in [2]. Here, the experiment is conducted using 2000 samples each from Reuters and 20Newsgroup datasets over 5 newsgroups. The triples are extracted from each sample using the method proposed in [2]. To speed up the execution, a parallel version of the proposed clustering algorithm is implemented using Message Passing Interface (MPI). The triples are processed in parallel to identify the clusters.

Table 4 shows the evaluation results for the proposed and the baseline approaches. The same performance metrics as mentioned and defined in [33], i.e. accuracy and purity, are used here for evaluation. From the table, it is clear that the proposed clustering algorithm performs better than baseline methods in terms of accuracy. However, due to the generation of more individual clusters, i.e. clusters with a single element, the purity of the proposed algorithm is less compared to [33].

5.2 Evaluation of Task-2: Extraction

To evaluate the results for the proposed KBC approach, an MT-based system is implemented which is considered as a gold standard. The gold standard system reduces the problem to monolingual information extraction and integration by translating the entire articles in the target language into English. Then the facts are extracted

Methods	Datasets	Accuracy	Purity
[33]	Reuters	0.3950	0.9418
	20 Newsgroups	0.3801	0.9200
[34]	Reuters	0.2320	0.5769
	20 Newsgroups	0.1659	0.4234
Proposed Algorithm	Reuters	0.5210	0.6200
	20 Newsgroups	0.4832	0.7398

Table 4. Monolingual evaluation of proposed clustering algorithm

from translated articles and are semantically merged with facts extracted from English news articles using the methods for IE and semantic merging proposed by [2]. Hence the gold standard system is named as Machine Translation based Monolingual Knowledge Base Completion (MTML_KBC). The quality of the proposed KBC approach is measured using the standard evaluation metrics precision and recall. Precision is calculated as the ratio of the number of valid new facts extracted to the total number of new facts extracted. The recall is calculated as the ratio of the number of valid new facts extracted to the total number of valid new facts available.

Table 5 shows the results recorded for five different clusters. Figures 8 and 9 show the performance of gold standard (MTML_KBC) and proposed approach in terms of precision and recall, respectively. From the figures, it is clear that the proposed approach achieves a better recall compared to precision. This is evident from the fact that the total number of new facts extracted by the proposed approach is more due to improper projection of bootstrapping triples with the triples extracted from candidate sentences.

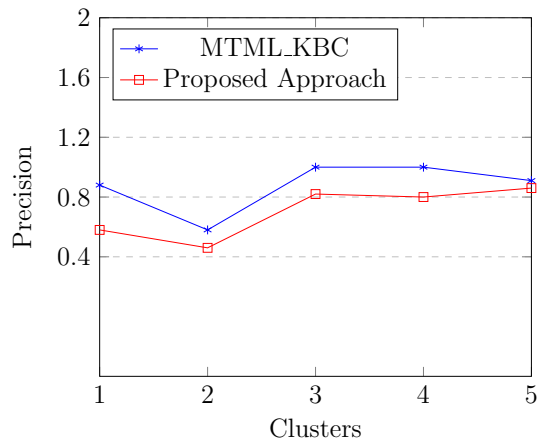


Figure 8. Precision for five clusters

Clusters	MTML_KBC			Proposed Approach		
	Total New Facts Extracted	Number of New Facts Available	Number of Valid New Facts Extracted	Total New Facts Extracted	Number of New Facts Available	Number of Valid New Facts Extracted
Cluster-1 (52 facts + 13 Hindi articles)	9	10	8	12	10	7
Cluster-2 (83 facts + 08 Hindi articles)	12	9	7	15	9	7
Cluster-3 (75 facts + 18 Hindi articles)	14	15	14	17	15	14
Cluster-4 (92 facts + 11 Hindi articles)	18	20	18	21	20	17
Cluster-5 (88 facts + 14 Hindi articles)	23	25	21	23	25	20

Table 5. Comparison of MTML_KBC and proposed approach

6 CONCLUSIONS AND FUTURE WORK

This work proposed a clustering and bootstrapping-based generic framework for knowledge base completion. Using the framework, any knowledge base created with the facts extracted from English news articles can be enriched with new facts available in low-resourced language articles without using language-specific tools. Here the experiment is conducted using the low resourced Indian language Hindi news articles. The redundancies that exist among the bilingual collection of articles are exploited by grouping the articles that are topically or sententially similar using the nearest neighborhood clustering. The proposed clustering algorithm makes use of knowledge base facts in terms of triples to represent the articles against the traditional Bag of Words representation, as the triples capture the high semantics. Empirical results show that clusters of high accuracy and quality are obtained for monolingual and bilingual facts, respectively. From each group of related articles, the facts related to English news articles are bootstrapped to extract the facts from Hindi news articles using Google translator API. This way of using the high-resource language facts as bootstrapping triples helps to extract the facts from articles related to the languages for which language processing tools like POS tags are neither available nor accurate. Experimental results for extraction show that using the framework a better recall is achieved in identifying the new facts compared to precision. A precision of high rate can be achieved by aligning the bootstrapped triples

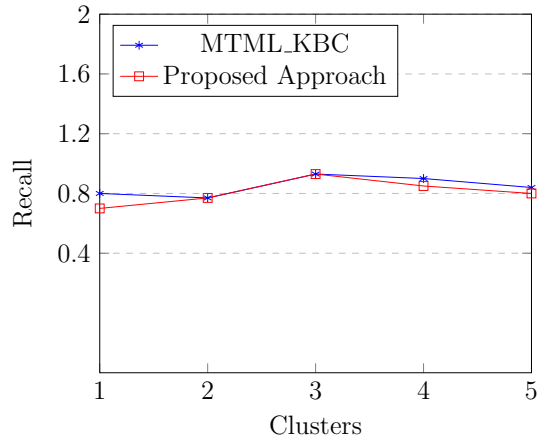


Figure 9. Recall for five clusters

with triples extracted from other languages more accurately, which will be considered in the future. In the future, the framework will be examined for other Indian languages also.

REFERENCES

[1] Paulheim, H.: Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web, Vol. 8, 2017, No. 3, pp. 489–508, doi: 10.3233/sw-160218.

[2] Srinivasa, K.—Thilagam, P. S.: Crime Base: Towards Building a Knowledge Base for Crime Entities and Their Relationships from Online News Papers. Information Processing and Management, Vol. 56, 2019, No. 6, Art. No. 102059, doi: 10.1016/j.ipm.2019.102059.

[3] Elsayed, A.—Mokhtar, H. M. O.—Ismail, O. : Ontology Based Document Clustering Using MapReduce. International Journal of Database Management Systems, Vol. 7, 2015, No. 2, doi: 10.5121/ijdms.2015.7201.

[4] Louhi, I.—Boudjeloud-Assala, L.—Tamisier, T.: Incremental Nearest Neighborhood Graph for Data Stream Clustering. 2016 International Joint Conference on Neural Networks (IJCNN'16), Vancouver, Canada, 2016, pp. 2468–2475, doi: 10.1109/ijcnn.2016.7727506.

[5] Claro, D. B.—Souza, M.—Castellã Xavier, C.—Oliveira, L.: Multilingual Open Information Extraction: Challenges and Opportunities. Information, Vol. 10, 2019, No. 7, Art. No. 228, 25 pp., doi: 10.3390/info10070228.

[6] Rospocher, M.—van Erp, M.—Vossen, P.—Fokkens, A.—Aldabe, I.—Rigau, G.—Soroa, A.—Ploeger, T.—Bogaard, T.: Building Event-Centric

Knowledge Graphs from News. Journal of Web Semantics, Vol. 37–38, 2016, pp. 132–151, doi: 10.1016/j.websem.2015.12.004.

[7] Gerber, D.—Ngomo, A.-C. N.: Extracting Multilingual Natural-Language Patterns for RDF Predicates. In: ten Teije, A. et al. (Eds.): Knowledge Engineering and Knowledge Management (EKAW 2012). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7603, 2012, pp. 87–96, ISBN: 978-3-642-33876-2, doi: 10.1007/978-3-642-33876-2_10.

[8] Dasgupta, T.—Naskar, A.—Saha, R.—Dey, L.: CrimeProfiler: Crime Information Extraction and Visualization from News Media. Proceedings of the International Conference on Web Intelligence (WI '17), 2017, pp. 541–549, doi: 10.1145/3106426.3106476.

[9] Buitelaar, P.—Cimiano, P.—Frank, A.—Hartung, M.—Racioppa, S.: Ontology-Based Information Extraction and Integration from Heterogeneous Data Sources. International Journal of Human-Computer Studies, Vol. 66, 2008, No. 11, pp. 759–788, doi: 10.1016/j.ijhcs.2008.07.007.

[10] Alani, H.—Kim, S.—Millard, D. E.—Weal, M. J.—Lewis, P. H.—Hall, W.—Shadbolt, N. R.: Automatic Extraction of Knowledge from Web Documents. 2nd International Semantic Web Conference – Workshop on Human Language Technology for the Semantic Web and Web Services, Sanibel Island, Florida, USA, 2003. Available at: https://eprints.soton.ac.uk/258194/.

[11] Wu, Z.—Liang, C.—Giles, C. L.: Storybase: Towards Building a Knowledge Base for News Events. In: Chen, H. H., Markert, K. (Eds.): Proceedings of ACL-IJCNLP 2015 System Demonstrations. ACL, 2015, pp. 133–138, doi: 10.3115/v1/p15-4023.

[12] Stern, R.—Sagot, B.: Population of a Knowledge Base for News Metadata from Unstructured Text and Web Data. Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction (AKBC-WEKEX), Montréal, Canada, ACL, 2012, pp. 35–40. Available at: https://aclanthology.org/W12-30.pdf.

[13] Dong, X.—Gabrilovich, E.—Heitz, G.—Horn, W.—Lao, N.—Murphy, K.—Strohmann, T.—Sun, S.—Zhang, W.: Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 601–610, doi: 10.1145/2623330.2623623.

[14] Fan, J.—Kalyanpur, A.—Gondek, D. C.—Ferrucci, D. A.: Automatic Knowledge Extraction from Documents. IBM Journal of Research and Development, Vol. 56, 2012, No. 3-4, pp. 5:1–5:10, doi: 10.1147/jrd.2012.2186519.

[15] Carlson, A.—Betteridge, J.—Kisiel, B.—Settles, B.—Hruschka, E. R.—Mitchell, T. M.: Toward an Architecture for Never-Ending Language Learning. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI '10), Atlanta, Georgia, 2010, pp. 1306–1313.

[16] Gerber, D.—Hellmann, S.—Bühmann, L.—Soru, T.—Usbeck, R.—Ngomo, A.-C. N.: Real-Time RDF Extraction from Unstructured Data Streams. In: Alani, H. et al. (Eds.): The Semantic Web – ISWC 2013. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 8218, 2013, pp. 135–150, doi: 10.1007/978-3-642-41335-3_9.

[17] Wiedemann, G.—Yimam, S. M.—Biemann, C.: A Multilingual Information Extraction Pipeline for Investigative Journalism. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, 2018, pp. 78–83, doi: 10.18653/v1/D18-2014.

[18] Hecking, M.—Schwerdt, C.: Multilingual Information Extraction for Intelligence Purposes. 13th International Command and Control Research and Technology Symposium (ICCRTS): "C2 for Complex Endeavors", Seattle, WA, 2008. Available at: http://dodccrp.org/events/13th_iccrts_2008/CD/html/papers/025.pdf.

[19] Akbik, A.—Chiticariu, L.—Danilevsky, M.—Kbrom, Y.—Li, Y.—Zhu, H.: Multilingual Information Extraction with PolyglotIE. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan, 2016, pp. 268–272. Available at: https://aclanthology.org/C16-2056.pdf.

[20] Oramas, S.—Espinosa-Anke, L.—Sordo, M.—Saggion, H.—Serra, X.: Information Extraction for Knowledge Base Construction in the Music Domain. Data and Knowledge Engineering, Vol. 106, 2016, pp. 70–83, doi: 10.1016/j.datak.2016.06.001.

[21] Rebele, T.—Suchanek, F.—Hoffart, J.—Biega, J.—Kuzey, E.—Weikum, G.: YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In: Groth, P. et al. (Eds.): The Semantic Web – ISWC 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9982, 2016, pp. 177–185, doi: 10.1007/978-3-319-46547-0_19.

[22] Lehmann, J.—Isele, R.—Jakob, M.—Jentzsch, A.—Kontokostas, D.—Mendes, P. N.—Hellmann, S.—Morsey, M.—van Kleef, P.—Auer, S.—Bizer, C.: DBpedia – A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web, Vol. 6, 2015, No. 2, pp. 167–195, doi: 10.3233/sw-140134.

[23] Navigli, R.—Ponzetto, S. P.: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, Vol. 193, 2012, pp. 217–250, doi: 10.1016/j.artint.2012.07.001.

[24] Erxleben, F.—Günther, M.—Krötzsch, M.—Mendez, J.—Vrandečić, D.: Introducing Wikidata to the Linked Data Web. In: Mika, P. et al. (Eds.): The Semantic Web – ISWC 2014. Springer, Cham, Lecture Notes in Computer Science, Vol. 8796, 2014, pp. 50–65, doi: 10.1007/978-3-319-11964-9_4.

[25] Gangemi, A.—Presutti, V.—Reforgiato Recupero, D.—Nuzzolese, A. G.—Draicchio, F.—Mongiovì, M.: Semantic Web Machine Reading with FRED. Semantic Web, Vol. 8, 2017, No. 6, pp. 873–893, doi: 10.3233/sw-160240.

[26] Chen, M.—Tian, Y.—Yang, M.—Zaniolo, C.: Multilingual Knowledge Graph Embeddings for Cross-Lingual Knowledge Alignment. Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, 2017, pp. 1511–1517, doi: 10.24963/ijcai.2017/209.

[27] Klein, P.—Ponzetto, S. P.—Glavaš, G.: Improving Neural Knowledge Base Completion with Cross-Lingual Projections. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017, pp. 516–522, doi: 10.18653/v1/e17-2083.

[28] Rousseeuw, P. J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, Vol. 20, 1987, pp. 53–65, doi: 10.1016/0377-0427(87)90125-7.

[29] Malaviya, C.—Bhagavatula, C.—Bosselut, A.—Choi, Y.: Commonsense Knowledge Base Completion with Structural and Semantic Context. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, No. 03: AAAI-20 Technical Tracks 3, pp. 2925–2933, doi: 10.1609/aaai.v34i03.5684.

[30] Pezeshkpour, P.—Tian, Y.—Singh, S.: Revisiting Evaluation of Knowledge Base Completion Models. Automated Knowledge Base Construction (AKBC 2020), 2020, doi: 10.24432/C53S3W.

[31] Benetka, J. R.—Balog, K.—Norvag, K.: Towards Building a Knowledge Base of Monetary Transactions from a News Collection. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017, pp. 1–10, doi: 10.1109/jcdl.2017.7991575.

[32] Kertkeidkachorn, N.—Ichise, R.: T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. AAAI Workshops 2017, San Francisco, CA, USA, 2017.

[33] Bisandu, D. B.—Prasad, R.—Liman, M. M.: Clustering News Articles Using Efficient Similarity Measure and N-Grams. International Journal of Knowledge Engineering and Data Mining, Vol. 5, 2018, No. 4, pp. 333–348, doi: 10.1504/IJKEDM.2018.095525.

[34] Sohangir, S.—Wang, D.: Improved SQRT-Cosine Similarity Measurement. Journal of Big Data, Vol. 4, 2017, No. 1, Art. No. 25, 13 pp., doi: 10.1186/s40537-017-0083-6.

[35] English News Article. Available at: `https://indianexpress.com/article/india/gurgaon-police-arrests-key-lawrence-bishnoi-gang-member-sampat-nehra-underworld-gangster-5207714/`.

[36] Hindi News Article. Available at: `https://www.livehindustan.com/ncr/story-lawrence-bishnoi-gang-gangster-sampath-nehra-arrested-from-hyderabad-2000546.html`.

K. Srinivasa received his B.Eng. degree in computer science and engineering in 2004 from the Vijayanagara Engineering College, Bellary, Visvesvaraya Technological University, Belgaum, India and his M.Tech. degree in computer science and engineering from the National Institute of Technology Karnataka (NITK), Surathkal, India, in 2010, and he has been pursuing his Ph.D. degree in the Department of Computer Science and Engineering, at the National Institute of Technology Karnataka (NITK), Surathkal, India, from 2017. Since 2005, he has been with the Department of Computer Science and Engineering at Siddaganga Institute of Technology, Tumakuru, Karnataka, India where he is currently Assistant Professor. His current research interests include information extraction, natural language processing and knowledge management.

P. Santhi Thilagam received her B.Eng. degree in computer science and engineering in 1991, and the M.Eng. degree in computer science and engineering from College of Engineering, Guindy, Anna University, Chennai, India, in 1999, and her Ph.D. degree in information technology from the National Institute of Technology Karnataka (NITK), Surathkal, India, in 2008. Since 1996, she has been with the Department of Computer Science and Engineering at NITK Surathkal, where she is Professor. Her current research interests include database security, data management, data analysis, and distributed computing. She is Member of several technical associations, scientific committees and editorial boards. She was the recipient of the best Ph.D. thesis award in the Computer Science and Engineering Category of the Board of IT Education Standards in 2009, Ramanujan Lecture presenter award of the Institution of Engineers India (IEI-India) in 2015.

Computing and Informatics, Vol. 40, 2021, 341–367, doi: 10.31577/cai_2021_2_341

WEAKLY COMPLETE EVENT LOGS IN PROCESS MINING

Julijana Lekić

Faculty of Technical Sciences, University of Pristina in Kosovska Mitrovica
Kneza Milosa 7, 38220 Kosovska Mitrovica, Serbia
e-mail: julijana.lekic@pr.ac.rs

Dragan Milićev

Faculty of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia
e-mail: dmilicev@etf.bg.ac.rs

Abstract. Many information systems have a possibility to record their execution, and, in this way, to generate a trace about events describing the real system behaviour. From behaviour example records in traces of the event log, the α-algorithm automatically generates a process model that belongs to a subclass of Petri nets, known as workflow nets. One of the basic limiting assumptions of α-algorithm is that the event log needs to be complete. As a result of attempting to overcome the problem of completeness of the event log, we introduced the notion of weakly complete event logs, from which our modified technique and algorithm can produce the same result as the α-algorithm from complete logs on parallel processes. Thereby weakly complete logs can be significantly smaller than complete logs, considering the number of traces they consist of. Weakly complete logs were used for the realization of our idea of interactive parallel business process model generation.

Keywords: Process mining, business process model discovery, block-structured parallel process models, α-algorithm, $\alpha^{\|}$-algorithm, complete log, weakly complete event log

1 INTRODUCTION

Extraction of knowledge from traces recorded in event logs which are available in today's information systems, and use of that knowledge for discovering, monitoring and improvement of bussiness proces models, are basis for occurrence of different techniques of process mining (PM) areas [1, 2, 3]. α-algorithm is able to discover a large class of workflow (WF) nets [4] based on the behaviour recorded in event logs, with the main limiting assumption that the event log is complete. The property of completeness of the log often implies the necessity of having a large number of traces in the log on which the "representative" model for the behaviour seen in the log has to be constructed. Therefore, our challenge was to find logs with potentially much lower number of traces, which may not be complete, but are sufficiently valid so that, using the appropriate algorithm based on the evidences recorded in such logs, a "representative" model can be obtained.

To achieve this, we have partially modified the technique of process model discovering, and also the α-algorithm itself [1, 2, 3] by introducing the relation of *indirect precedence* as another basic relation between the activities recorded in the event log [5, 6]. Within that goal we firstly defined the so called *causally complete* event logs, which do not fulfil conditions of completeness, and yet using our modified algorithm $\alpha^{||}$-algorithm) we can reconstruct the original network of a parallel bussiness process. Obtaining block-structured parallel business process model with our modified method based on the event logs which fulfil the requirements of casual completeness is presented in [5].

The discovery of the parallel bussiness process model from event logs with as few recorded traces as possible was a part of our research, and the ultimate goal was finding ways for those models to be created interactively. With defining of causally complete event logs that goal was not achieved in full. Namely, even though parallel bussiness process models could be successfully discovered from causally complete event logs based on significantly smaller number of traces than from complete event logs, still we could not achieve the interactive model creation. Further research in that direction leads to defining of the so called *weakly complete* event log by which our goals could be achieved. Weakly complete event logs may contain a significantly smaller number of traces needed for obtaining the appropriate parallel business process model than in case with complete or causally complete logs.

The property of $\alpha^{||}$-algorithm, where its application on the weakly complete event logs may lead to the model discovery, was succesfully used for the creation of the parallel business process models by demonstration. For this purpose, its own demonstration graphical user interface has been created, which enables user to perform different scenarios of activity execution process using direct manipulation. The graphical user interface we created is a tool that visually shows steps of $\alpha^{||}$-algorithm. Such tool could serve as a learning tool and playground for those who want to learn more about the general α-algorithm, which is based on the same principles like its modified version $\alpha^{||}$-algorithm, and how it works. This has been

achieved through the idea in which a user enters log entries (scenario) step by step and observes a current model which is shown and obtained based on the log entries entered.

The created demonstration user interface contains the components of artificial inteligence which are reflected in the fact that the system itself suggests the order of process activitiy performance (in order to discover the process model as soon as possible) and "infers" some relations between the actvities which were not performed during the demonstration procedure. In order to find and infer relations which were not performed (inferred relations), the system uses the event log footprint[1] and certain rules described later in this paper. The results and ways of realization of this idea are presented in our paper [7].

Besides defining weakly complete event logs, this paper will present the results of experimental analysis carried out on examples of real business processes with the goal to discover the size of weakly complete event logs for observed processes expressed in the number of traces. The obtained results will be compared to the results of the experimental analysis carried out for the same processes on complete and causally complete event logs, which is presented in detail in [5]. In that way we will be able to show the improvement we made on weakly complete event logs in comparison to causally complete event logs [5] in regard of obtaining the valid model of parallel business processes based on the least possible number of traces from the event logs.

Our assumptions and preconditions for process models (that have to be block-structured parallel models), to which our algorithm is applicable, may look as a serious restriction. It really is for real-world process models in general. However, our solution still covers a respectably wide subclass of process models and represents a first step in a more ambitious attempt to solve the very serious problem of log completeness. Although having a strong limitation, we still deem that even a partial solution to such a serious problem can be beneficial for future research and is worth reporting. In our future research, we will try to expand our work to other categories of processes.

The paper is organized as follows. The next section provides an overview of the existing literature from the area of interest for our work and this paper. Section 3 gives some preliminaries about modified PM technique, modified $\alpha^{||}$-algorithm and weakly complete event logs necessary for someone to further follow the content of this paper. The problem with dangling nods in WF-net that is encountered during the discovery of the business process model from weakly complete event logs as well as the way of resolving that problem are presented. An example of the $\alpha^{||}$-algorithm application on the weakly complete event log is presented at the end of this section. Section 4 describes the ProM framework for applying the $\alpha^{||}$-algorithm on weakly complete event logs, within which the plug-ins we have developed for the needs of the

[1] Footprint of the event log is a matrix in which the defined relations between any two activities can be represented [2, 3, 5].

experimental analysis are presented. The results of the application of α-algorithm and $\alpha^{||}$-algorithm on weakly complete log which is used as a running example in the paper are also shown at the end of the section. In Section 5 we present the results of the performed experimental analysis. Section 6 contains some conclusions and guidelines for the future work and research.

2 RELATED WORK

The process mining idea is not a new one. One of the earliest examples of usage of PM in the context of workflow management, based on workflow graphs, was presented in [8]. Cook and Wolf have investigated similar issues in the context of software engineering, looking in particular sequential [9] and parallel behaviour of the process [10]. Dealing with sequential processes, in [9] they describe three methods for detection, one of which uses a neural network, while the other one is entirely based on the algorithmic approach, and the third one uses a Markovian approach. In [10], the same authors extended their work to concurrent processes, where they propose specific metrics to discover models out of event streams, but they do not provide an approach to generate explicit process models. Herbst also dealt with the No. of PM in the context of workflow management [11, 12] using an inductive approach, observing sequential [12] and parallel models [11] separately. A notable difference between his work and other approaches is that the same task can appear multiple times in the workflow model, i.e., the approach allows for duplicate tasks. Schimm [13] has developed a mining tool suitable for discovering hierarchically structured workflow processes.

Although many researchers have dealt with the ideas of process mining, the most comprehensive study is presented in the works of W. M. P. van der Aalst and his collaborators [1, 2, 3, 14, 15, 16, 17]. In [2, 3] a detailed description and formalization of techniques for discovering processes from workflow logs is presented, the α-algorithm for extracting process models from such logs is defined, and a representation of the model obtained in the form of a *sound*[2] WF-net is shown. An introduction to PM and an overview of existing techniques for discovering processes, and the problems which have been encountered in the application of the α-algorithm have been most fully presented in [1]. An overview of best practices and challenges is presented in [18], which is the work of a group of experts that was created in order to promote research, development and understanding of process mining as well as its implementation and evolution.

To discover process models from traces recorded in workflow logs, many techniques have been proposed [2, 3, 8, 9, 16, 19]. Many of these techniques use Petri nets in the process of discovering and presenting the discovered process model. However, other very different approaches are also used for the same purpose.

[2] Soundness corresponds to liveness and safeness of the corresponding short-circuited net [2, 3]. The set of all sound WF-nets (SWF) is denoted with `W`.

Although the original α-algorithm is able to discover a large class of WF-nets, there are problems it cannot cope with: incompleteness of the event log, rare behaviours, complex routing constructions and others [1]. As a consequence, there is a large number of algorithms that overcome lacks of the basic α-algorithm [20]. Some of them are variants of the original α-algorithm, such as, for example, the α^+ algorithm [15] and α^{++} algorithm [17], while others use a completely different approach, such as: heuristic mining [9], genetic mining [16], fuzzy mining [21], process mining from a basis of regions [19] or flexible heuristics miner [22]. The inductive miner (IM) [23], aims to discover block-structured process models that are sound and fitting to the behavior represented in the event log.

The above mentioned algorithms and techniques have overcome some of problems, but the problem of incompleteness of logs, to the best of our knowledge, has not been overcome. This fact still remains a challenge for future research.

The subject of our research presented in this paper is the problem of completeness of event logs [1]. Part of this research was preliminarily announced in our previous paper [24], and has been presented in [5], where we also dealt with logs that do not meet the requirement of completeness. In this paper, the more detailed description of weakly complete log is given than in [24], based on the definitions and rules. Moreover, the comparison of results obtained by applying our algorithm [5] with the results of other process mining algorithms on weakly complete logs was done. Finally, another big difference from the paper [24] is that this paper brings an extensive experimental analysis and its results are presented here. Unlike [5] where we dealt with causally complete event logs, this paper focuses on weakly complete event logs, that we obtained in the attempt to improve the characteristics of causally complete event logs. One of those improved characteristics is a smaller size of event logs, which was a basic parameter for the experimental analysis that was performed. In this analysis, values of the minimal sizes of complete logs, causaly complete logs and weakly complete logs are compared for 100 real examples of parallel business processes. In this paper, plug-ins are also presented in the existing ProM framework designed for the needs of the experimental analysis.

Another improvement that we have achieved with weakly complete event logs compared to causally complete event logs is the property of weakly complete event logs that based on them, the models of parallel business processes can be interactively (using demonstration) generated, which is presented in detail in our paper [7].

3 PRELIMINARIES

In this section, we give some definitions of concepts used throughout this paper. We must note that the detailed description of the concepts that will be mentioned, as well as how they were obtained, is shown in our paper [5], therefore we will here provide only the short review necessary for someone to further follow the content of this paper.

3.1 Modified Technique and Algorithm for Discovering Process Models

For the purposes of this paper, we will show only the basic characteristics of our modified PM technique and modified α-algorithm.

The log-based relations that are used to indicate the relevant patterns in the log in our modified technique of discovering process models are defined by Definition 1.

Definition 1. (Log-based ordering relations, in the modified PM technique of discovering process models). Let L be an event log over $\mathcal{A}$[3], i.e., $L \in \mathcal{P}\ (\mathcal{A}^*)$[4]. Let $a, b \in \mathcal{A}$ be two activities. Then, by definition:

- $a >_L b$ if and only if there is a trace $\sigma = \{t_1, t_2, t_3, \ldots, t_n\}$ and $i \in \{1, \ldots, n-1\}$ such that $\sigma \in L$ and $t_i = a$ and $t_{i+1} = b$,
- $a \gg_L b$ if and only if there is a trace $\sigma = \{t_1, t_2, t_3, \ldots, t_n\}$ and there are $i, j \in \{1, \ldots, n\}$ such that $i + 2 \leq j$, where $\sigma \in L$ and $t_i = a, t_j = b$, and it is not that $a >_L b$,
- $a \rightarrow_L b$ if and only if $a >_L b$, and it is not $b >_L a$, and it is not $b \gg_L a$,
- $a \Rightarrow_L b$ if and only if $a \gg_L b$, and it is not $b >_L a$, and it is not $b \gg_L a$,
- $a \sharp_L b$ if and only if it is not $a >_L b$, and it is not $b >_L a$, and it is not $a \gg_L b$, and it is not $b \gg_L a$,
- $a ||_L b$ if and only if $a >_L b$ and $b >_L a$, or $a >_L b$ and $b \gg_L a$, or $a \gg_L b$ and $b >_L a$, or $a \gg_L b$ and $b \gg_L a$.

The defined relations can be represented with a matrix, which represents a footprint of the event log.

The $\alpha^{||}$-algorithm, where "||" reflects the fact that the algorithm targets parallel business processes, can be described by Definition 2.

Definition 2. ($\alpha^{||}$-algorithm). Let L be an event log over $T \subseteq \mathcal{A}$. $\alpha^{||}(L)$ is defined as follows:

1. $T_L = \{t \in T \mid (\exists \sigma \in L)\ t \in \sigma\}$,
2. $T_I = \{t \in T \mid (\exists \sigma \in L)\ t = \mathit{first}(\sigma)\}$,
3. $T_O = \{t \in T \mid (\exists \sigma \in L)\ t = \mathit{last}(\sigma)\}$,
4. $X_L = \{(A, B) \mid A \subseteq T_L \wedge A \neq \emptyset \wedge B \subseteq T_L \wedge B \neq \emptyset \wedge (\forall a \in A)(\forall b \in B)(a \rightarrow_L b)\}$,
5. $P_L = \{p_{(A,B)} \mid (A, B) \in X_L\} \cup \{i_L, o_L\}$,

[3] $\mathcal{A}$ is the set of business process activities.

[4] $\mathcal{A}^*$ is the set of all finite sequences (traces) of the elements of $\mathcal{A}$, and $\mathcal{P}(\mathcal{A}^*)$ is the powerset of $\mathcal{A}^*$.

6. $F_L = \{(a, p_{(A,B)}) \mid (A, B) \in X_L \wedge a \in A\} \cup \{(p_{(A,B)}, b) \mid (A, B) \in X_L \wedge b \in B\} \cup \{(i_L, t) \mid t \in T_I\} \cup \{(t, o_L) \mid t \in T_O\}$,
7. $\alpha^{||}(L) = (P_L, T_L, F_L)$.

A necessary condition for discovering the original network by the α-algorithm is that the log on which the algorithm is applied needs to be complete, where the condition of completeness is based on the relation $>_L$ [1, 2, 3]. The condition of completeness in our modified PM technique for model discovering is related to the causality relation $\rightarrow_L$.

For a particular process model to be discovered, there may be a large (in general, an unlimited) number of different complete logs. However, all these complete logs have the same footprint, i.e., the same causality relation. We call this relation the *basic causality relation.*

Definition 3. (The basic causality relation). Let $N = (P, T, F)$[5] be a sound WF-net, i.e., $N \in \mathcal{W}$, and let L be a complete workflow log of N. $\rightarrow^B{}_N$ is the basic causality relation of network N iff $\rightarrow^B{}_N = \rightarrow_L$.

On the other hand, there may exist other logs for the same process model that are not complete or causally complete [5], but which have the same causality relation obtained from those logs. We are focused on investigating such logs, which we refer to as *weakly complete logs.* Obviously, the idea is to find weakly complete logs that may be, in general, significantly smaller than fully complete logs (in the terminology of the original α-algorithm) and causally complete logs (in the terminology of the modified $\alpha^{||}$-algorithm).

3.2 Weakly Complete Event Logs

As it has already been shown in [5], by using the modified PM technique on parallel processes we can obtain the original network from any event log in which $\rightarrow_L = \rightarrow^B{}_N$. Therefore, the main task is to find a log with the causality relation that is equal to the basic causality relation, and then apply the $\alpha^{||}$ algorithm, which leads to the original network of the parallel processes. There are event logs in which the causality relation is not equal to $\rightarrow^B{}_N$, but from the footprint of that log we can subsequently conclude the elements of the causality relation on whose joining the causality relation of the log becomes equal to $\rightarrow^B{}_N$. The log with such property has been named *weakly complete event log – L_w.*

Examples we have analyzed show that the original network can be discovered from an incomplete log L for which the following holds: $\rightarrow^B{}_N \subset (\rightarrow_L \cup \Rightarrow_L)$, and $\rightarrow_L \subset \rightarrow^B{}_N$; if the causality relation of that log is joined by the elements of causality relation which can be inferred from the footprint of the log. Such a log we call a weakly complete log (L_w) [24].

[5] Tuple (P, T, F) originates from the Place/Transition net [2, 3] definition.

Definition 4. (Weakly complete event log). Let $N = (P, T, F)$ be a sound WF-net, i.e., $N \in \mathcal{W}$, and let $\rightarrow^B{}_N$ be the basic causality relation of N. L_w is a *weakly complete* workflow log of N iff:

1. $\rightarrow^B{}_N \subset (\rightarrow_{L_w} \cup \Rightarrow_{L_w})$, and $\rightarrow_{L_w} \subset \rightarrow^B{}_N$, and
2. for any $t \in T$ [6] there is $\sigma \in L_w$ so that $t \in \sigma$.

Even though in weakly complete event logs it cannot be inferred based on log traces that $\rightarrow_{L_w} = \rightarrow^B{}_N$ (but $\rightarrow_{L_w} \subset \rightarrow^B{}_N$), the elements of the causality relation which are not in $\rightarrow_{L_w}$, but are in $\rightarrow^B{}_N$, may be subsequently inferred from the log footprint. Those elements create causality relation called the *inferred causality relation*, denoted with $\rightarrow^i$. The causality relation that inserted the final appearance of the log footprint (denoted with $\rightarrow_{Lf}$), and on which the $\alpha^{||}$-algorithm is applied, becomes: $\rightarrow_{Lf} = \rightarrow_{Lw} \cup \rightarrow^i$ by which we get that $\rightarrow_{Lf} = \rightarrow^B{}_N$. Finding of the *footprint causality relation* $\rightarrow_{Lf}$ which equals to basic causal relation $\rightarrow^B{}_N$ is the condition enough for discovering the original network based on $\rightarrow_{Lf}$, using the $\alpha^{||}$-algorithm in a manner shown in [24].

3.3 The Dangling Nodes Problem

A network obtained from a weakly complete log often contains dangling nodes, i.e., activities (node in Petri net) without predecessors and/or successors. The occurrence of dangling nods in the network obtained based on the traces recorded in the weakly complete event log is due to a large number of activities that can be performed simultaneously and due to the rapid detection of parallelism by modified PM technique. Besides that, the occurrence of dangling nods in the network is also due to the property of weakly complete logs that all the causality relation elements cannot be discovered from the record written in the log traces, alone.

Definition 5. (Dangling node). Let $N = (P, T, F)$ be a network with initial place i and ending place o, and let a be such an activity that $a \in T$ and $a \notin \{i, o\}$. Activity $a \in T$ is a *dangling* node in network N if there is no activity $b \in T$ such that $a\bullet$ [7] $\cap \bullet b \neq \emptyset$ or there is no activity $b \in T$ such that $b \bullet \cap \bullet a \neq \emptyset$.

If activity $a \in T$ is a dangling node in network N and if there is no activity $b \in T$, such that $a \bullet \cap \bullet b \neq \emptyset$, then it can be said that activity a does not have its *successor* in network N. If activity $a \in T$ is a dangling node in network N and there is no activity $b \in T$ such that $b \bullet \cap \bullet a \neq \emptyset$, than it can be said that activity a does not have its predecessor in network N.

[6] $T \subseteq \mathcal{A}$, $\mathcal{A}$ is a set of activities, T is a finite set of transitions in the Petri net [4], σ is a trace such that $\sigma \in L$ [2, 3].

[7] The concepts of marking and tokens $\bullet$ are well known for Petri nets and precise definitions can be found in [2, 3].

The definition of a WF-net [4], includes an assumption of network connectivity, which means connectivity of all nodes in the network, and which prohibits the existence of dangling nodes. Therefore, the network obtained based only on the records written in weakly complete event log which contains dangling nods in itself is not a proper SWF-net, and it is not equal to the original network. To overcome the problem of dangling nodes, we observed the relations in footprints, and based on those we have defined the rules of inference of the direct successors and predecessors from the indirect ones. Thus, for each activity that is a dangling node, a successor and/or predecessor can be found.

Network obtained based on weakly complete event log which contains dangling nods has in the event log footprint at least one activity that in its table row does not have relation $\rightarrow$ (if the activity has no successor) or relation $\leftarrow$ (if the activity has no predecessor) [24].

Rule 1 (Determining the inferred causality relation $\rightarrow^i$ when activity has no successor). Let a be activity which in its log footprint row has no relation $\rightarrow$ then by definition: $a \rightarrow^i c$ iff in footprint $a \Rightarrow c$, and there is b such that $b \rightarrow c$, where $a||b$.

Rule 2 (Determining the inferred causality relation $\leftarrow^i$ when activity has no predecessor). Let a be activity which in its log footprint row has no relation $\leftarrow$, then by definition: $a \leftarrow^i c$ iff in footprint $a \Leftarrow c$ and there is b such that $b \leftarrow c$, where $a||b$.

Using Rule 1 and Rule 2, the elements of the inferred causality relation $\rightarrow^i$ ($\leftarrow^i$) can be determined based on the footprint of the event log. This contributes to the discovery of those causality relations which cannot be discovered from the weakly complete event log traces alone. The causality relation which enters the final log footprint appearance and on which $\alpha^{||}$-algorithm applies becomes: $\rightarrow_{L_f} = \rightarrow_{L_w} \cup \rightarrow^i$. In that way $\rightarrow_{L_f}$ becomes equal to the basic causality relation, i.e. $\rightarrow_{L_f} = \rightarrow^B{}_N$, which allows the discovery of the original network. The larger the number of elements $\rightarrow^i$ in relation $\rightarrow_{L_f}$, the poorer the weakly complete event log on basis of which the model can be constructed, i.e., with a smaller number of the traces recorded.

3.4 An Example of the $\alpha^{||}$-Algorithm Application on the Weakly Complete Event Log

Let us consider a parallel process model shown in Figure 1, and a log L with records obtained after several executions of the process.

$$L = [\langle a, d, b, e, c, f, h, g, k\rangle^4, \langle d, e, a, c, b, f, g, h, k\rangle^3].$$

The basic causality relation for this example is:

$$\rightarrow^B{}_N = \{(a, b), (a, c), (d, e), (b, f), (c, f), (e, f), (f, g), (f, h), (g, k), (h, k)\}.$$

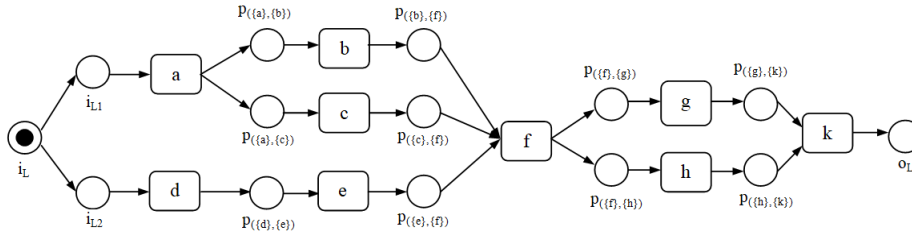


Figure 1. Example of a block-structured parallel process model

The log-based ordering relations for this example are:

$$>_L = \{(a,b),(b,c),(c,d),(d,e),(e,f),(f,g),(g,h),(h,k),(d,e),(e,a),(a,c),(c,b),(b,f),(f,h),(h,g),(g,k)\},$$

$$<_L = \{(b,a),(c,b),(d,c),(e,d),(f,e),(g,f),(h,g),(k,h),(e,d),(a,e),(c,a),(b,c),(f,b),(h,f),(g,h),(k,g)\},$$

$$\gg_L = \{(a,d),(a,e),(a,f),(a,g),(a,h),(a,k),(b,d),(b,e),(b,g),(b,h),(b,k),(c,e),(c,f),(c,g),(c,h),(c,k),(d,f),(d,g),(d,h),(d,k),(e,g),(e,h),(e,k),(f,k),(d,b),(d,f),(d,h),(d,g),(d,k),(e,c),(e,b),(e,h),(e,g),(e,k),(a,f),(a,h),(a,g),(a,k),(d,a),(d,c),(c,f),(c,h),(c,g),(c,k),(b,h),(b,g),(b,k),(f,k)\},$$

$$\ll_L = \{(d,a),(e,a),(f,a),(g,a),(h,a),(k,a),(d,b),(e,b),(g,b),(h,b),(k,b),(e,c),(f,c),(g,c),(h,c),(k,c),(f,d),(g,d),(h,d),(k,d),(g,e),(h,e),(k,e),(k,f),(a,d),(c,d),(b,d),(f,d),(h,d),(g,d),(k,d),(c,e),(b,e),(h,e),(g,e),(k,e),(f,a),(h,a),(g,a),(k,a),(f,c),(h,c),(g,c),(k,c),(h,b),(g,b),(k,b),(k,f)\},$$

$$||_L = \{(b,c),(c,b),(c,d),(d,c),(g,h),(h,g),(e,a),(a,e),(a,d),(d,a),(b,d),(d,b),(b,e),(e,b),(c,e),(e,c)\},$$

$$\rightarrow_L = \{(a,b),(d,e),(e,f),(f,g),(h,k),(a,c),(b,f),(f,h),(g,k)\},$$

$$\leftarrow_L = \{(b,a),(e,d),(f,e),(g,f),(k,h),(c,a),(f,b),(h,f),(k,g)\},$$

$$\Rightarrow_L = \{(a,f),(a,g),(a,h),(a,k),(b,g),(b,h),(b,k),(c,f),(c,g),(c,h),(c,k),$$
$$(d,f),(d,g),(d,h),(d,k),(e,g),(e,h),(e,k),(f,k)\},$$

$$\Leftarrow_L = \{(f,a),(g,a),(h,a),(k,a),(g,b),(h,b),(k,b),(f,c),(g,c),(h,c),(k,c),$$
$$(f,d),(g,d),(h,d),(k,d),(g,e),(h,e),(k,e),(k,f)\},$$

$$\sharp_L = \{(a,a),(b,b),(c,c),(d,d),(e,e),(f,f),(g,g),(h,h),(k,k)\}.$$

The footprint of the event log L is given in Table 1.

	a	b	c	d	e	f	g	h	k
a	$\sharp$	$\rightarrow$	$\rightarrow$	$\|\|$	$\|\|$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
b	$\leftarrow$	$\sharp$	$\|\|$	$\|\|$	$\|\|$	$\rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
c	$\leftarrow$	$\|\|$	$\sharp$	$\|\|$	$\|\|$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
d	$\|\|$	$\|\|$	$\|\|$	$\sharp$	$\rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
e	$\|\|$	$\|\|$	$\|\|$	$\leftarrow$	$\sharp$	$\rightarrow$	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
f	$\Leftarrow$	$\leftarrow$	$\Leftarrow$	$\Leftarrow$	$\leftarrow$	$\sharp$	$\rightarrow$	$\rightarrow$	$\Rightarrow$
g	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\leftarrow$	$\sharp$	$\|\|$	$\rightarrow$
h	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\leftarrow$	$\|\|$	$\sharp$	$\rightarrow$
k	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\Leftarrow$	$\leftarrow$	$\leftarrow$	$\sharp$

Table 1. Footprint of the log L

It can be noted that the causality relation of log L is not equal to the basic causality relation, i.e., $\rightarrow_L \neq \rightarrow^B{}_N$, but it holds: $\rightarrow^B{}_N \subset (\rightarrow_L \cup \Rightarrow_L), \rightarrow_L \cup \rightarrow^B{}_N$, which makes L a weakly complete log. It can also be seen from the footprints that the activity c does not have its direct successors (the rows c do not have $\rightarrow$), which means that there will be dangling nodes in the network.

According to the Rule 1, i.e., the rule of inference of direct from indirect successors in networks with dangling nodes, we obtain:

$$c \Rightarrow_L f, \quad b \rightarrow_L f, \quad c \|_L b$$

then

$$c \rightarrow^i{}_L f,$$

or

$$c \rightarrow_L f, \quad e \rightarrow_L f, \quad c \|_L e$$

then

$$c \rightarrow^i{}_L f$$

i.e.,

$$f \leftarrow^i{}_L c.$$

Thus: $\leftarrow^{i}{}_{L} = (c, f)$, i.e.:

$$\rightarrow_{L_f} = \rightarrow_L \cup \rightarrow^{i}{}_{L} = \{(a,b),(a,c),(d,e),(b,f),(c,f),(e,f),(f,g),(f,h),$$
$$(g,k),(h,k)\}.$$

It can be noted that now it holds $\Rightarrow_{Lf} = \rightarrow^{B}{}_{N}$. By applying the $\alpha^{||}$-algorithm to the given log L, we obtain the following:

1. $T_L = \{a, b, c, d, e, f, g, h, k\}$,
2. $T_I = \{a, d\}$,
3. $T_O = k$,
4. $X_L = \{(\{a\},\{b\}),(\{a\},\{c\}),(\{d\},\{e\}),(\{b\},\{f\}),(\{c\},\{f\}),(\{e\},\{f\}),(\{f\},\{g\}),(\{f\},\{h\}),(\{g\},\{k\}),(\{h\},\{k\})\}$,
5. $P_L = \{p(\{a\},\{b\}),p(\{a\},\{c\}),p(\{d\},\{e),p(\{b\},\{f\}),p(\{c\},\{f\}),p(\{e\},\{f\}),p(\{f\},\{g\}),p(\{f\},\{h\}),p(\{g\},\{k\}),p(\{h\},\{k\}),i_{L1},i_{L2},o_L\}$,
6. $F_L = \{(a,p_{(\{a\},\{b\})}),(p_{(\{a\},\{b\})},b),(a,p_{(\{a\},\{c\})}),(p_{(\{a\},\{c\})},c),(d,p_{(\{d\},\{e\})}),(p_{(\{d\},\{e\})},e),(b,p_{(\{b\},\{f\})}),(p_{(\{b\},\{f\})},f),(c,p_{(\{c\},\{f\})}),(p_{(\{c\},\{f\})},f),(e,p_{(\{e\},\{f\})}),(p_{(\{e\},\{f\})},f),(f,p_{(\{f\},\{g\})}),(p_{(\{f\},\{g\})},g),(f,p_{(\{f\},\{h\})}),(p_{(\{f\},\{h\})},h),(g,p_{(\{g\},\{k\})}),(p_{(\{g\},\{k\})},k),(h,p_{(\{h\},\{k\})}),(p_{(\{h\},\{k\})},k),(i_{L1},a),(i_{L2},d),(k,o_L)\}$,
7. $\alpha^{||}(L) = (P_L, T_L, F_L)$.

4 PROM FRAMEWORK FOR APPLYING THE $\alpha^{||}$-ALGORITHM

For the need of discovering original networks of block-structured parallel business processes by the modified PM method and the $\alpha^{||}$-algorithm based on causally complete [5] and weakly complete logs, we have developed a plug-in *Alpha||-algorithm* (Figure 5) for the existing ProM framework [26]. The program code of the plug-in is located in a separate, new package, *alpha_parallel_algorithm* and is located at address [27].

At the same address there is a program code of the *Alpha||-algorithm – helper plug-in* (Figure 5), which is given in a separate package *alpha_parallel_algorithm_basic_causal_relation*. The mentioned plug-in is created for the purpose of extraction of basic causal relations from a complete event log.

It should be noted that this utility of extracting the basic causality relation is not used by the $\alpha^{||}$-algorithm. As it has been explained, the point is that our algorithm is guaranteed to discover the original process model if the input log is causally or weakly complete, just as the original α-algorithm is guaranteed to restore the original model if the log is fully complete; on the opposite, none of these algorithms can guarantee that the obtained model is the original one if the log is not causally, weakly or fully complete, respectively. The plug-in is just an independent, helper utility that can be used during experimentation to check whether the given log

is weakly complete or not, for the given known process model. Of course, in the procedure of process discovery, the model is unknown.

In the procedure of discovering original networks from weakly complete event logs performance of additional operations by Rules 1 and 2 is needed, which allows inferring of direct successors based on indirect ones or predecessors respectively, in network with dangling nods. For each indirect relation a check is being made to determine if there is such an event in the set of events which meets any of the conditions from the above mentioned rules. If any such event is found the newly discovered causality relation is added to the *inferred causal_relations* collection.

Inferring direct successors based on indirect successors is implemented in the *weakly_completed_logs_find_casual_from_indirect_succesor* function, as shown in Appendix 1, which is located at address [27]. In addition, inferring of direct predecessors based on indirect predecessors is implemented in the *weakly_completed_logs_find_casual_from_indirect_predecessor* function as shown in [27, Appendix 1].

Figure 2 shows an example of a block-structured model of a parallel business process [5, 25], presented in a form of a Petri net, which represents our running example in this paper (as well as in [5] and [24]). Discovering the original model from Figure 2 from a weakly complete event log $L_w = [\langle a, b, c, d, e, f, g, h\rangle^3, \langle a, f, g, c, e, d, b, h\rangle^2]$ is presented in detail in [24].

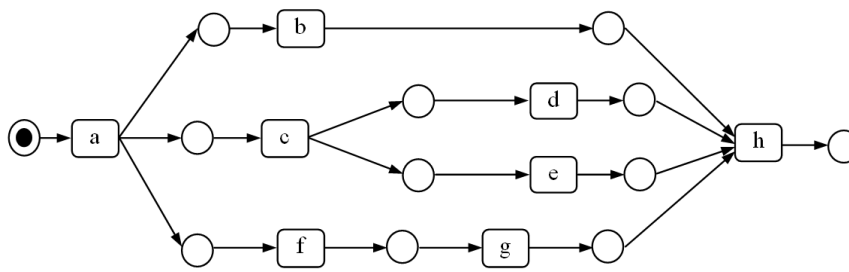


Figure 2. Example of a block-structured parallel process model

Figure 3 shows the $N^{||}$ [8] network obtained by applying the $\alpha^{||}$-algorithm over log $L_w = [\langle a, b, c, d, e, f, g, h\rangle^3, \langle a, f, g, c, e, d, b, h\rangle^2]$ and using the plug-in *Alpha||-algorithm*.

It can be noticed that the network $N^{||} = \alpha^{||}(L_w)$ in Figure 3 is equal to the original network in Figure 2, although it is obtained from a weakly complete log which is not complete and which is smaller than the complete log and causaly complete log, as it will be shown later in this paper.

[8] $N^{||} = (P, T, F)$ denotes a parallel process network [5].

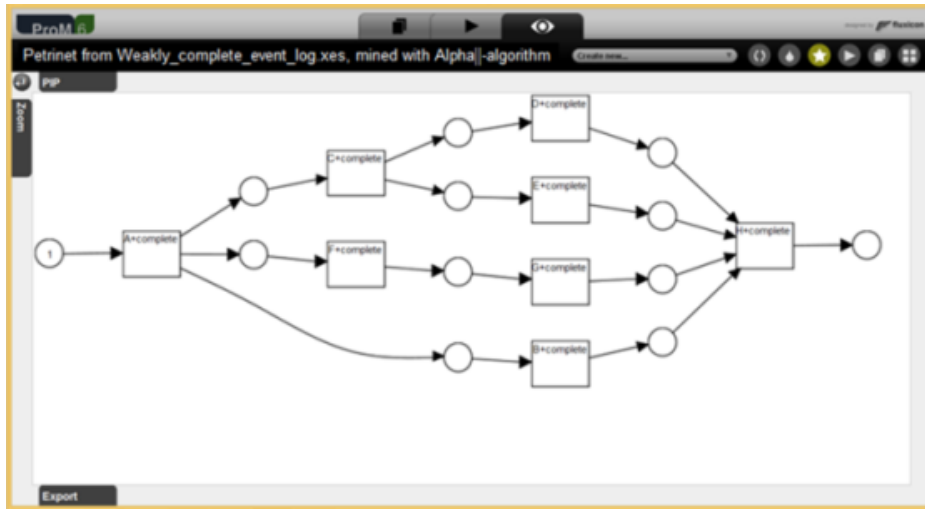


Figure 3. WF-net $N^{\|} = \alpha^{\|}(L_w)$

4.1 Comparison of Application of the $\alpha^{\|}$-Algorithm and Other Algorithms on Weakly Complete Event Logs

It can be seen from the above said that we deal with the problem of completeness which originates from the original α-algorithm and is present in all other its modifications or other algorithms for discovering process models. As other algorithms resulted mainly in the attempt to overcome some other problems, but not the problem of completeness, and as our algorithm was created by a modification of the basic α-algorithm, it would be most appropriate to compare it with the α-algorithm in the first place.

In order to evaluate the potential and effectiveness of our algorithm, we have applied several other algorithms to the same sample log $L_w = [\langle a, b, c, d, e, f, g, h\rangle^3, \langle a, f, g, c, e, d, b, h\rangle^2]$ and checked their ability to rediscover the original model. The sample log L_w is not complete, because there are missing elements in the relation $>_L$: $a > c, b > d, b > e, b > g, c > b, c > f, c > g, d > b, d > f, d > g, d > h, e > b, e > g, e > h, f > b, f > c, f > d, f > e, g > d$ and $g > e$, which could be potentially performed on the basis of the process model given in Figure 2, and the resulting WF-net $N^{\|}$.

When the original α-algorithm is applied on this weakly complete log L_w, the model shown in Figure 4 is obtained.

Due to the lack of elements of relations: $b > d$, $b > e$, $b > g$, $c > b$, $d > b$, $e > b$ and $f > b$, Alpha Miner was unable to detect that the activity b is parallel to the activities c, d, e, f and g. Due to the lack of elements of relations: $f > b$, $f > c$, $f > d$, $f > e$, $c > f$ and $d > f$, Alpha Miner was unable to detect that the activity f is parallel to the activities b, c, d and e. Due to the lack of elements of

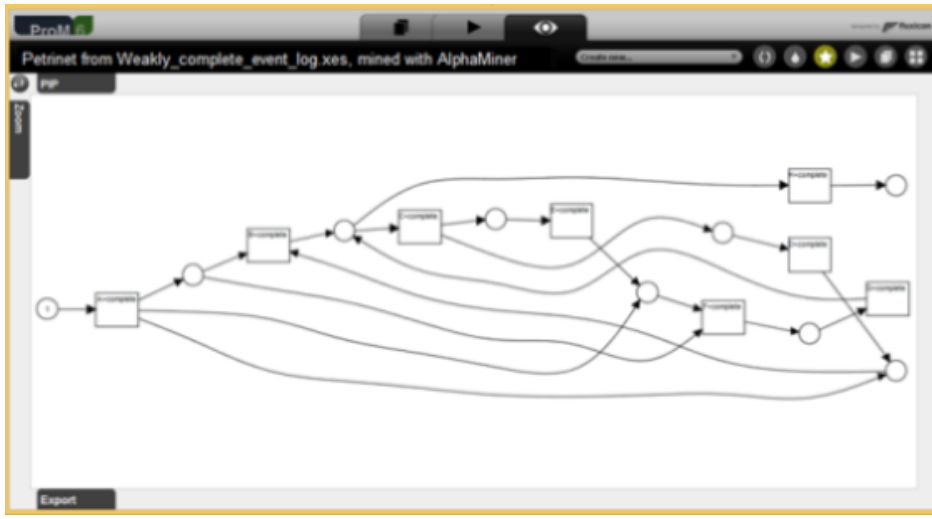


Figure 4. The result of the application of Alpha Miner to the weakly complete log L_w

relations: $c > b$, $c > f$, $c > g$ and $f > c$, Alpha Miner was unable to detect that the activity c is parallel to the activities b, f and g. Due to the lack of elements of relations: $b > g$, $c > g$, $d > g$, $e > g$, $g > d$ and $g > e$, Alpha Miner was unable to detect that the activity g is parallel to the activities b, c, d and e. For these reasons, the model obtained by Alpha Miner is so complex and different from the original network.

The Appendix 2 which is located at address [27] presents the results of the application of the available plug-ins for several other algorithms on the same given weakly complete log L_w: Alpha++ Miner, Heuristics Miner, Fuzzy Miner, Genetic Miner, ILP Miner, Mine transition system and Inductive Miner. As it can be seen from [27, Appendix 2], neither of the selected algorithms have succeeded to rediscover the original model from the given weakly complete log L_w. On the contrary, in most cases, the rediscovered models were rather complex and very far from the original model. We also give our opinion about the reasons of the inability to rediscover the original model for these algorithms.

5 EXPERIMENTAL ANALYSIS

Our experimental analysis was performed on real examples, where the size of minimal complete, minimal causaly complete and minimal weakly complete logs were compared. In order to achieve this within the existing ProM framework [21], another plug-in has been developed *Alpha‖-algorithm – minimal logs from complete log* which, from the given complete log extracts complete, causally complete and weakly complete logs with minimal possible number of traces, comparing their size

(Figure 5). The program code of the plug-in is located in a separate, new package, *alpha_parallel_algorithm_minimal_logs_from_complete_log*, and is located at address [27].

Figure 5. View of the ProM framework with an active plug-in Alpha||-algorithm – minimal logs from complete log

5.1 Procedure for Carrying Out Experimental Analysis

We will show the procedure that has been done in the experimental analysis in the example which model is shown in Figure 2. Let us observe the L event log with records obtained after several executions of the process in Figure 2.

$$\begin{aligned} L = [&\langle a,b,c,d,e,f,g,h\rangle, \langle a,f,g,b,c,e,d,h\rangle, \langle a,c,d,e,f,g,b,h\rangle,\\ &\langle a,b,f,g,c,d,e,h\rangle, \langle a,c,b,d,e,f,g,h\rangle, \langle a,c,b,e,d,f,g,h\rangle,\\ &\langle a,f,b,g,c,d,e,h\rangle, \langle a,c,f,b,g,e,d,h\rangle, \langle a,f,c,g,b,e,d,h\rangle,\\ &\langle a,f,g,c,d,b,e,h\rangle, \langle a,b,f,c,d,g,e,h\rangle, \langle a,c,e,b,d,f,g,h\rangle,\\ &\langle a,b,c,f,e,g,d,h\rangle, \langle a,c,f,d,g,e,b,h\rangle]. \end{aligned}$$

Only a variety of traces are displayed in the log, with no indication of the number of their occurrences, since the frequency of their occurrence is not relevant to this research. The log L has 14 traces and fulfils the conditions of completeness [2, 3] for the model in Figure 2.

When an active plug-in: Alpha||-algorithm – minimal logs from complete log (Figure 5), is started on the imported log L, we get the size and the appearance

of the minimal complete event log (has 14 traces), minimal causaly complete event log (has 6 traces) and minimal weakly complete event log (has two traces). The procedure shown is applied to all selected examples used in experimental analysis.

The experimental analysis was performed on a sample of 100 real examples obtained by arbitrary manual search of the Internet and selecting publicly available models of business processes, which fulfill our conditions of block-structured models of parallel processes[9]. The considered examples with their .xes files complete, causaly complete and weakly complete logs can be found at the address given in [27].

Some characteristics that reflect the network structure and size of the analysed examples are given in Tables 2 and 3 in [5]. These characteristics are expressed by the total number of activities in the network and a number of branches in the network. As with block-structured parallel processes there is one input and one output [5, 25] and it can be presented by the structure of the tree, by "branch" we meant a direct route from the entrance to the exit of the network.

5.2 Analysis Results

Table 2 presents results of the performed comparative analysis of the minimal size of complete, causaly complete and weakly complete logs, needed for discovering original networks of the considered examples.

For easier understanding of Tables 2 and 3 as well as Figures 6, 7 and 8, the following notations are used:

- N_{mcl} denotes the number of traces in minimal complete logs,
- N_{mccl} denotes the number of traces in minimal causaly complete logs,
- N_{mwcl} denotes the number of traces in minimal weakly complete logs,
- N_a denotes the total number of activities in parallel branches,
- N_b denotes the number of parallel branches in the network.

The performed experimental analysis has shown that the size of weakly complete logs from which the original networks of the observed parallel business processes can be discovered are lower, or (in the worst case) equal to the size of complete and causaly complete logs.

From Table 2 it can be seen that in 99 examples (from the examined 100 examples), the size of the minimal weakly complete logs is less than the size of minimal complete logs is lower than the size of minimal complete by an average of 52.74 %, while in one example only their values are equal. Besides that, Table 2 also shows that the size of the minimal weakly complete logs in the observed examples is only 2 or 3 traces.

[9] The models were found by searching the Web for the keywords: block-structured parallel process, parallel business process, activity diagram, BPMN diagrams etc.

N	Number of Traces in Logs			N_{mwcl} Less Than N_{mcl}	N_{mwcl} Less Than N_{mccl}
	N_{mcl}	N_{mccl}	N_{mwcl}	%	%
1	2	2	2	0.00	0.00
12	3	2	2	33.33	0.00
13	4	2	2	50.00	0.00
39	4	3	2	50.00	33.33
1	4	3	3	25.00	0.00
5	5	2	2	60.00	0.00
6	5	3	2	60.00	33.33
3	5	4	2	60.00	50.00
4	6	2	2	66.67	0.00
3	6	3	2	66.67	33.33
1	6	5	2	66.67	60.00
3	7	3	2	71.43	33.33
3	8	4	2	75.00	50.00
1	9	4	2	77.78	50.00
1	9	4	3	66.67	25.00
2	10	3	3	70.00	0.00
1	10	5	3	70.00	40.00
1	11	3	2	81.82	33.33
Total				On average less by	
100				52.742 %	22.08 %

Table 2. Results of the comparative analysis of the minimal size of complete, causaly complete and weakly complete logs

It can also be seen from Table 2 that the minimal weakly complete event logs are in average smaller than the minimal causaly complete logs by 22.08 %, observed in a sample of 100 examples. In none of the observed examples the number of traces in the minimal causally complete event log is lower than the number of traces in the minimal weakly complete log.

In order to confirm that the hypothesis that the size of weakly complete logs is lower than the size of complete logs and causaly complete logs is statistically relevant, we have applied the Wilcoxon-Mann-Whitney rank-sum nonparametric test [28] on the results from Table 2.

5.2.1 Proof of the Hypothesis That the Minimal Weakly Complete Event Logs Are Smaller Than the Minimal Complete Event Logs

If we denote: X = size of minimal weakly complete logs, and Y_1 = size of minimal complete logs, it is needed to test the null hypothesis that distributions of these two marks (labels) are equal, i.e., $H_0 : F_X = F_{Y1}$, against the alternative hypothesis H_1: "The size of minimal weakly complete logs X is lower than the size

of minimal complete logs Y_1". The corresponding critical area C in this case is in Table 3.

H_0	H_1	C
$F_X = F_{Y1}$	X is lower than Y_1	$z_0 \leq -z_{0.5-\alpha}$

Table 3.

Applying the Wilcoxon-Mann-Whitney test on the obtained experimental analysis results, the following values are obtained:

$$n_1 = 100; \quad n_2 = 100; \quad n = n_1 + n_2 = 200;$$

$$V = 0; \quad E(V) = n_1 n_2/2 = 4\,900.5;$$

$$D(V) = E(V)(n+1)/6 = 162\,533.2;$$

$$z_0 = (V - E(V))/[D(V)]^{\frac{1}{2}} = -12.125.$$

For the level of significance $\alpha = 0.05$, the critical area of this test is $C = (-\infty, -1.645]$. Since the realized value of the test statistic z_0 belongs to the critical area C, the null hypothesis H_0 is rejected in favour of alternative hypothesis H_1. In other words, for the level of significance $\alpha = 0.05$, it can be concluded that the assertion that the size of minimal weakly complete logs is lower than the size of minimal complete logs is statistically significant.

5.2.2 Proof of the Hypothesis That the Minimal Weakly Complete Event Logs Are Smaller Than the Minimal Causally Complete Event Logs

If we denote: X = size of minimal weakly complete logs, and Y_2 = size of minimal causaly complete logs, it is needed to test the null hypothesis that distributions of these two marks (labels) are equal, i.e., $H_0 : F_X = F_{Y2}$, against the alternative hypothesis H_1: "The size of minimal weakly complete logs X is lower than the size of minimal causaly complete logs Y_2". The corresponding critical area C in this case is in Table 4.

H_0	H_1	C
$F_X = F_{Y2}$	X is lower than Y_2	$z_0 \leq -z_{0.5-\alpha}$

Table 4.

Applying the Wilcoxon-Mann-Whitney test on the obtained experimental analysis results, the following values are obtained:

$$n_1 = 100; \quad n_2 = 100; \quad n = n_1 + n_2 = 200;$$

$$V = 35 + 35 + 35 + 35 + 35 = 175; \quad E(V) = n_1 n_2/2 = 5\,000;$$

$$D(V) = E(V)(n+1)/6 = 167\,500;$$
$$z_0 = (V - E(V))/[D(V)]^{\frac{1}{2}} = -11.789.$$

For the level of significance $\alpha = 0.05$, the critical area of this test is $C = (-\infty, -1.645]$. Since the realized value of the test statistic z_0 belongs to the critical area C, the null hypothesis H_0 is rejected in favour of alternative hypothesis H_1. In other words, for the level of significance $\alpha = 0.05$, it can be concluded that the assertion that the size of minimal weakly complete logs is lower than the size of minimal causaly complete logs is statistically significant.

5.2.3 Influence of the Structure of Networks on the Event Log Size

Observing the structure of networks in the considered examples, the experimental analysis has shown that the size of the event log, from which the original networks can be discovered, can depend on the number of parallel branches in the network, as well as on the total number of activities in mutually parallel branches.

In Table 3 the results of the performed experimental analysis are presented, in which considered examples are grouped according to the total number of activities in parallel branches and the number of parallel branches in the network. Considering such groups of examples, sizes minimal complete, minimal causaly complete and minimal weakly complete logs are presented, as well as the difference between them, and the difference between the total number of activities in parallel branches and the number of parallel branches in the network.

From Figure 6 (and from Table 3) it can be seen that the size of minimal complete logs is proportional to the total number of activities in mutually parallel branches. It can also be seen that the number of activities in parallel branches does not affect the size of minimal causaly complete and minimal weakly complete logs.

From Figure 7 (and from Table 3) it can be seen that the number of parallel branches in the network does not affect the size of minimal complete and minimal weakly complete event logs. It can also be seen that the size of minimal causaly complete logs is proportional (nearly equal) to the total number of parallel branches in the network.

From Figure 8 (and from Table 3) it can be seen that the difference between the size of the minimal complete logs and the size of the minimal weakly complete logs, expressed in the number of traces, is proportional to the difference between the total number of activities in parallel branches and the number of parallel branches in the network. The difference between the total number of activities in parallel branches and the number of parallel branches in the network does not affect the relationship between the sizes of the minimal causaly complete and minimal weakly complete event logs.

N	N_a	N_b	$N_a - N_b$	N_{mcl}	N_{mccl}	N_{mwcl}	$N_{mcl} - N_{mwcl}$	$N_{mccl} - N_{mwcl}$
1	2	2	0	2	2	2	0	0
12	3	2	1	3	2	2	1	0
39	3	3	0	4	3	2	2	1
13	4	2	2	4	2	2	2	0
1	4	3	1	4	3	2	2	1
3	4	3	1	5	3	2	3	1
1	4	3	1	6	3	2	4	1
2	4	4	0	5	4	2	3	2
4	5	2	3	5	2	2	3	0
1	5	3	2	5	2	2	3	0
3	5	3	2	5	3	2	3	1
1	5	3	2	5	4	2	3	2
1	5	5	0	6	5	2	4	3
4	6	2	4	6	2	2	4	0
2	6	3	3	6	3	2	4	1
1	6	3	3	7	3	2	5	1
1	6	4	2	8	4	2	6	2
2	7	3	4	7	3	2	5	1
2	7	4	3	8	4	2	6	2
1	8	4	4	9	4	2	7	2
2	8	3	5	10	3	3	7	0
1	8	4	4	9	4	3	6	1
1	9	4	5	9	4	3	6	1
1	9	5	4	10	5	3	7	2
1	10	3	7	11	3	2	9	1

Table 5. The influence of the number of parallel branches in the network and the total number of activities in each parallel branch on the event log size

6 CONCLUSIONS

The examples show that with our modification of the PM discovering technique, we are able to reduce the problem of completeness of logs in parallel processes that occur in the basic α algorithm. That way, we can improve the efficiency of obtaining a block-structured process model, with the meaning that our algorithm can guarantee the discovery of the original model from significantly smaller logs, which do not satisfy the condition of completeness. For that reason, we first defined the causally complete logs [5] and then the weakly complete logs presented in this paper.

The contribution that we have made with the paper goes in two directions. The first is that by defining a new type of event logs – weakly complete event logs, we have made an improvement with regards to overcoming of the conditions of completeness and causall completeness. The weakly complete logs were cre-

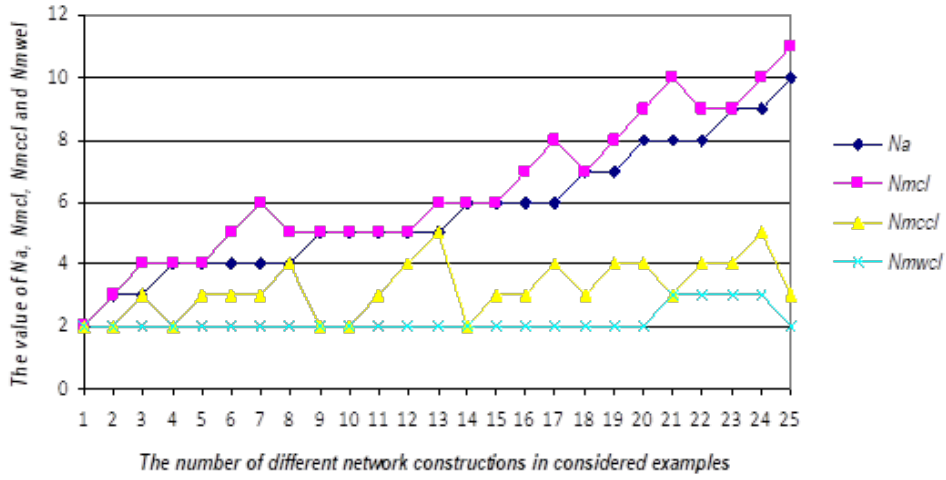


Figure 6. The relation between N_{mcl}, N_{mccl}, N_{mwcl} and N_a

ated as a consequence of our efforts to improve the properties of causally complete logs, primarily in terms of their size. The detailed experimental analysis presented in this paper has just shown that such an improvement has been realized.

Comparative analysis of the results showed that weakly complete event logs can be significantly smaller than both complete and causally complete event logs by the number of traces from which the process model can be discovered.

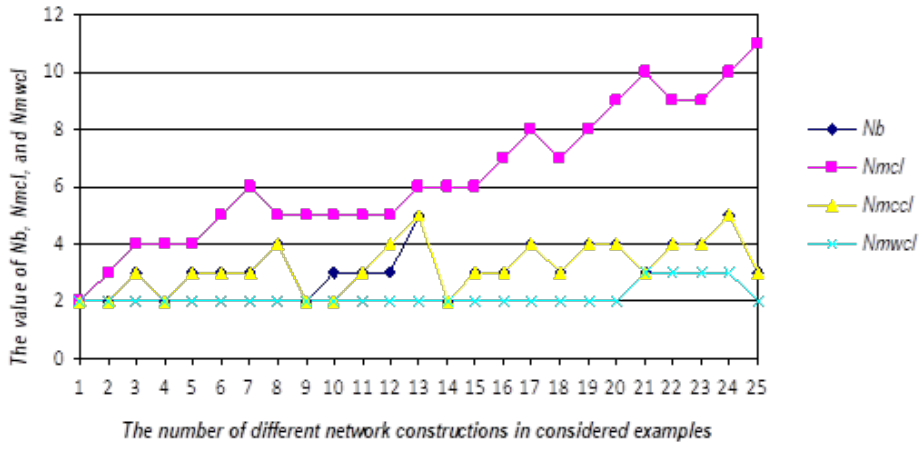


Figure 7. The relation between N_{mcl}, N_{mccl}, N_{mwcl} and N_b

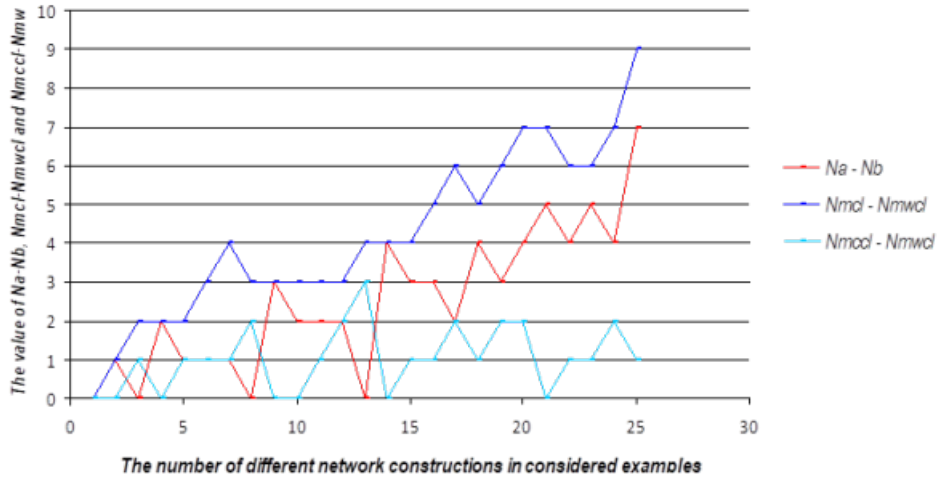


Figure 8. The relation between differences: $N_{mcl} - N_{mwcl}$, $N_{mccl} - N_{mwcl}$ and $N_a - N_b$

The other contribution of defining weakly complete event logs presented in this paper is that they enabled the interactive generation of parallel business process models by demonstration, which was our primary goal. To accomplish our goal, a graphical user interface (GUI) was created, through which the user demonstrates different scenarios of process execution. The graphical user interface that we created is a tool that visually shows steps of $\alpha^{\|}$-algorithm. Such tool could serve as a learning tool and playground for those who want to learn more about how the much better known and more general α-algorithm, which is based on the same principles, functions.

Our assumptions and preconditions for process models (that have to be block-structured parallel models), to which our algorithm and technique are applicable, may look as a serious restriction. However, our solution still covers a respectably wide subclass of process models and represents a first step in a more ambitious attempt to solve the very serious problem of log completeness. In our future research, we will try to expand our work to other categories of processes.

REFERENCES

[1] van der Aalst, W. M. P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011, doi: 10.1007/978-3-642-19345-3.

[2] van der Aalst, W.—Weijters, T.—Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering, Vol. 16, 2004, No. 9, pp. 1128–1142, doi: 10.1109/TKDE.2004.47.

[3] van der Aalst, W. M. P.—Weijters, A. J. M. M.—Maruster, L.: Workflow Mining: Which Processes Can Be Rediscovered? BETA Working Paper Series, Vol. 74, Eindhoven University of Technology, Eindhoven, 2002.

[4] van der Aalst, W. M. P.: Verification of Workflow Nets. In: Azéma, P., Balbo, G. (Eds.): Application and Theory of Petri Nets 1997 (ICATPN 1997). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 1248, 1997, pp. 407–426, doi: 10.1007/3-540-63139-9_48.

[5] Lekić, J.—Milićev, D.: Discovering Block-Structured Parallel Process Models from Causally Complete Event Logs. Journal of Electrical Engineering, Vol. 67, 2016, No. 2, pp. 111–123, doi: 10.1515/jee-2016-0016.

[6] Sun, H.—Du, Y.—Qi, L.—He, Z.: A Method for Mining Process Models with Indirect Dependencies via Petri Nets. IEEE Access, Vol. 7, 2019, pp. 81211–81226, doi: 10.1109/ACCESS.2019.2923624.

[7] Lekić, J.—Milićev, D.—Stanković, D.: Generating Block-Structured Parallel Process Models by Demonstration. Applied Sciences, Vol. 11, 2021, No. 4, Art. No. 1876, doi: 10.3390/app11041876.

[8] Agrawal, R.—Gunopulos, D.—Leymann, F.: Mining Process Models from Workflow Logs. In: Schek, H. J., Alonso, G., Saltor, F., Ramos, I. (Eds.): Advances in Database Technology (EDBT '98). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 1377, 1998, pp. 467–483, doi: 10.1007/BFb0101003.

[9] Cook, J. E.—Wolf, A. L.: Discovering Models of Software Processes from Event-Based Data. ACM Transactions on Software Engineering and Methodology, Vol. 7, 1998, No. 3, pp. 215–249, doi: 10.1145/287000.287001.

[10] Cook, J. E.—Wolf, A. L.: Event-Based Detection of Concurrency. Proceedings of the Sixth International Symposium on the Foundations of Software Engineering (FSE-6), ACM SIGSOFT Software Engineering Notes, Vol. 23, 1998, No. 6, pp. 35–45, doi: 10.1145/291252.288214.

[11] Herbst, J.: Dealing with Concurrency in Workflow Induction. In: Baake, U., Zobel, R., Al-Akaidi, M. (Eds.): Proceedings of the European Concurrent Engineering Conference (ECEC 2000), Leicester, UK, 2000.

[12] Herbst, J.—Karagiannis, D.: Integrating Machine Learning and Workflow Management to Support Acquisition and Adaptation of Workflow Models. Intelligent Systems in Accounting, Finance, and Management, An International Journal, Vol. 9, 2000, No. 2, pp. 67–92, doi: 10.1002/1099-1174(200006)9:2<67::AID-ISAF186>3.0.CO;2-7.

[13] Schimm, G.: Process Miner – A Tool for Mining Process Schemes from Event-Based Data. In: Flesca, S., Greco, S., Ianni, G., Leone, N. (Eds.): Logics in Artificial Intelligence (JELIA 2002). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 2424, 2002, pp. 525–528, doi: 10.1007/3-540-45757-7_47.

[14] van der Aalst, W. M. P.—Stahl, C.: Modeling Business Processes: A Petri Net-Oriented Approach. MIT Press, Cambridge, MA, 2011, doi: 10.7551/mitpress/8811.001.0001.

[15] de Medeiros, A. K. A.—van der Aalst, W. M. P.—Weijters, A. J. M. M.: Workflow Mining: Current Status and Future Directions. In: Meersman, R., Tari, Z., Schmidt, D. C. (Eds.): On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (OTM 2003). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 2888, 2003, pp. 389–406, doi: 10.1007/978-3-540-39964-3_25.

[16] de Medeiros, A. K. A.—Weijters, A. J. M. M.—van der Aalst, W. M. P.: Genetic Process Mining: An Experimental Evaluation. Data Mining and Knowledge Discovery, Vol. 14, 2007, No. 2, pp. 245–304, doi: 10.1007/s10618-006-0061-7.

[17] Wen, L.—van der Aalst, W. M. P.—Wang, J.—Sun, J.: Mining Process Models with Non-Free-Choice Constructs. Data Mining and Knowledge Discovery, Vol. 15, 2007, No. 2, pp. 145–180, doi: 10.1007/s10618-007-0065-y.

[18] van der Aalst, W.—Adriansyah, A.—de Medeiros, A. K. A.—Arcieri, F.—Baier, T. et al.: Process Mining Manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (Eds.): Business Process Management Workshops (BMP 2011). Springer, Berlin, Heidelberg, Lecture Notes in Business Information Processing, Vol. 99, 2012, pp. 169–194, doi: 10.1007/978-3-642-28108-2_19.

[19] Carmona, J.—Cortadella, J.—Kishinevsky, M.: A Region-Based Algorithm for Discovering Petri Nets from Event Logs. In: Dumas, M., Reichert, M., Shan, M. C. (Eds.): Business Process Management (BPM 2008). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 5240, 2008, pp. 358–373, doi: 10.1007/978-3-540-85758-7_26.

[20] van Dongen, B. F.—Alves de Medeiros, A. K.—Wen, L.: Process Mining: Overview and Outlook of Petri Net Discovery Algorithms. In: Jensen, K., van der Aalst, W. M. P. (Eds.): Transactions on Petri Nets and Other Models of Concurrency II. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 5460, 2009, pp. 225–242, doi: 10.1007/978-3-642-00899-3_13.

[21] Günther, C. W.—van der Aalst, W. M. P.: Fuzzy Mining – Adaptive Process Simplification Based on Multi-Perspective Metrics. In: Alonso, G., Dadam, P., Rosemann, M. (Eds.): Business Process Management (BPM 2007). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 4714, 2007, pp. 328–343, doi: 10.1007/978-3-540-75183-0_24.

[22] Weijters, A. J. M. M.—Ribeiro, J. T. S.: Flexible Heuristics Miner (FHM). Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2011), Paris, France, 2011, pp. 310–317, doi: 10.1109/CIDM.2011.5949453.

[23] Leemans, S. J. J.—Fahland, D.—van der Aalst, W. M. P.: Discovering Block-Structured Process Models from Incomplete Event Logs. In: Ciardo, G., Kindler, E. (Eds.): Applications and Theory of Petri Nets and Concurrency (PETRI NETS 2014). Springer, Cham, Lecture Notes in Computer Science, Vol. 8489, 2014, pp. 91–110, doi: 10.1007/978-3-319-07734-5_6.

[24] Lekic, J.—Milicev, D.: Discovering Models of Parallel Workflow Processes from Incomplete Event Logs. In: Hammoudi, S., Pires, L. F., Desfray, P., Filipe, J. (Eds.): Proceedings of the 3rd International Conference on Model-Driven Engineering and Software Development (MODELSWARD 2015), Angers, Loire Valley, France, 2015, pp. 477–482, doi: 10.5220/0005242704770482.

[25] Ling, J. M.—Zhang, L.—Feng, Q.: An Improved Structure-Based Approach to Measure Similarity of Business Process Models. Proceedings of the 26th International Conference on Software Engineering and Knowledge Engineering (SEKE 2014), 2014, pp. 377–380.

[26] van der Aalst, W. M. P.—van Dongen, B. F.—Günther, C. W.—Mans, R. S.—Alves de Medeiros, A. K.—Rozinat, A.—Rubin, V.—Song, M.—Verbeek, H. M. W.—Weijters, A. J. M. M.: ProM 4.0: Comprehensive Support for Real Process Analysis. In: Kleijn, J., Yakovlev, A. (Eds.): Petri Nets and Other Models of Concurrency – ICATPN 2007. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 4546, 2007, pp. 484–494, doi: 10.1007/978-3-540-73094-1_28.

[27] Weakly Complete Event Logs. `https://drive.google.com/drive/u/0/folders/1TKM-6T03i4qZNYZ6a-6yAB7K1x5GHsDD`.

[28] Popovic, C. B.: Mathematical Statistics. Faculty of Sciences and Mathematics, University of Niš, 2009 (in Serbian).

Julijana B. Lekić received her dipl. Eng. degree in electrical engineering from the Faculty of Electrotechnical Engineering, University of Priština (1989), M.Sc. degree from the Faculty of Electrotechnical Engineering, University of Belgrade (1998), and Ph.D. from the Faculty of Technical Sciences, University of Priština (2016), Serbia. She is Assistant Professor at the Faculty of Technical Sciences, University of Priština (temporarily displaced in Kosovska Mitrovica). Her current research interests are in the field of computer science, particularly information systems, software engineering, and business process modeling and mining.

Dragan S. Milićev is Professor at the University of Belgrade, Faculty of Electrical Engineering. He received his dipl. Eng. degree in 1993, M.Sc. in 1995, and Ph.D. in 2001, all from the University of Belgrade. He is specialized in software engineering, model-based engineering, model-driven development, UML, software architecture and design, information systems, and real-time systems. He is a member of the Editorial Board of Springer's Software and Systems Modeling journal (SoSyM). He authored three books on object-oriented programming and UML, published in Serbian, and a book in English, published by Wiley/Wrox, entitled "Model-Driven Development with Executable UML". With thirty years of extensive industrial experience in building complex commercial software systems, he has been serving as the chief software architect, project manager, or consultant in a number of international projects.

Computing and Informatics, Vol. 40, 2021, 368–386, doi: 10.31577/cai_2021_2_368

NON-REDUNDANT IMPLICATIONAL BASE OF MANY-VALUED CONTEXT USING SAT

Taufiq Hidayat

Fakulti Teknologi Maklumat dan Komunikasi (FTMK)
Universiti Teknikal Malaysia Melaka (UTeM)
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
&
Department of Informatics
Islamic University of Indonesia
55584 Yogyakarta, Indonesia
e-mail: p031710045@student.utem.edu.my, taufiq.hidayat@uii.ac.id

Asmala bin Ahmad, Mohammad Ishak bin Desa

Fakulti Teknologi Maklumat dan Komunikasi (FTMK)
Universiti Teknikal Malaysia Melaka (UTeM)
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
e-mail: {asmala, mohammad.ishak}@utem.edu.my

Abstract. Some attribute implications in an implicational base of a derived context of many-valued context can be inferred from some other attribute implications together with its scales. The scales are interpretation of some values in the many-valued context therefore they are a prior or an existing knowledge. In knowledge discovery, the such attribute implications are redundant and cannot be considered as new knowledge. Therefore the attribute implicational should be eliminated. This paper shows that the redundancy problem exists and formalizes a model to check the redundancy.

Keywords: Attribute implication, background knowledge, SAT problem

1 INTRODUCTION

Formal context is a simple data structure, which is defined as a triple (G, M, I) where G is a set of objects, M is a set of attributes, and $I \subseteq G \times M$. If $(g, m) \in I$ where $g \in G$ and $m \in M$ then (g, m) is read as "object g has attribute m" [1, 2]. Figure 1 is an example of formal context represented by a cross table. The formal context is about small natural number. In the formal context,

$$G = \{1, 2, \ldots, 10\},$$

$$M = \{odd, even, greater\ than\ 2, greater\ than\ 5, prime, square\}.$$

	odd	even	greater than 2	greater than 5	prime	square
1	×					×
2		×			×	
3	×		×		×	
4		×	×			×
5	×		×		×	
6		×	×	×		
7	×		×	×	×	
8		×	×	×		
9	×		×	×		×
10		×	×	×		

Figure 1. Formal context of small natural number

Formal context is also able to represent a data table (relational data). A data table will be represented by many-valued context. By scaling, the many-valued context will be transformed into a one-valued context [1, 2, 3]. The one-valued context is called a derived context. In this form, the many-valued context will be analyzed.

Formal Concept Analysis (FCA) is a study to extract knowledge from the formal context. The study is useful in knowledge discovery of data. Three forms of knowledge discovery offered by FCA are clusters (which are called formal concepts), data dependencies (which are called attribute implications), and visualization of formal concepts by single hierarchical diagram (which is called concept lattice) [4]. Many researches are conducted in application of formal concepts analysis to knowledge discovery [5, 6, 7, 8, 9, 10, 11].

An attribute implication of formal context (G, M, I) is an implication in a form $A \rightarrow B$ where $A, B \subseteq M$. The attribute implication means that all objects which have all attributes in A also have all attributes in B. It holds in the formal context (G, M, I) if it holds in each object $g \in G$. These following attribute implications hold in the formal context in Figure 1:

1. $\{even, square\} \Rightarrow \{greater\ than\ 2\}$,
2. $\{prime, greater\ than\ 2\} \Rightarrow \{odd\}$,
3. $\{prime, greater\ than\ 5\} \Rightarrow \{odd\}$.

A set of attribute implications is an implicational base of a formal context (G, M, I) if the attribute implications are sound, complete, and non-redundant with respect to the formal context [2]. There are some algorithms to generate an implicational base.

However, regarding the implicational base, sometimes there are attribute implications which are already known or can be inferred from other attribute implications together with our existing knowledge. We call the existing knowledge as background knowledge. This following simple example illustrates the problem. Recall the formal context in Figure 1. From our knowledge, regarding the formal context we already know that:

1. Every odd number is not even, and every even number is not odd.
2. Every number which is greater than 5 is also greater than 2.

Recall also the three attribute implications holding in the formal context. If we consider the second knowledge, the third attribute implication can be inferred from the first attribute implication together with this knowledge.

An attribute implication could be inferred from other attribute implications with backgroud knowledge considered unimportant knowledge or redundant. Therefore, the attribute implication could be ignored. Ignoring an attribute implication will also reduce the size of knowledge extracted from a formal context to obtain only the important knowledge.

Reducing size of knowledge extracted from a formal context is also a recent issue in this research area because the size is sometime very large. The research in [12] reduced the size by congruent relations whereas in [13] by block relations. A research in [14] summarized this issue and classified all recent techniques in reducing the size of knowledge of concept lattice into 3: redundant information removal, simplification, and selection.

Our research could be considered as another technique in redundant information removal. The redundant information means attribute implications which could be inferred from other atribute implications using background knowledge.

Some recent researches in knowledge discovery and data mining had considered background knowledge to ignore or eliminate extracted knowledge which could be inferred using the background knowledge [15, 16, 17, 18, 19, 20]. The inferred exctracted knowledge is also called redundant knowledge. The redundant knowledge

have to be eliminated since it becomes a handicap and harder for using it in decision making [15, 17, 20, 21].

Regarding to a formal context, the background knowledge relating with formal context exists. A kind of the background knowledge exists in analysis of many-valued context. As stated earlier, a many-value context has to be transformed into a derived context before being analyzed. The transformation process is called scaling. The scaling needs some scales which are one-valued contexts. A scale can be considered as interpretation of attribute values in the many-valued context. Thus, the scales are representations of prior knowledge to the interpretation. Therefore the scales contain some information which can be seen as background knowledge. Interestingly, many sets of data are in the form of many-valued context [8, 9, 10, 11, 22, 23, 24, 25, 26, 27, 28, 29].

Another kind of background knowledge is from our prior knowledge. The kind of background knowledge exists and some researches used it for formal concept analysis [4, 12, 13, 30, 31, 32, 33]. Some of the researches used such background knowledge to remove or reject some extracted knowledges which are incompatible with it [4, 30, 31, 32, 33] where the extracted knowledge is in the form of attribute implications [4, 30] and concepts [31, 32, 33]. The other researches used such background knowledge to reduce the size of extracted knowledge in the form of concept lattice [12, 13].

To know whether an attribute implication of implicational base can be inferred from some other attribute implications using some background knowledge is a hard problem. However, it probably can be solved using SAT approach. The problem will be encoded into SAT Problem and solved by SAT Solver.

SAT Problem (satisfiability problem) is to determine whether a given propositional formula is satisfiable or not. If it is not, we say that the propositional formula is unsatisfiable. A propositional formula is satisfiable if there is an assignment for all propositonal variables in that formula where the assignment makes the evaluation of the formula to true value. If there is no such assignment, the formula is unsatisfiable [34, 35, 36].

Some algorithms have been developed to solve the SAT Problem and implemented in SAT Solver software. The algorithm which is implemented in many modern SAT Solvers is DPLL algorithm [37, 38, 39]. The DPLL algorithm is a backtracking-based algorithm for deciding the satisfiability of propositional formula in conjunctive normal form. It was introduced in 1962 by Martin Davis, Hilary Putnam, George Logemann and Donald W. Loveland [38] and is a refinement of the earlier Davis-Putnam algorithm, which is a resolution-based procedure developed by Davis and Putnam in 1960 [37].

The recent SAT Solvers are able to solve a propositional formula in millions number of both clauses and variables in reasonable time. It gives a chance to make SAT applicable in real world. Therefore, the current researches in the SAT area are not only focusing in the algorithm [40, 41] and solver [42, 43, 44, 45, 46] but also application of SAT [47].

This paper introduces non-redundant implicational base using scales as background knowledge in many-valued context, models the problem, and formalizes it in the satisfiability problem.

2 FOUNDATIONS

2.1 Formal Context

Definition 1 (Formal Context)**.** A **formal context** (G, M, I) consists of two non-empty sets G and M, and a relation $I \subseteq G \times M$. We call the set G a set of objects, whereas the set M a set of attributes. For $g \in G$ and $m \in M$, $(g, m) \in I$ or gIm is read as the object g has the attribute m [1].

A cross table can represent a formal context. The rows of the cross table represent the objects, and the columns represent the attributes. The headers of the rows are object names, whereas the headers of the columns are attribute names. If an object g has an attribute m, then we cross the table in row g and column m. Figure 1 is a formal context in the cross table.

Definition 2 (Derivation Operator)**.** If $A \subseteq G$ is a set of objects, then we define [1]:

$$A^I = \{m \mid (g, m) \in I \text{ for all } g \in A\}. \tag{1}$$

Reversely, if $B \subseteq M$ is a set of attributes, then we define:

$$B^I = \{g \mid (g, m) \in I \text{ for all } m \in B\}. \tag{2}$$

Notation A^{II} refers to $(A^I)^I$.

2.2 Attribute Implication

Let M a set of attributes in (G, M, I). $A \Rightarrow B$ where $A, B \subseteq M$ is an attribute implication over the formal context. The attribute implication holds in the formal context if each object of the formal context respects the attribute implication. An object $g \in G$ respects the attribute implication iff its attributes set is a model of the implication [2].

Definition 3 (Model of Attribute Implication)**.** Let $A, B, T \subseteq M$. T is a **model of attribute implication** $A \Rightarrow B$ iff $A \nsubseteq T$ or $B \subseteq T$ [2].

Definition 4 (Respecting Object)**.** An object $g \in G$ **respects to** $A \Rightarrow B$ over (G, M, I) iff $\{g\}^I$ is a model of the attribute implication. An object $g \in G$ **respects to** a set $\mathcal{L}$ of attribute implications iff g respects all attribute implications in $\mathcal{L}$ [2].

Definition 5 (Holding Attribute Implication)**.** An attribute implication $A \Rightarrow B$ holds in a formal context (G, M, I) iff all $g \in G$ respect the attribute implication.

Algorithm: Implicational Base
Input : A formal context (G,M,I)
Output: The implicational base, $\mathcal{L}$

begin
  $X \leftarrow \emptyset$
  $\mathcal{L} \leftarrow \emptyset$
  repeat
    if $(X \neq X^{II})$ ***then***
      $\mathcal{L} \leftarrow \mathcal{L} \cup \{X \Rightarrow X^{II}/X\}$
    $X \leftarrow$ Next_Closure(X) from $\mathcal{L}$
  until $(X = M)$
  return $\mathcal{L}$
end

Figure 2. Implicational Base algorithm [1, 2]

A set $\mathcal{L}$ of attribute implications holds in a formal context (G, M, I) iff all attribute implications in $\mathcal{L}$ holds in (G, M, I) [2].

Definition 6 (Inference)**.** An implication $A \Rightarrow B$ can be **inferred** from $\mathcal{L}$, denoted by [2]

$$\mathcal{L} \vDash A \Rightarrow B \tag{3}$$

iff all models of $\mathcal{L}$ are also models of $A \Rightarrow B$.

Definition 7 (Implicational Base)**.** A set $\mathcal{L}$ of attribute implications is an **implicational base** of a formal context, if the followings hold [2]:

- **Sound**, if $\mathcal{L}$ holds in the formal context.
- **Complete**, if the following holds. If there is an attribute implication which holds in the formal context, it can be inferred from $\mathcal{L}$.
- **Non-redundant**, if there is no attribute implication in $\mathcal{L}$ that can be inferred from the others.

Figure 2 shows an algorithm to generate an implicational base of a formal context. **Next_Closure(X) from** $\mathcal{L}$ is the lexically smallest model of $\mathcal{L}$ which is lexically larger than X. Let $A, B \subseteq M = \{m_1, m_2, \ldots, m_n\}$ and $m_1 < m_2 < \cdots < m_n$. We define $A < B$, which means "A smaller than B" or "B larger than A", iff $A <_i B$, which is defined as follows, there is i such that

- $i \notin A$ and $i \in B$, and
- for all $j < i$, $j \in A$ iff $j \in B$.

Example 1. Recall a formal context in Figure 1. The implicational base of the formal context generated by algorithm in Figure 2 contains the following attribute implications:

- $\{greater\ than\ 5\} \Rightarrow \{greater\ than\ 2\}$,
- $\{greater\ than\ 2, prime\} \Rightarrow \{odd\}$,
- $\{greater\ than\ 2, greater\ than\ 5, square\} \Rightarrow \{odd\}$,
- $\{odd, prime\} \Rightarrow \{greater\ than\ 2\}$,
- $\{odd, even\} \Rightarrow \{greaterthan\ 2, greaterthan\ 5, prime, square\}$.

2.3 Attribute Implication of Many-Valued Context

Definition 8 (Many-valued Context)**.** A **many-valued context** (G, M, W, I) consists of a set of objects G, a set of attributes M, a set of attribute values W, and a ternary relation $I \subseteq G \times M \times W$ where $(g, m, w) \in I$ and $(g, m, v) \in I$ imply $w = v$ [2, 3].

In the attribute exploration of a many-valued context, we have to transform the many-valued context into one-valued context. The transformation is called scaling. In the scaling, we need some scales, which are also formal contexts [2].

Definition 9 (Scale)**.** A **scale** for attribute $m \in M$ of a many-valued context (G, M, W, I) is a one-valued context $S_m = (G_m, M_m, I_m)$ with $\{w \mid (g, m, w) \in I \text{ and } g \in G\} \subseteq G_m$ [2].

	Final	Written	Practical
1	Pass	Pass	Pass
2	Fail	Pass	Fail
3	Fail	Fail	Pass
4	Fail	Fail	Fail

Figure 3. Many-valued context

	Final:Pass	Final:Fail
Pass	×	
Fail		×

	Written:Pass	Written:Fail
Pass	×	
Fail		×

	Practical:Pass	Practical:Fail
Pass	×	
Fail		×

Figure 4. Scales for attributes: Final, Written, and Practical, respectively

Definition 10 (Derived Context)**.** The **derived context** in scaling of the many-valued context (G, M, W, I) and scales S_m for all $m \in M$ is the context (G, N, J) where

$$N = \bigcup_{m \in M} M_m \tag{4}$$

and for $g \in G$ and $n \in N$, $(g, n) \in J$ iff $(m, g, w) \in I$ and $(w, n) \in I_m$ [2].

Example 2. Figure 3 is an example of a many-valued context with

$$M = \{Final, Written, Practical\}.$$

The many-valued context shows all possible results of driving test. The driving test consists of two parts which are written and practical part showed by attribute $Written$ and $Practical$, respectively. The final result which depends on both test parts is showed by attribute $Final$.

By scaling with a formal context in Figure 4 for all attributes in M, we obtain a derived context in Figure 5.

	Final:Pass	Final:Fail	Written:Pass	Written:Fail	Practical:Pass	Practical:Fail
1	×		×		×	
2		×	×			×
3		×		×	×	
4		×		×		×

Figure 5. The derived context

2.4 SAT Problem

We take some notations from [36, 45] and [48] to formulate the propositional formula and the SAT problem.

A propositional formula is a logical formula based on proposition. An atomic (simple) formula consists of a single propositional variable whereas a complex formula is a composition of connectors and propositional variable(s). The connectors are $\wedge$ (conjunction), $\vee$ (disjunction), $\rightarrow$ (implication), $\leftrightarrow$, (biimplication), and $\neg$ (negation).

Definition 11 (Propositional Formula)**.** A **propositional formula** F is recursively defined as follows:

$$F = \begin{cases} p, \\ \neg F', \\ F_1 \circ F_2, \quad \text{where } \circ \in \{\wedge, \vee, \rightarrow, \leftrightarrow\}, \end{cases}$$

where

- p is a propositional variable, possibly with indices,
- F_1, F_2, and F' are propositional formulas.

Definition 12 (Interpretation)**.** An **interpretation** *Int* is a mapping of propositional formulas to truth values $\{\top, \bot\}$.

An interpretation *Int* will uniquely act on each variable occurring in F. Let p a propositional variable. *Int* will be either $Int(p) = \top$ or $Int(p) = \bot$. An interpretation *Int* will be a model of formula F if and only if $Int(F) = \top$. F is satisfiable if and only if F has some models, and F is unsatisfiable if and only if F has no models.

Given a propositional formula F, the goal of the SAT Problem is to determine whether the formula F is satisfiable or unsatisfiable.

3 BACKGROUND KNOWLEDGE IN MANY-VALUED CONTEXT

3.1 Background-Inferring Problem

Given an attribute implication which holds in a derived context, the question is whether the attribute implication can be implied by the other attribute implications, which also hold in the derived context, together with information in its scales.

Definition 13 (Background-inferring Problem)**.** Scales can be considered as interpretations of values in a many-valued context. Those are already some existing knowledges which are used to derive the many-valued context to obtained a derived one-valued context. The implicational base algorithm in Figure 2 does not considered the existing knowledge. The following shows that an attribute implication probably can be inferred from some others attribute implications in the impliational base together with the knowledge in those scales.

Let $\mathcal{L}$ a set of attributes implications which hold in the derived context from a many-valued context (G, M, W, I) and scales S_m for all $m \in M$, $\mathcal{H}$ knowledge represents the scales, and $A \Rightarrow B$ an attribute implication which also holds in the derived context. The **background-inferring problem** is whether:

$$(\mathcal{L} \cup \mathcal{H}) \text{ implies } A \Rightarrow B. \tag{5}$$

It means that all models of $\mathcal{L}$ and $\mathcal{H}$ are also models of $A \Rightarrow B$. Since a scale $S_m = (G_m, M_m, I_m)$ consists of all possible combination values of attributes in M_m, a model T of $\mathcal{L}$ is also a model of $\mathcal{H}$ iff for each S_m, T is compatible with S_m. T is compatible with S_m iff there is $g \in G_m$ such that $\{g\}^{I_m} \subseteq T$ [30].

Example 3. These attribute implications hold in the derived context showed in Figure 5:

- {Practical:Fail} ⇒ {Final:Fail},
- {Written:Fail} ⇒ {Final:Fail},
- {Written:Pass, Practical:Pass} ⇒ {Final:Pass},
- {Final:Fail, Practical:Pass} ⇒ {Written:Fail}.

Let $\mathcal{L}$ consist of the three first-attribute-implications and $\mathcal{H}$ represent information from scales in Figure 4. All models of $\mathcal{L}$ containing {Final:Fail, Practical:Pass} are

- {Final:Fail, Final:Pass, Practical:Pass, Written:Pass}, and
- {Final:Fail, Practical:Pass, Written:Fail}.

Because of the scale of attribute Practical (Figure 4), the first model is not the model of $\mathcal{H}$. Thus, only the second model is the model of $(\mathcal{L} \cup \mathcal{H})$. It is also a model of

- {Final:Fail, Practical:Pass} ⇒ {Written:Fail}.

Therefore, $(\mathcal{L} \cup \mathcal{H})$ implies the attribute implication.

For the next examples, we will use the natural numbers 1, 2, ... to refer attribute names Final:Pass, Final:Fail, ..., respectively.

4 BACKGROUND-INFERRING PROBLEM IN SAT

The followings are some corresponding notations between formal context and propositional formula in this encoding:

- An attribute $m \in M$ corresponds to a propositional variable p_m.
- $T \subseteq M$ corresponds to an interpretation Int_T. $m \in T$ iff $Int_T(p_m) = \top$.

Proposition 1. Let $T, A, B \subseteq M$. T is a model of $A \Rightarrow B$ iff

$$Int_T \left(\bigwedge_{b \in B} \left(\left(\bigwedge_{a \in A} p_a \right) \rightarrow p_b \right) \right) = \top.$$

Proof.

1. T is a model of $A \Rightarrow B$. There are two possibilities:

   (a) $A \nsubseteq T$
   $\hookrightarrow$ there is $c \in A$, but $c \notin T$
   $\hookrightarrow$ $Int_T(p_c) = \bot$
   $\hookrightarrow$ $Int_T\left(\bigwedge_{a\in A} p_a\right) = \bot$
   $\hookrightarrow$ For all $b \in B$, $Int_T\left(\left(\bigwedge_{a\in A} p_a\right) \rightarrow p_b\right) = \top$
   $\hookrightarrow$ $Int_T\left(\bigwedge_{b\in B}\left(\left(\bigwedge_{a\in A} p_a\right) \rightarrow p_b\right)\right) = \top$

   (b) $B \subseteq T$
   $\hookrightarrow$ For all $b \in B$, $Int_T(p_b) = \top$
   $\hookrightarrow$ For all $b \in B$, $Int_T\left(\left(\bigwedge_{a\in A} p_a\right) \rightarrow p_b\right) = \top$
   $\hookrightarrow$ $Int_T\left(\bigwedge_{b\in B}\left(\left(\bigwedge_{a\in A} p_a\right) \rightarrow p_b\right)\right) = \top$

2. $Int_T\left(\bigwedge_{b\in B}\left(\left(\bigwedge_{a\in A} p_a\right) \rightarrow p_b\right)\right) = \top$
   $\hookrightarrow$ For all $b \in B$, $Int_T\left(\left(\bigwedge_{a\in A} p_a\right) \rightarrow p_b\right) = \top$
   $\hookrightarrow$ There are also two possibilities:

   (a) For all $b \in B$, $Int_T(p_b) = \top$
   $\hookrightarrow$ $B \subseteq T$
   $\hookrightarrow$ T is a model of $A \Rightarrow B$

   (b) $Int_T\left(\bigwedge_{a\in A} p_a\right) = \bot$
   $\hookrightarrow$ There is $c \in A$, such that $Int_T(p_c) = \bot$
   $\hookrightarrow$ There is $c \in A$, but $c \notin T$
   $\hookrightarrow$ $A \nsubseteq T$
   $\hookrightarrow$ T is a model of $A \Rightarrow B$.

□

From Proposition 1, $A \Rightarrow B$ corresponds to a propositional formula:

$$\bigwedge_{b\in B}\left(\left(\bigwedge_{a\in A} p_a\right) \rightarrow p_b\right).$$

We will use $F_{A\Rightarrow B}$ to refer the formula.

Example 4. Recall Example 3. We have the following correspond formulas, respectively:

- $p_6 \rightarrow p_2$,
- $p_4 \rightarrow p_2$,
- $(p_3 \wedge p_5) \rightarrow p_1$,
- $(p_2 \wedge p_5) \rightarrow p_4$.

Proposition 2. Let $S_m = (G_m, M_m, I_m)$ a scale to obtain a derived context (G, N, J) and $T \subseteq N$. T is compatible with S_m iff

$$Int_T \left(\bigvee_{g \in G_m} \left(\bigwedge_{a \in \{g\}^{I_m}} p_a \wedge \bigwedge_{a \in M_m / \{g\}^{I_m}} \neg p_a \right) \right) = \top.$$

Proof.

1. T is compatible with $S_m = (G_m, M_m, I_m)$
   $\hookrightarrow$ There is $g_c \in G_m$, such that $\{g_c\}^{I_m} \subseteq T$
   $\hookrightarrow$ $Int_T(\bigwedge_{a \in \{g_c\}^{I_m}} p_a \wedge \bigwedge_{a \in M_m / \{g_c\}^{I_m}} \neg p_a) = \top$
   $\hookrightarrow$ $Int_T \left(\bigvee_{g \in G_m} (\bigwedge_{a \in \{g\}^{I_m}} p_a \wedge \bigwedge_{a \in M_m / \{g\}^{I_m}} \neg p_a) \right) = \top$
2. $Int_T \left(\bigvee_{g \in G_m} (\bigwedge_{a \in \{g\}^{I_m}} p_a \wedge \bigwedge_{a \in M_m / \{g\}^{I_m}} \neg p_a) \right) = \top$
   $\hookrightarrow$ There is $g_c \in G_m$, such that $Int_T(\bigwedge_{a \in \{g_c\}^{I_m}} p_a \wedge \bigwedge_{a \in M_m / \{g_c\}^{I_m}} \neg p_a) = \top$
   $\hookrightarrow$ $\{g_c\}^{I_m} \subseteq T$
   $\hookrightarrow$ T is compatible with $S_m = (G_m, M_m, I_m)$.

□

From Proposition 2, we know that the information related with a scale $S_m = (G_m, M_m, I_m)$ corresponds to a propositional formula:

$$\bigvee_{g \in G_m} \left(\bigwedge_{a \in \{g\}^{I_m}} p_a \wedge \bigwedge_{a \in M_m / \{g\}^{I_m}} \neg p_a \right).$$

We will use H_m to refer the propositional formula which a scale S_m corresponds to.

Example 5. Recall Example 3. From scale of attribute Final, Written, and Practical in Figure 4, we have the following formulas:

- $(p_1 \wedge \neg p_2) \vee (\neg p_1 \wedge p_2)$,
- $(p_3 \wedge \neg p_4) \vee (\neg p_3 \wedge p_4)$,
- $(p_5 \wedge \neg p_6) \vee (\neg p_5 \wedge p_6)$.

Proposition 3. T is a model of a set of attribute implications $\mathcal{L}$, iff

$$Int_T \left(\bigwedge_{A \Rightarrow B \in \mathcal{L}} F_{A \Rightarrow B} \right) = \top. \tag{6}$$

Proof. T is a model of $\mathcal{L}$
iff For all $A \Rightarrow B \in \mathcal{L}$, T is also a model of $A \Rightarrow B$
iff For all $A \Rightarrow B \in \mathcal{L}$, $Int_T(F_{A \Rightarrow B}) = \top$ {from Proposition 1}
iff $Int_T \left(\bigwedge_{A \Rightarrow B \in \mathcal{L}} F_{A \Rightarrow B} \right) = \top$. □

Proposition 4. T is a model of $\mathcal{H}$, which is information representing scales $S_m = (G_m, M_m, I_m)$ for all $m \in M$, iff

$$Int_T\left(\bigwedge_{m\in M} H_m\right) = \top. \tag{7}$$

Proof. T is a model of $\mathcal{H}$
iff For all $m \in M$, T is compatible with $S_m = (G_m, M_m, I_m)$
iff For all $m \in M$, $Int_T(H_m) = \top$ {from Proposition 2}
iff $Int_T\left(\bigwedge_{m\in M} H_m\right) = \top$. □

Let $F_{\mathcal{L}} = \bigwedge_{A\Rightarrow B\in\mathcal{L}} F_{A\Rightarrow B}$ and $F_{\mathcal{H}} = \bigwedge_{m\in M} H_m$. $\mathcal{L}$ corresponds to $F_{\mathcal{L}}$, whereas $\mathcal{H}$ corresponds to $F_{\mathcal{H}}$.

Proposition 5. T is a model of $(\mathcal{L} \cup \mathcal{H})$ iff $Int_T(F_{\mathcal{L}} \wedge F_{\mathcal{H}}) = \top$.

Proof. T is a model of $(\mathcal{L} \cup \mathcal{H})$
iff T is a model of both $\mathcal{L}$ and $\mathcal{H}$
iff $Int_T(F_{\mathcal{L}}) = \top$ and $Int_T(F_{\mathcal{H}}) = \top$ {from Proposition 3 and Proposition 4}
iff $Int_T(F_{\mathcal{L}} \wedge F_{\mathcal{H}}) = \top$. □

Proposition 6. $(\mathcal{L} \cup \mathcal{H})$ does not imply $A \Rightarrow B$, iff $F_{\mathcal{L}} \wedge F_{\mathcal{H}} \wedge \neg F_{A\Rightarrow B}$ is satisfiable.

Proof. $(\mathcal{L} \cup \mathcal{H})$ does not imply $A \Rightarrow B$
iff There is $T \in M$, T is a model of $(\mathcal{L} \cup \mathcal{H})$, but T is not a model of $A \Rightarrow B$
iff There is $T \in M$, $Int_T(F_{\mathcal{L}} \wedge F_{\mathcal{H}}) = \top$ (Proposition 5) and $Int_T(F_{A\Rightarrow B}) = \bot$ (Proposition 1)
iff There is $T \in M$, $Int_T(F_{\mathcal{L}} \wedge F_{\mathcal{H}} \wedge \neg F_{A\Rightarrow B}) = \top$
iff $F_{\mathcal{L}} \wedge F_{\mathcal{H}} \wedge \neg F_{A\Rightarrow B}$ is satisfiable. □

Example 6. Recall Example 3, 4, and 5. Let $\mathcal{L} = \{\{6\} \Rightarrow \{2\}, \{4\} \Rightarrow \{2\}, \{3,5\} \Rightarrow \{1\}\}$ and $\mathcal{H}$ information from scales in Figure 4. We want to check whether $(\mathcal{L} \cup \mathcal{H})$ does not imply $\{2,5\} \Rightarrow \{4\}$. Then, we obtain the following propositional formula:

1. $p_6 \rightarrow p_2$,
2. $\wedge p_4 \rightarrow p_2$,
3. $\wedge(p_3 \wedge p_5) \rightarrow p_1$,
4. $\wedge((p_1 \wedge \neg p_2) \vee (\neg p_1 \wedge p_2))$,
5. $\wedge((p_3 \wedge \neg p_4) \vee (\neg p_3 \wedge p_4))$,
6. $\wedge((p_5 \wedge \neg p_6) \vee (\neg p_5 \wedge p_6))$,
7. $\wedge\neg((p_2 \wedge p_5) \rightarrow p_4)$.

Let F be the propositional formula. If we consider the conjunct 4 then we only have two possible interpretations e.g. Int_{T_1} and Int_{T_2}, where:

- $Int_{T_1}(p_1) = \top$ and $Int_{T_1}(p_2) = \bot$,
- $Int_{T_2}(p_1) = \bot$ and $Int_{T_2}(p_2) = \top$.

$Int_{T_1}(F) = \bot$ since $Int_{T_1}(\neg((p_2 \wedge p_5) \rightarrow p_4)) = \bot$ (conjunct 7).

Whereas Int_{T_2} will be a model of F, if $Int_{T_2}(p_3) = \bot$ or $Int_{T_2}(p_5) = \bot$ because of conjunct 3. Suppose $Int_{T_2}(p_3) = \bot$. Because of conjunct 5, $Int_{T_2}(p_4) = \top$. It makes Int_{T_2} over conjunct 7 be $\bot$. Thus, $Int_{T_2}(F) = \bot$.

Also Int_{T_2} over conjuct 7 will be $\bot$ if $Int_{T_2}(p_5) = \bot$.

We can conclude that neither Int_{T_1} nor Int_{T_2} will be a model of F. Thus, F is unsatisfiable. Therefore, $(\mathcal{L} \cup \mathcal{H})$ implies $\{2, 5\} \Rightarrow \{4\}$. It is the same conclusion obtained in Example 3.

5 CONCLUSION

We showed that some attribute implications in an implicational base of derived context of many-valued context can be inferred from the many-valued context's scales. Even though, the scales are intepretation of some values in the many-valued context, therefore the scales are an existing knowledge. Some literatures proposed that knowledge in knowledge discovery from data, an implicational base in case of a formal context, which can be inferred from existing or background knowledge should be eliminated. They will be redundant knowledge.

We also formalized a model to check the redundancy in SAT Problem. The formulation has also been proven.

In the next research we will develop an algorithm to obtain non-redundant implicational base of many-valued context using scales as background knowledge based on the proposed model. Some experiments with real data also will be conducted using the algorithm.

REFERENCES

[1] Ganter, B.—Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag, Berlin, Germany, 1999, doi: 10.1007/978-3-642-59830-2.

[2] Ganter, B.: Formal Concept Analysis: Algorithmic Aspects. Available at: `http://www.math.tu-dresden.de/~ganter/cl03/cl02.pdf`, 2002.

[3] Gugisch, R.: Many-Valued Context Analysis Using Descriptions. In: Delugach, H. S., Stumme, G. (Eds.): Conceptual Structures: Broadening the Base (ICCS 2001). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 2120, 2001, pp. 157–168, doi: 10.1007/3-540-44583-8_12.

[4] Bělohlávek, R.—Vychodil, V.: Formal Concept Analysis with Background Knowledge: Attribute Priorities. IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol. 39, 2009, No. 4, pp. 399–409, doi: 10.1109/TSMCC.2008.2012168.

[5] Xu, Y.—Wang, K.—Zhang, B.—Chen, Z.: Privacy-Enhancing Personalized Web Search. Proceedings of the 16th International Conference on World

Wide Web (WWW '07), Alberta, Canada, ACM, 2007, pp. 591–600, doi: 10.1145/1242572.1242652.

[6] Kumar, C. A.: Knowledge Discovery in Data Using Formal Concept Analysis and Random Projections. International Journal of Applied Mathematics and Computer Science, Vol. 21, 2011, No. 4, pp. 745–756, doi: 10.2478/v10006-011-0059-1.

[7] Klimeš, J.: Using Formal Concept Analysis for Control in Cyber-Physical Systems. Procedia Engineering, Vol. 69, 2014, pp. 1518–1522, doi: 10.1016/j.proeng.2014.03.149.

[8] Kaytoue, M.—Kuznetsov, S. O.—Napoli, A.—Duplessis, S.: Mining Gene Expression Data with Pattern Structures in Formal Concept Analysis. Information Sciences, Vol. 181, 2011, No. 10, pp. 1989–2001, doi: 10.1016/j.ins.2010.07.007.

[9] Du, Y.—Li, H.: Strategy for Mining Association Rules for Web Pages Based on Formal Concept Analysis. Applied Soft Computing, Vol. 10, 2010, No. 3, pp. 772–783, doi: 10.1016/j.asoc.2009.09.007.

[10] Lee, C.—Jeon, J.—Park, Y.: Monitoring Trends of Technological Changes Based on the Dynamic Patent Lattice: A Modified Formal Concept Analysis Approach. Technological Forecasting and Social Change, Vol. 78, 2011, No. 4, pp. 690–702, doi: 10.1016/J.TECHFORE.2010.11.010.

[11] Beydoun, G.: Formal Concept Analysis for an e-Learning Semantic Web. Expert Systems with Applications, Vol. 36, 2009, No. 8, pp. 10 952–10 961, doi: 10.1016/j.eswa.2009.02.023.

[12] Viaud, J.-F.—Bertet, K.—Missaoui, R.—Demko, C.: Using Congruence Relations to Extract Knowledge from Concept Lattices. Discrete Applied Mathematics, Vol. 249, 2018, pp. 135–150, doi: 10.1016/j.dam.2016.11.021.

[13] Konecny, J.—Krupka, M.: Block Relations in Formal Fuzzy Concept Analysis. International Journal of Approximate Reasoning, Vol. 73, 2016, pp. 27–55, doi: 10.1016/j.ijar.2016.02.004.

[14] Dias, S. M.—Vieira, N. J.: Concept Lattices Reduction: Definition, Analysis and Classification. Expert Systems with Applications, Vol. 42, 2015, No. 20, pp. 7084–7097, doi: 10.1016/j.eswa.2015.04.044.

[15] Diaz, J.—Molina, C.—Vila, M. A.: A Model for Redundancy Reduction in Multidimensional Association Rules. In: dos Reis, A. P., Abraham, A.P. (Eds.): Proceedings of IADIS European Conference Data Mining 2013 (part of MCCSIS 2013). IADIS Press, Lisbon, Portugal, 2013, pp. 89–93.

[16] Nguyen, D.—Nguyen, L. T. T.—Vo, B.—Hong, T.-P.: A Novel Method for Constrained Class Association Rule Mining. Information Sciences, Vol. 320, 2015, pp. 107–125, doi: 10.1016/j.ins.2015.05.006.

[17] Ashrafi, M. Z.—Taniar, D.—Smith-Miles, K.: A New Approach of Eliminating Redundant Association Rules. In: Galindo, F., Takizawa, M., Traunmüller, R. (Eds.): Database and Expert Systems Applications (DEXA 2004). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 3180, 2004, pp. 465–474, doi: 10.1007/978-3-540-30075-5_45.

[18] Baralis, E.—Cagliero, L.—Cerquitelli, T.—Garza, P.: Generalized Association Rule Mining with Constraints. Information Sciences, Vol. 194, 2012, pp. 68–84, doi: 10.1016/j.ins.2011.05.016.

[19] Chandanan, A. K.—Shukla, M. K.: Removal of Duplicate Rules for Association Rule Mining from Multilevel Dataset. Procedia Computer Science, Vol. 45, 2015, pp. 143–149, doi: 10.1016/j.procs.2015.03.106.

[20] Diaz, J.—Molina, C.—Vila, M. A.: Using Imprecise User Knowledge to Reduce Redundancy in Association Rules. Proceedings of the 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology, Advances in Intelligent Systems Research Series, 2015, pp. 1098–1105, doi: 10.2991/ifsa-eusflat-15.2015.155.

[21] Gondek, D.—Hofmann, T.: Non-Redundant Data Clustering. Knowledge and Information Systems, Vol. 12, 2007, No. 1, pp. 1–24, doi: 10.1007/s10115-006-0009-7.

[22] Muangprathub, J.—Boonjing, V.—Pattaraintakorn, P.: A New Case-Based Classification Using Incremental Concept Lattice Knowledge. Data and Knowledge Engineering, Vol. 83, 2013, pp. 39–53, doi: 10.1016/j.datak.2012.10.001.

[23] Kaytoue-Uberall, M.—Duplessis, S.—Napoli, A.: Using Formal Concept Analysis for the Extraction of Groups of Co-Expressed Genes. In: Le Thi, H. A., Bouvry, P., Pham Dinh, T. (Eds.): Modelling, Computation and Optimization in Information Systems and Management Sciences (MCO 2008). Springer, Berlin, Heidelberg, Communications in Computer and Information Science, Vol. 14, 2008, pp. 439–449, doi: 10.1007/978-3-540-87477-5_47.

[24] Castellanos, A.—Cigarrán, J.—García-Serrano, A.: Formal Concept Analysis for Topic Detection: A Clustering Quality Experimental Analysis. Information Systems, Vol. 66, 2017, pp. 24–42, doi: 10.1016/j.is.2017.01.008.

[25] Cigarrán, J.—Castellanos, Á.—García-Serrano, A.: A Step Forward for Topic Detection in Twitter: An FCA-Based Approach. Expert Systems with Applications, Vol. 57, 2016, pp. 21–36, doi: 10.1016/j.eswa.2016.03.011.

[26] Fu, G.: FCA Based Ontology Development for Data Integration. Information Processing and Management, Vol. 52, No. 5, 2016, pp. 765–782, doi: 10.1016/j.ipm.2016.02.003.

[27] Valverde-Albacete, F. J.—González-Calabozo, J. M.—Peñas, A.—Peláez-Moreno, C.: Supporting Scientific Knowledge Discovery with Extended, Generalized Formal Concept Analysis. Expert Systems with Applications, Vol. 44, 2016, pp. 198–216, doi: 10.1016/j.eswa.2015.09.022.

[28] Lihonosova, A.—Kaminskaya, A.: Using Formal Concept Analysis for Finding the Closest Relatives Among a Group of Organisms. Procedia Computer Science, Vol. 31, 2014, pp. 860–868, doi: 10.1016/j.procs.2014.05.337.

[29] Keller, B. J.—Eichinger, F.—Kretzler, M.: Formal Concept Analysis of Disease Similarity. Proceedings of the 2012 AMIA Joint Summits on Translational Science, 2012, pp. 42–51.

[30] Ganter, B.: Attribute Exploration with Background Knowledge. Theoretical Computer Science, Vol. 217, 1996, No. 2, pp. 215–233, doi: 10.1016/S0304-3975(98)00271-0.

[31] Bělohlávek, R.—Sklenář, V.—Zacpal, J.: Concept Lattices Constrained by Attribute Dependencies. In: Pokorný, J., Richta, K. (Eds.): Proceedings of the Dateso 2004 Annual International Workshop on Databases, Texts, Specifications and Objects, Desna, Czech Republic, 2004. CEUR Workshop Proceedings, Vol. 98, 2004, pp. 63–73.

[32] Bělohlávek, R.—Sklenář, V.: Formal Concept Analysis Constrained by Attribute-Dependency Formulas. In: Ganter, B., Godin, R. (Eds.): Formal Concept Analysis (ICFCA 2005). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 3403, 2005, pp. 176–191, doi: 10.1007/978-3-540-32262-7_12.

[33] Belohlavek, R.—Vychodil, V.: Adding Background Knowledge to Formal Concept Analysis via Attribute Dependency Formulas. Proceedings of the 2008 ACM Symposium on Applied Computing (SAC '08), Fortaleza, Ceara, Brazil, 2008, pp. 938–943, doi: 10.1145/1363686.1363900.

[34] Posthoff, C.—Steinbach, B.: SAT-Problems – New Findings. Proceedings of 9th WSEAS International Conference on Data Networks, Communications, Computers, WSEAS Press, 2007, pp. 339–344.

[35] Posadas-Causor, A.—Torres-Jimenez, J.: SAT-DB: An Integrated System for Satisfiability Problem Study. Proceedings of 6th WSEAS International Multiconference on Circuits, Systems, Communications and Computers, WSEAS Press, 2002, pp. 4951–4956.

[36] Biere, A.—Heule, M.—van Maaren, H.—Walsh, T.: Handbook of Satisfiability. Amsterdam, IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 185, 2009, 980 pp.

[37] Davis, M.—Putnam, H.: A Computing Procedure for Quantification Theory. Journal of the ACM, Vol. 7, 1960, No. 3, pp. 201–215, doi: 10.1145/321033.321034.

[38] Davis, M.—Logemann, G.—Loveland, D.: A Machine Program for Theorem-Proving. Communications of the ACM, Vol. 5, 1962, No. 7, pp. 394–397, doi: 10.1145/368273.368557.

[39] Nieuwenhuis, R.—Oliveras, A.—Tinelli, C.: Solving SAT and SAT Modulo Theories: From an Abstract Davis–Putnam–Logemann–Loveland Procedure to DPLL(T). Journal of the ACM, Vol. 53, 2006, No. 6, pp. 937–977, doi: 10.1145/1217856.1217859.

[40] Martinez-Rios, F.—Frausto-Solis, J.: An Hybrid Simulated Annealing Threshold Accepting Algorithm for Satisfiability Problems Using Dynamically Cooling Schemes. WSEAS Transactions on Computers, Vol. 7, 2008, No. 5, pp. 374–386.

[41] Vázquez-Morán, I.—Torres-Jiménez, J.: A SAT Instances Construction Based on Hypergraphs. Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA '03), 2003, pp. 244–247.

[42] Moskewicz, M. W.—Madigan, C. F.—Zhao, Y.—Zhang, L.—Malik, S.: Chaff: Engineering an Efficient SAT Solver. Proceedings of the 38th Annual Design Automation Conference, ACM, 2001, pp. 530–535, doi: 10.1145/378239.379017.

[43] Manthey, N.: Coprocessor – A Standalone SAT Preprocessor. In: Tompits, H. et al. (Eds.): Applications of Declarative Programming and Knowledge Management (INAP 2011, WLP 2011). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7773, 2013, pp. 297–304, doi: 10.1007/978-3-642-41524-1_18.

[44] Manthey, N.—Saptawijaya, A.: Towards Improving the Resource Usage of SAT-Solvers. In: Le Berre, D. (Ed.): POS-10. Pragmatics of SAT, Edinburgh, UK, 2010, EasyChair, Vol. 8, 2012, pp. 28–40, doi: 10.29007/3vwv.

[45] Hyvärinen, A. E. J.—Manthey, N.: Designing Scalable Parallel SAT Solvers. In: Cimatti, A., Sebastiani, R. (Eds.): Theory and Applications of Satisfiability Testing (SAT 2012). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7317, 2012, pp. 214–227, doi: 10.1007/978-3-642-31612-8_17.

[46] Manthey, N.: Towards Next Generation Sequential and Parallel SAT Solvers. KI – Künstliche Intelligenz, Vol. 30, 2016, No. 3, pp. 339–342, doi: 10.1007/s13218-015-0406-8.

[47] Lynce, I.—Ouaknine, J.: Sudoku as a SAT Problem. Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics (AIMATH 06). Springer, Berlin, 2006.

[48] Huth, M.—Ryan, M.: Logic in Computer Science: Modelling and Reasoning about Systems. 2nd Edition. Cambridge University Press, Cambridge, 2004, doi: 10.1017/cbo9780511810275.

Taufiq Hidayat is Ph.D. candidate at the Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. He currently works as Junior Lecturer at Universitas Islam Indonesia, Indonesia. His main research areas include formal concept analysis, mathematical logic, machine learning and data mining.

Asmala bin Ahmad is Associate Professor at the Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. He also serves as the research coordinator for the faculty. His research interest includes remote sensing, image processing, artificial intelligence and applied mathematics.

Mohammad Ishak bin Desa is Professor at the Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. His main research areas include operations research, computational intelligence, data science and analytics.

Computing and Informatics, Vol. 40, 2021, 387–407, doi: 10.31577/cai_2021_2_387

HUMAN POSE ESTIMATION USING PER-POINT BODY REGION ASSIGNMENT

Dana Škorvánková, Martin Madaras

Faculty of Mathematics, Physics and Informatics
Comenius University
Mlynská dolina F1, 842 48 Bratislava, Slovakia
e-mail: {dana.skorvankova, madaras}@fmph.uniba.sk

Abstract. In recent years, the task of human pose estimation has become increasingly important, due to the large scale of usage, including VR applications, as well as higher-level tasks, such as human behavior understanding. In this paper, we introduce a novel two-stage deep learning approach named Segmentation-Guided Pose Estimation (SGPE). The pipeline is based on two neural networks working in a sequential fashion, while both models effectively process unorganized point clouds on the input. First, the segmentation network performs a pointwise classification into the corresponding body regions. In the next step, the point cloud with the per-point region assignment, forming the fourth input channel, is passed to the regression network. This way, both local and global features of the point cloud are preserved, helping the model fully maintain the body pose structure. Our strategy achieves competitive results on all of the examined benchmark datasets, and outperforms state-of-the-art methods.

Keywords: Machine vision, deep learning, neural networks, pose estimation, point clouds

Mathematics Subject Classification 2010: 68-T45

1 INTRODUCTION

One of many fields where the neural networks are applicable is the human motion analysis. Some of the most frequent motion tasks include skeleton tracking, human motion prediction and pose estimation. The motion tasks using either data-based or

physics-based methods still remain a challenge these days. The data-driven methods rely mostly on motion capture systems, while the physics-based methods depend on optimization to predict motion. The task of human pose estimation attracts a lot of attention among deep learning researchers, mainly because of its frequent usage in virtual and augmented reality, ergonomic body posture analysis, action recognition, surveillance, human-robot interaction, trajectory prediction or motion-based human identification. Although a lot has been achieved in the 3D human pose estimation task, there are still many challenges nowadays, which are not easy to overcome.

Analyzing previous human pose estimation methods based on deep learning, the pipeline is usually formed by passing a single 2D image to the network, which directly regresses the 3D skeletal joint coordinates. Single-person pose estimation forms a basis for a number of related tasks, such as multi-person pose estimation [3, 18, 28], pose tracking [38, 40] or video pose estimation [24]. Most of the research is currently focused on estimating the pose from RGB data [2, 18, 25, 28, 33], mainly due to easily obtainable data which can be captured using a conventional RGB camera, without requiring a special hardware setup for recording. On the other hand, methods processing depth data on the input proved to be beneficial in terms of accuracy, by providing the additional spatial information.

Since most of the research is currently focused on estimating the pose from RGB data [5, 17, 18, 20, 23, 28], one of the most critical challenges of pose estimation from 3D input is data availability. To successfully train a neural network of reasonable size, a large and well labeled dataset is crucial. Currently, there is a very limited set of publicly available 3D human pose estimation databases. Moreover, even among the available datasets, it is hard to find one that is both large enough in its scale, and accurate enough to avoid overfitting of the neural model. There are several large action recognition datasets with motion capture ground truth, but since providing the exact skeleton joint locations is not their primal purpose, the ground truth is often not accurate enough for the task of pose estimation. Due to the lack of the accessible depth data, many researchers have recently used their own recorded depth datasets to evaluate the results of their proposed method. However, this leads to the fact, that it is difficult to objectively compare the particular methods, because the recorded databases are often not published. It is important to mention that recording of a quality depth dataset is not a trivial task, mainly since the expensive motion capture system is usually required to obtain accurate ground truth labels, which also limits us to indoor scenes. The usual workaround is to use the Kinect camera for recording, which can also directly extract the 3D skeleton joint coordinates, even though still working well only in indoor scenes.

Another issue concerning pose estimation from 3D data is the actual type of 3D data that is passed as input to the neural network. The most frequent option is to use depth maps [14, 15, 30, 39], thus encoding the third dimension into the 2D image. The depth maps are the very dense representation of a human pose, which results in expensive computations and lowering the time efficiency, while also processing the

seemingly redundant data. Furthermore, since depth maps are usually treated by neural networks as 2D images, there arises the same problem as in estimating 3D pose from RGB data, i.e. the need for highly non-linear operations. Additionally, because of the projection of an object in 3D space onto a 2D image plane, the actual shape of the human pose can be distorted in the depth map, which means the network has to perform the perspective distortion-invariant estimation [21]. In an attempt to overcome these drawbacks, voxelized grids have been used in several solutions [9, 13, 21] to provide sparser 3D data representation. Despite that, voxels have their shortcomings, too. First of all, voxels require 3D convolution operations, which are rather demanding in terms of memory, time, and computing power. Moreover, the conversion of point clouds or depth maps into the voxelized grids can be time-consuming itself.

Sparser 3D representations of the human pose, like voxels or point clouds, are usually employed to perform the classification, segmentation, or related tasks. They are rarely used in pose estimation, mainly because the common 2D convolutions cannot be used on this type of data in the same way as on RGB or depth images. Treating point clouds as unorganized sets of points, this type of data can be processed inside the network either by extracting features for each point separately, which yields exclusively local information, or by aggregating the features of all points, which gives us global information about the whole point cloud. Alternatively, the data can be clustered in particular point sets, which are treated as local regions [37]. While in the classification tasks, the global features are those needed to predict the correct class scores, both local and global information is essential in pose estimation task. Hence, the main issue with performing local context-driven tasks on point clouds is often related to poor propagation of local features inside the network.

Our work solves the task of single-person human pose estimation from depth data using a novel two-stage deep learning method called Segmentation-Guided Pose Estimation (SGPE). To avoid the projection of 3D human pose to 2D image space, we employ unorganized and unordered point clouds on the input to compute 3D skeletal joint coordinates as a result. We enhance the local and global feature propagation by performing an auxiliary semantic segmentation into the body regions. First, a corresponding body region is assigned to each point of the point cloud in a segmentation stage. To enable the network to fully perceive the data in its local as well as global context, we also make use of the intermediate concatenation of pointwise and aggregated features inside the model. Second, the input point cloud containing the point coordinates is concatenated with the per-point body region labels, adding the fourth channel to the data. Afterwards, the four-channel point clouds are fed into the regression model, which is where the resulting joint coordinates are estimated as an output. The main contributions of this work can be summarized as follows.

- We cope with the excessive number of network parameters and computational cost by processing depth data in a form of sparse unordered point clouds, instead

of the commonly used depth maps. This way, we also avoid the need for the model to perform a distortion-invariant estimation.

- Our two-stage pipeline deals with the issues related to poor propagation of local context through the networks, by concatenation of features extracted in intermediate layers before and after pooling aggregation, and by incorporating residual connections in-between the layers. Thus, we improve the gradient flow inside the models. Furthermore, to increase the accuracy of estimated joint coordinates, we augment the initial 3D point clouds with a per-point body region segmentation predicted in the first stage of the pipeline.
- To evaluate our approach, we conduct experiments on a number of depth-based human pose benchmark datasets, including both synthetic and real data. Our strategy achieves competitive results on all of the examined datasets, and outperforms state-of-the-art methods.

2 RELATED WORK

Nowadays, neural networks are widely used in the field of image processing, pattern recognition, human movement analysis and many more. There are numerous types of tasks concerning human movement analysis, where the neural networks proved to be beneficial, e.g. action recognition [36], action classification, body-movement-based human identification, pose estimation etc. Focusing on the pose estimation task, there have been many different methods and approaches presented in recent years. Based on the type of the input data, the studies can be divided into approaches inferring from two-dimensional data (RGB images) [5, 17, 18, 20, 23, 28, 33, 42], and three-dimensional data (depth maps, point clouds, voxelized grids etc.) [1, 6, 7, 8, 11, 15, 21, 31, 32, 39, 41]. The two-dimensional approaches are far more usable and easily accessible in real-time applications, being able to run without any special devices, using only the RGB camera. On the other hand, the regression of 3D joint positions from 2D input data requires highly non-linear operations, which can lead to many difficulties in the learning procedure. The three-dimensional approaches provide the additional depth information, which can significantly simplify the task for the network, and thus improve the estimation accuracy.

2.1 Human Pose Estimation from RGB Data

We can divide studies working with the RGB input data in two main groups based on whether they directly regress the 3D pose coordinates [16, 34] or use the 2D pose to infer the 3D pose [4, 17, 19, 20, 28]. Among those employing the 2D pose, many approaches make use of lifting the estimated 2D pose to 3D [4, 10, 16, 22] by direct regression, database matching etc.

One of the first real-time approaches was proposed by Mehta et al. [20]. They introduced a system to obtain real-time full global 3D skeletal pose, combining

a pose regressor based on convolutional neural network with kinematic skeleton fitting. They parametrized each 3D skeletal joint by a confidence heatmap and three location maps, one for each axis. However, the stated model was unable to handle occlusions. Thus, they removed the restrictions in the follow-up work [18], where the model is also extended to capture multiple people in the scene by a single RGB camera. Unlike the previous work, the model outputs full skeletal pose in joint angles and global body positions of a coherent skeleton in real-time.

2.2 Depth-Based Human Pose Estimation

The depth data used as the input to the neural networks comes in various forms. Most frequently, the 3D input data is in a form of a depth map (RGB-D image). Depth maps are actually encoding 3D space into 2D image, where the value at each pixel position represents the corresponding depth value (third axis coordinate). Marin-Jimenez et al. [15] proposed a technique where the final estimated pose is computed as the weighted sum of the predefined set of prototype poses. The weights corresponding to the prototypes are directly regressed from input depth maps by a convolutional neural network. The stated approach is an example of a single-stage method.

The two-stage methods generally consists of the segmentation stage and the regression stage. First, the input data is segmented to the corresponding body-parts. Then, the segmented input data is used to infer 3D joint coordinates. An example of a two-stage method was proposed by Shafaei and Little [31]. They treat the problem of 3D pose estimation from depth data through a two-stage pipeline, where in the first stage the body parts are identified in the input depth maps by a dense classifier. In the second stage, all camera views are merged, and a set of statistics concerning a created unified 3D point cloud is collected and passed as features to a linear regressor to compute 3D body joint locations.

Aside from depth maps, some of the methods make use of the voxelized grids, made by discretizing a given point cloud in a predefined set of values. However, voxels require use of three-dimensional convolutions, which makes operations with them very time-consuming and computationally expensive. *V2V PoseNet* [21] operates with this kind of data and regresses joint locations with 3D CNN-autoencoders. They first use 3D CNN encoder and decoder to estimate per-voxel likelihood of each skeleton joint from voxelized input. Afterwards, they refine the target object localization with a 2D CNN which takes a cropped depth map and output an offset from its reference point to the center of ground truth joint positions. This way, they obtain an accurate reference point.

2.3 Point Cloud Input Data

As an alternative to depth images or voxels, there are several networks proposed which work directly with unordered point clouds as input data, yet implement the convolution operations on the point clouds without using computationally expensive

3D convolutions. Some of the methods decided to use shared multi-layer perceptrons and max-pooling layers to obtain the features of a point cloud. Although they manage to extract global features, since the max-pooling layers are applied on the whole set of points, it is hard to capture the local context. Qi et al. [26] proposed a classification and segmentation model called *PointNet*, where they intend to incorporate the local features by an aggregation of the intermediate outputs from the classification network, before and after max-pooling. Afterwards, they fed the aggregated local and global features into the segmentation network. Later, Qi et al. [27] introduced *PointNet++* model, which has similar key structure as the previous PointNet, but it improves the model by utilizing a hierarchical structure, similar to the one used in image processing convolutional neural networks. It recursively applies PointNet on a nested partitioning of the input point cloud, starting from small local patches and gradually extending to bigger regions.

In another study, Wu et al. [37] presented a new convolution operation called *PointConv*, which can be applied on unordered and irregular point clouds. They treat convolution kernels as nonlinear weight and density functions of the local coordinates of 3D points. The weight functions are learned with multi-layer perceptron networks and density functions through kernel density estimation. Such learned kernels can be used for translation-invariant and permutation-invariant convolutions on any 3D point set.

It is worth mentioning, that all of the stated methods processing unordered point clouds perform object classification or segmentation task, which is not an aim of this work. Concerning pose estimation task, Ali [1] introduced a novel one-stage approach in his thesis, called *Point-Based Pose Estimation* (PBPE), using point clouds directly as input data to the model which outputs 3D skeleton joint coordinates. He concludes, that since point clouds are able to provide sparser representation of the human body, compared to depth maps, the operations on them would be much easier, and thus, the computational complexity would be reduced. The inspiration for the model was in the PointNet architecture. Besides the proposed PBPE model, the contribution of his work also consists of the refinement of several two-stage methods by using an automatic annotation mechanism for labeling body regions in real data. Next, the study presents the benefits of fusion of the real training data and more complex synthetic training data. The poses in the synthetic dataset are much more varied, so by adding certain amount of the synthetic data to the real dataset during the training phase, they extend the diversity of the training set. As a result, the model is able to generalize better. On the other hand, the synthetic data is also useful for pre-training a model, reducing the computational cost and time of the real data annotation. Thus, such pre-trained model can be fine-tuned on a relatively small part of the real dataset, yet achieving reasonable results.

As a part of our previous research, we re-implemented the method from [1], while slightly modifying the model architecture to improve the final estimations.

We enhanced the part of the network which extracts local features of the input point cloud, and reduced the amount of batch normalization in the model.

In this paper, we solve the problem of depth-based human pose estimation using unordered point clouds as the input data type. However, unlike the previous approaches processing point clouds, our pipeline works in two subsequent stages, instead of a direct regression, to effectively merge both local and global features of the data without losing any contextual information. Thus, the resulting pose coordinates can be regressed from a point cloud enhanced by additional regional information, helping the network fully maintain the body pose structure.

3 OVERVIEW

We introduce the Segmentation-Guided Pose Estimation (SGPE) – a two-stage pipeline which takes a point cloud as an input, and outputs the 3D coordinates of the estimated skeletal joint positions. Incorporating the idea of handling unorganized and permutation-invariant point clouds, both stages of the pipeline are based on pseudo-convolutions, which operate in the filter dimension. The first stage of our pipeline involves a segmentation network, which classifies the points representing a human pose into the corresponding body regions. In the second stage, the original input point cloud containing the point coordinates is concatenated with the output regions from the segmentation network, thus forming a four-channel point cloud input. Such produced data, conserving together the local as well as the global information, is then fed into the second model – the regression network, where the joint coordinates are finally regressed. The architecture of both networks, as depicted in Figure 1, makes use of residual connections added to the shared multi-layer perceptron blocks, to strengthen the feature propagation.

4 SEGMENTATION-GUIDED POSE ESTIMATION

This section describes our proposed method in detail, providing further information on the training procedures. Our pipeline takes a point cloud on the input, passes it through two subsequent neural networks, and outputs the 3D coordinates of the skeletal joints, defining the estimated human pose. Prior to sending the input point cloud to the first neural network, the background scene is segmented out – the ground floor and the surrounding walls are removed using RANSAC plane fitting algorithm, and the biggest cluster of the point cloud is extracted, being considered the captured human subject. To unify the dimension of the model input, the point cloud is subsampled to a fixed number of points using the farthest point sampling. We set the hyperparameter determining the number of points in each point cloud to $p = 2\,048$, yielding a fair density of the input data. Both the ground truth skeleton coordinates, as well as the input point clouds, are normalized to the range

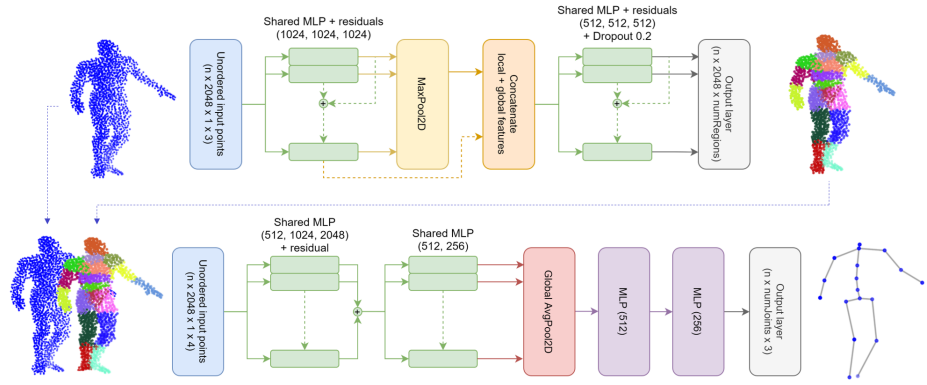


Figure 1. The overview of the proposed Segmentation-Guided Pose Estimation pipeline: First, the point clouds are segmented into body regions in the segmentation network (*top*), then the input point clouds are concatenated with the predicted per-point body region assignment as a fourth channel, and fed into the regression network (*bottom*)

$[-1, 1]$ along each axis, using minimum and maximum values of the whole training set.

4.1 Shared Multi-Layer Perceptron Module

The shared multi-layer perceptron (MLP), introduced in [26], is a stack of convolutional layers with kernel size 1×1. Unlike the standard convolutional layers, they do not affect the dimension of the input, but instead expand (or shrink) the dimension of the filters. By operating in the filter space, the 1×1 convolutions allow us to process unorganized and permutation-invariant sets of points. The points passed to the shared MLP module are treated as 2D input with dimensions 1×3.

4.2 Body Region Segmentation Network

As a part of our pipeline, we propose a segmentation network with an architecture similar to the one of the regression model, instead of making use of one of the existing segmentation methods (e.g. U-net [29] or PointNet [26]). Instead of using an exhausting segmentation architecture, which has high memory and time requirements, we decided to utilize the same main modules in the segmentation and regression model. This is partly because segmentation is not the main task of this work, and is strictly in role of an auxiliary subtask, therefore the absolute segmentation accuracy is not crucial in our study. Also, we believe preserving a similar network-specific rep-

resentation of the body pose in both models works for the benefit of more accurate pose estimation.

In the first stage, the pre-processed point clouds are fed into the segmentation network, which performs a pointwise classification into the corresponding body regions. The architecture of the model, as shown in Figure 1 (top), is based on the shared multi-layer perceptron modules. To obtain global features, the output vector of the first shared MLP is aggregated in a pooling layer across all points of the point cloud. Since the local information is essential in the task of semantic segmentation as well, we want to avoid losing the local context after the max pooling aggregation. Therefore, the local features extracted from the intermediate layers of the shared MLP are concatenated with the aggregated global features and sent off to the second shared MLP module. After the second shared MLP, the model outputs the predicted per-point classification probabilities for each body region.

In order to help the gradient flow, and enhance the feature propagation, we improved the shared MLP modules in our approach by adding residual connections in-between the convolutional layers. Referring to the figure, the numbers in the brackets near the shared MLP blocks describe the number of filters in the respective 1×1 convolutional layers.

Since the real data does not come with body-parts segmentation, we perform an automatic annotation of the point clouds to acquire ground truth body region classification of the data. The number of regions matches the number of joints in skeleton, each region being associated with the particular joint. Every single point of the point cloud is then assigned to the region corresponding to the nearest skeleton node in terms of Euclidean distance.

4.3 Regression Network

The second stage of our pipeline is based on the regression network. To incorporate the idea of retaining both local and global context of the input point clouds, the initial 3D point cloud is concatenated with the predicted pointwise region assignment after the body region segmentation, forming a four-channel input point cloud, which is passed to the regression model (as indicated in Figure 1, bottom). Again, the network incorporates two shared MLP blocks. The first one contains three convolutional layers with 1×1 kernels, followed by one residual connection adding up the outputs of the three preceding layers. To control the number of parameters of the network, the second shared MLP includes two layers and no additional skip connections. To avoid having majority of the model parameters concentrated in the first fully-connected layer, the global average pooling is utilized instead of a simple flattening layer to spatially average across all points right before the fully-connected layers. Finally, the model estimates the 3D skeletal joint coordinates of the captured human subject as the output.

5 RESULTS

5.1 Benchmark Datasets

ITOP. The ITOP dataset [8] contains 40 K training and 10 K testing depth frames recorded from two viewpoints (front-view and top-view). The dataset captures 20 different subjects, each performing 15 sequences. The ground truth skeleton is defined by 3D coordinates of 15 skeletal joints.

UBC3V. The UBC3V [31] is a synthetically made human pose dataset. It contains around 6 M synthetic depth frames structured in three parts according to the complexity of the human postures – easy, medium and hard pose, each with its train, validation and test split. The pose in each frame is represented by the position of 18 skeletal joints. It captures a total of 16 characters and each frame is observed from three different viewpoints.

MHAD. The MHAD dataset [35] consists of 11 actions performed by 7 male and 5 female subjects. Each subject performed each of the actions 5 times, which yields about 660 action sequences corresponding to about 82 minutes of total recording time. The total number of depth frames is over 250 K. The skeleton structure in this dataset contains 35 joints.

CMU Panoptic dataset. The CMU Panoptic [12] is a large scale multi-modal human pose dataset containing video recordings from 480 VGA cameras and more than 30 HD cameras, RGB and depth data from 10 Kinect v2 sensors, and 3D body poses. The full dataset yields around 6 hours of recordings. The synchronization of the devices is hardware-based, although, as the authors state in the database description, there is no way to perfectly synchronize multiple Kinects. However, most of the data is aligned accurately by hardware modifications for time-stamping. The skeleton structure consists of 15 joint locations. The database captures multiple actors of different gender, age and body shape.

5.2 Evaluation Metrics

In the process of evaluation, we used mean per joint position error (MPJPE) and mean average precision (mAP) as metrics, following [1, 8, 15, 31]. Mean average precision is defined as percentage of all skeletal joints predicted under 10 cm threshold from ground truth.

5.3 Implementation Details

We conduct experiments on NVIDIA GTX 1070. Both networks are trained using the Adam optimizer with the initial learning rate equal to 10^{-3}, and an exponential decay rate of $d = 0.2$ applied at the end of each epoch. All weights are initial-

ized with Xavier normal initializer. The batch size is fixed to $b = 32$ for both models.

Regarding segmentation network, the categorical cross-entropy is employed as a loss function, to measure the accuracy of the body part classification. In the case of the regression network, mean absolute error between the predicted locations and the ground truth labels of all skeletal joints is used to determine the model loss. We have also evaluated the performance of the regression network using huber loss with a regularization term, yielding approximately the same estimation accuracy.

For the regularization purposes, a single dropout layer with rate of 0.2 is included before the output layer of the segmentation network (as shown in Figure 1 (top)).

5.4 Experiments

For the purpose of evaluation, we used several benchmark datasets, including the challenging ITOP front-view [8], UBC3V hard-pose [31], MHAD [35] and a subset of CMU Panoptic dataset [12]. On a test set of the ITOP front-view dataset, the mean per joint position error our method achieves is 6.40 cm (as shown in Figure 5, left). Using a 10 cm threshold, the mean average precision is 85.57 %, which is comparable to the state-of-the-art results.

Regarding the CMU dataset, we evaluated our method specifically on the *Range of motion* section of the dataset, yielding approximately 141 K frames, as it was the only section capturing a single person, having ground truth labels available at the time of this research. Since prior to our work, there was no protocol established for the utilized section of the dataset, and considering the amount of data in the selected section of the dataset, we marked 20 % of the data obtained by random sampling as the test set. There are also no existing results to compare to, concerning the single person pose estimation on this dataset (up to our knowledge). The mean per joint position error using our proposed approach is 2.11 cm (as shown in Figure 5, right), and the mean average precision at 10 cm is 98.39 %. Figure 2 illustrates the qualitative results on samples from CMU Panoptic dataset.

Similarly, the MHAD dataset does not originally come with a train and test split, thus we carried out experiments using two different protocols:

1. choosing the test set as randomly sampled 25 % of the dataset,
2. leave-one-subject-out cross validation.

In case of MHAD data, the original skeleton is rather complex, containing as many as 35 skeleton nodes. We have slightly modified the original skeleton structure by removing several redundant joints – one pair repeated at fingertips, two additional pairs present at toe tips. This way, we restricted the skeleton to the resulting 29 joints (as shown in Figure 3), in the same way as in [1]. However, we present

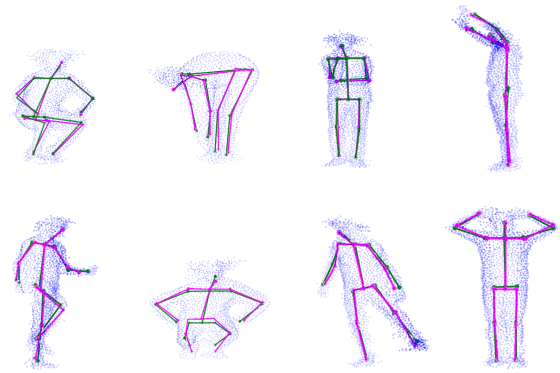


Figure 2. Qualitative results of our method on CMU Panoptic dataset [12]. The ground truth skeletons (*green*) vs. our estimation (*magenta*). Best viewed in color.

results of our approach also on the original full skeleton, to be able to compare our strategy to the existing methods (as shown in Table 1). Since in the case of the modified skeleton we have only removed the redundant skeletal nodes, we have not reduced the complexity of the skeleton in a significant manner, but rather increased the focus on more relevant joints in the skeleton. As it can be seen in Table 1, the mean per joint position error has visibly decreased after omitting the redundant skeletal joints.

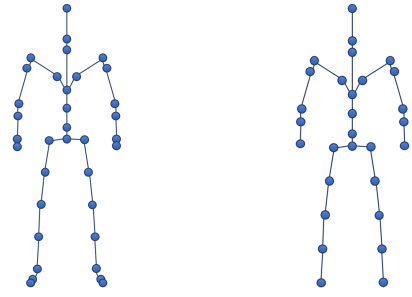


Figure 3. The original skeleton structure used in MHAD dataset (*left*) vs. the modified skeleton (*right*)

Following the first protocol, i.e. establishing the test set as 25 % of the data by random sampling, our method achieves the mean per joint position error as low as 1.39 cm for the multi-view approach, and 1.59 cm for the single-view approach (as shown in Figure 4, left), when using the modified skeleton structure. The achieved mean average precision at 10 cm is as high as 99.80 % and 99.21 % for the multi-view and single-view approach, respectively (Figure 6). We set a novel state-of-

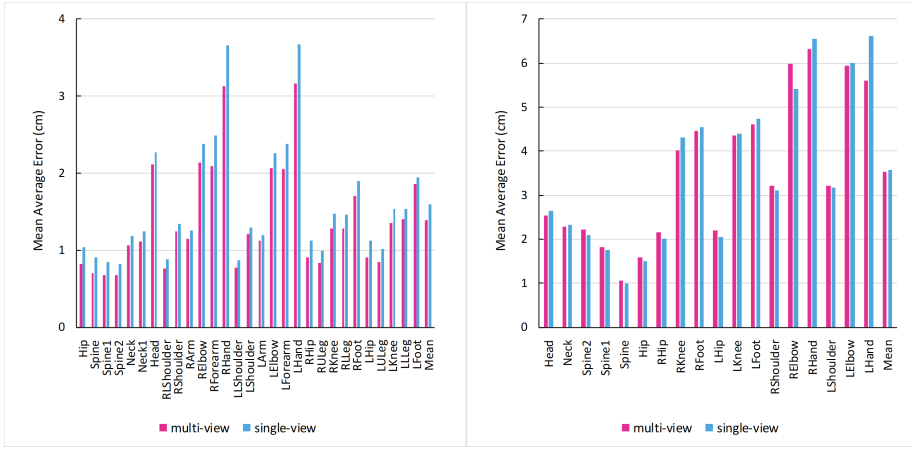


Figure 4. The mean per joint position error (MPJPE) on MHAD (*left*) and UBC3V (*right*) datasets, comparing multi-view and single-view approach

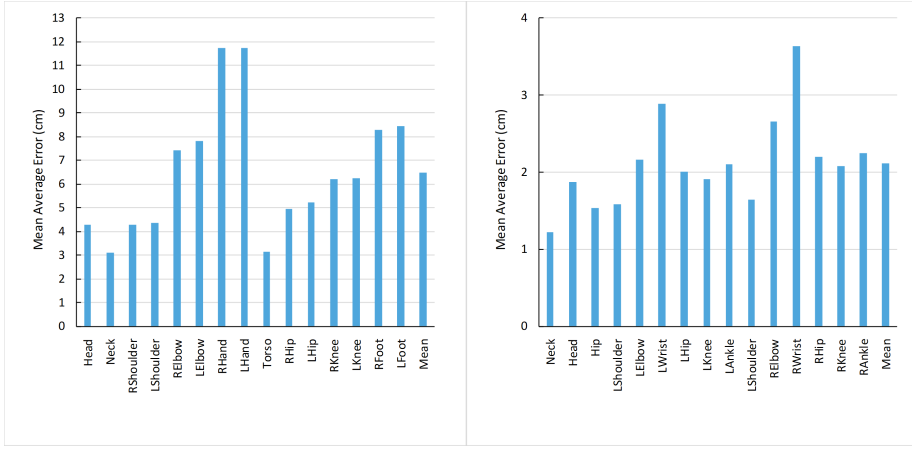


Figure 5. The mean per joint position error (MPJPE) on ITOP (*left*) and CMU (*right*) datasets

the-art for MHAD dataset, lowering the mean per joint position error by almost 65 % following the multi-view approach, and by approximately 50 % following the single-view approach.

Table 2 summarizes the mean per joint position error on UBC3V hard-pose dataset for both single-view and multi-view approach. Using our approach, the achieved mean per joint position error is 3.36 cm in the case of single-view data, and 3.53 cm with multi-view data (as shown in Figure 4, right). The mean average precision at 10 cm is 95.63 % and 95.71 % for the single-view and multi-view

Method	Eval. Protocol	MPJPE [cm] Single-View	MPJPE [cm] Multi-View
Shafei et. al [31]	LOSO	–	5.01
PBPE [1]	random 25 %	7.46	3.92
PBPE [1] (29 joints)	random 25 %	3.20	–
Ours – FCPE	LOSO	3.97	3.36
Ours – FCPE (29 joints)	LOSO	3.23	2.97
Ours – FCPE	random 25 %	1.85	1.62
Ours – FCPE (29 joints)	random 25 %	**1.59**	**1.39**

Table 1. The mean per joint position error (MPJPE) of our approach on MHAD dataset evaluated following the leave-one-subject-out (LOSO) cross validation strategy, as well as randomly sampled test set, compared to state-of-the-art methods

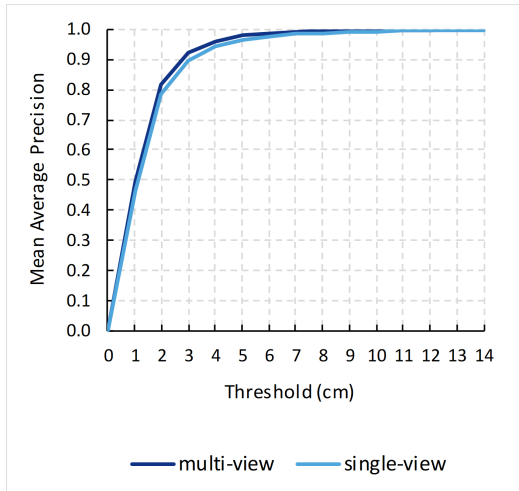


Figure 6. Mean average precision at 10 cm threshold on MHAD dataset for multi-view and single-view approaches

approach respectively. The claimed results of the Deep Depth Pose (DDP) model proposed in [15] are listed in italics, due to a number of unsuccessful attemps to reproduce them by various researchers. The observed results on the reproduced DDP model, implemented following the same training procedures as the original implementation, are indicated in the table as well. Sample qualitative results on UBC3V hard-pose test set are shown in Figure 7, predicted on merged multi-view point clouds.

We also present evaluation of the first stage of our pipeline. The accuracy of the semantic segmentation into the corresponding body regions over training epochs for all examined datasets is depicted in Figure 8. Our method achieves up to 95 % segmentation accuracy on CMU Panoptic dataset.

Method	MPJPE (cm) Single-View	MPJPE (cm) Multi-View
DDP (observed)	19.23	–
PBPE [1]	7.59	5.59
Shafei et. al [31]	–	5.64
DDP (claimed) [15]	*3.15*	*2.36*
Ours – FCPE	**3.57**	**3.53**

Table 2. The mean per joint position error (MPJPE) of the proposed method on the test set of the UBC3V hard-pose dataset compared to state-of-the-art methods

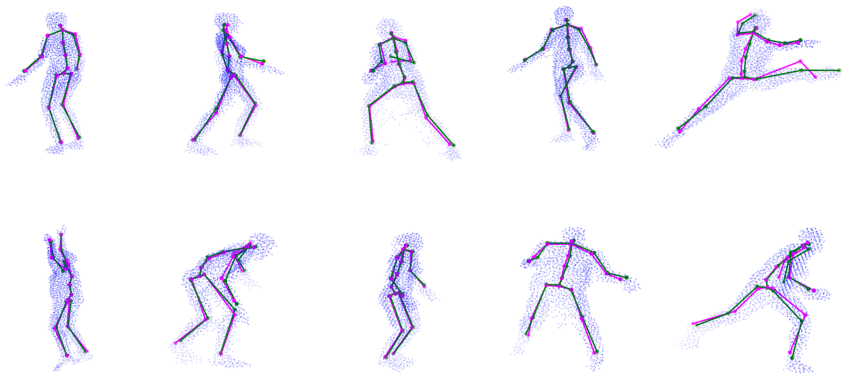


Figure 7. Qualitative results of our approach on test set of UBC hard-pose dataset [31]. The ground truth skeletons (*green*) vs. our estimation (*magenta*). Best viewed in color.

6 LIMITATIONS

We consider an important part of this study to point out the most relevant limitations we encountered during the experiments. Regarding the depth-based human pose estimation, we see the biggest shortage in the range and accuracy of the available datasets. The suitable public datasets, containing both depth data of a captured human subject and the ground truth skeletal joint coordinates, are either too small to be used as training data for a neural network, or the accuracy of the ground truth labels is not sufficient. Moreover, even in large datasets, the data is often incomplete for certain sections, so the valid subset of the dataset ends up of a too small range after all. The limited accuracy of the ground truth poses is usually caused by poor synchronization of a depth sensor and a motion capture system. The most commonly used depth sensors do not have a stable frame rate, which results in time delays and misalignment between frames, and makes the precise synchronization practically impossible. In some of the datasets, this issue is partly fixed by time-stamping technique, refining the frame alignment, and filtering out the mismatches. It is even harder considering the multi-view approach, when the multiple

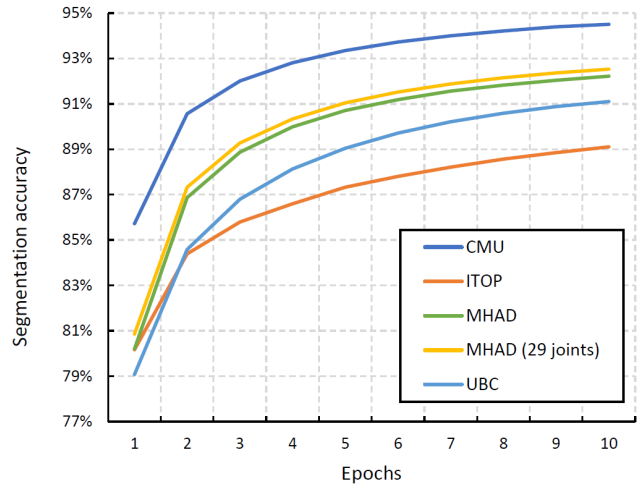


Figure 8. Accuracy of the body-parts segmentation performed in the first stage of our pipeline on all examined datasets

depth sensors need to be synchronized mutually as well as with the motion capture system.

7 CONCLUSIONS

We proposed a novel method for the accurate single-person depth-based human pose estimation called Segmentation-Guided Pose Estimation (SGPE). Main contribution of our work is the elimination of drawbacks related to the projection of 3D space to a 2D image, when estimating pose from depth maps, by introducing a concept of unordered point clouds as a permutation-invariant input to a neural network. To allow the network to maintain both local and global contextual information, we employ intermediate concatenation of extracted pointwise and aggregated features inside the model. Additionally, we perform semantic segmentation of the input point cloud into the corresponding body regions, and utilize the per-point region assignment as an extend of the input point cloud before the final regression. We believe engaging sparse point clouds as an input to the neural network instead of the commonly used depth maps allows us to provide a representation of the human body that is easier to be perceived by the network, while lowering memory requirements and computational cost at the same time. Moreover, to help preserve gradient flow throughout the entire depth of the network, we improved the shared multi-layer perceptron modules by additional skip-connections. Our strategy achieves competitive results on a number of benchmark datasets, and outperforms state-of-the-art approaches.

Acknowledgement

Our research was supported from the grant No. UK/91/2021.

REFERENCES

[1] Ali, A.: 3D Human Pose Estimation. M.Sc. Thesis, Georgia Institute of Technology, May 2019.

[2] Artacho, B.—Savakis, A.: UniPose: Unified Human Pose Estimation in Single Images and Videos. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 7033–7042, doi: 10.1109/CVPR42600.2020.00706.

[3] Briq, R.—Doering, A.—Gall, J.: Unifying Part Detection and Association for Recurrent Multi-Person Pose Estimation. CoRR, 2019, arXiv: 1904.11864.

[4] Chen, C. H.—Ramanan, D.: 3D Human Pose Estimation = 2D Pose Estimation + Matching. CoRR, 2016, arXiv: 1612.06524.

[5] Chou, C. J.—Chien, J. T.—Chen, H. T.: Self Adversarial Training for Human Pose Estimation. CoRR, 2017, arXiv: 1707.02439.

[6] Ge, L.—Liang, H.—Yuan, J.—Thalmann, D.: 3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 5679–5688, doi: 10.1109/CVPR.2017.602.

[7] Ge, L.—Liang, H.—Yuan, J.—Thalmann, D.: Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 3593–3601, doi: 10.1109/cvpr.2016.391.

[8] Haque, A.—Peng, B.—Luo, Z.—Alahi, A.—Yeung, S.—Fei-Fei, L.: Towards Viewpoint Invariant 3D Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.): Computer Vision – ECCV 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9905, 2016, pp. 160–177, doi: 10.1007/978-3-319-46448-0_10.

[9] Huang, F.—Zeng, A.—Liu, M.—Qin, J.—Xu, Q.: Structure-Aware 3D Hourglass Network for Hand Pose Estimation from Single Depth Image. CoRR, 2018, arXiv: 1812.10320.

[10] Iqbal, U.—Doering, A.—Yasin, H.—Krüger, B.—Weber, A.—Gall, J.: A Dual-Source Approach for 3D Human Pose Estimation from a Single Image. CoRR, 2017, arXiv: 1705.02883.

[11] Jiu, M.—Wolf, C.—Taylor, G.—Baskurt, A.: Human Body Part Estimation from Depth Images via Spatially-Constrained Deep Learning. Pattern Recognition Letters, Vol. 50 (C), 2014, pp. 122–129, doi: 10.1016/j.patrec.2013.09.021.

[12] Joo, H.—Simon, T.—Li, X.—Liu, H.—Tan, L.—Gui, L.—Banerjee, S.—Godisart, T. S.—Nabbe, B.—Matthews, I.—Kanade, T.—Nobuhara, S.—Sheikh, Y.: Panoptic Studio: A Massively Multiview System for Social Interaction

Capture. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, 2017, No. 1, pp. 190–204, doi: 10.1109/TPAMI.2017.2782743.

[13] Lekhwani, R.—Singh, B.: FastV2C-HandNet: Fast Voxel to Coordinate Hand Pose Estimation with 3D Convolutional Neural Networks. CoRR, 2019, arXiv: 1907.06327.

[14] Malik, J.—Elhayek, A.—Stricker, D.: Structure-Aware 3D Hand Pose Regression from a Single Depth Image. In: Bourdot, P., Cobb, S., Interrante, V., Kato, H., Stricker, D. (Eds.): Virtual Reality and Augmented Reality (EuroVR 2018). Springer, Cham, Lecture Notes in Computer Science, Vol. 11162, 2018, pp. 3–17, doi: 10.1007/978-3-030-01790-3_1.

[15] Marín-Jiménez, M. J.—Romero-Ramirez, F. J.—Muñoz-Salinas, R.—Medina-Carnicer, R.: 3D Human Pose Estimation from Depth Maps Using a Deep Combination of Poses. Journal of Visual Communication and Image Representation, Vol. 55, 2018, pp. 627–639, doi: 10.1016/j.jvcir.2018.07.010.

[16] Martinez, J.—Hossain, R.—Romero, J.—Little, J. J.: A Simple Yet Effective Baseline for 3D Human Pose Estimation. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2659–2668, doi: 10.1109/iccv.2017.288.

[17] Mehta, D.—Rhodin, H.—Casas, D.—Fua, P.—Sotnychenko, O.—Xu, W.—Theobalt, C.: Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. 2017 International Conference on 3D Vision (3DV), Qingdao, China, IEEE, 2017, pp. 506–516, doi: 10.1109/3dv.2017.00064.

[18] Mehta, D.—Sotnychenko, O.—Mueller, F.—Xu, W.—Elgharib, M.—Fua, P.—Seidel, H.-P.—Rhodin, H.—Pons-Moll, G.—Theobalt, C.: XNect: Real-Time Multi-Person 3D Human Pose Estimation with a Single RGB Camera. CoRR, 2019, arXiv: 1907.00837v1.

[19] Mehta, D.—Sotnychenko, O.—Mueller, F.—Xu, W.—Sridhar, S.—Pons-Moll, G.—Theobalt, C.: Single-Shot Multi-Person 3D Pose Estimation from Monocular RGB. 2018 International Conference on 3D Vision (3DV), Verona, Italy, IEEE, 2018, pp. 120–130, doi: 10.1109/3dv.2018.00024.

[20] Mehta, D.—Sridhar, S.—Sotnychenko, O.—Rhodin, H.—Shafiei, M.—Seidel, H.-P.—Xu, W.—Casas, D.—Theobalt, C.: VNect: Real-Time 3D Human Pose Estimation with a Single RGB Camera. ACM Transactions on Graphics, Vol. 36, 2017, No. 4, Art. No. 44, 14 pp., doi: 10.1145/3072959.3073596.

[21] Moon, G.—Chang, J. Y.—Lee, K. M.: V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 5079–5088, doi: 10.1109/CVPR.2018.00533.

[22] Moreno-Noguer, F.: 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1561–1570, doi: 10.1109/CVPR.2017.170.

[23] Newell, A.—Yang, K.—Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.): Computer Vision – ECCV 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9912, 2016, pp. 483–499, doi: 10.1007/978-3-319-46484-8_29.

[24] Pavllo, D.—Feichtenhofer, C.—Grangier, D.—Auli, M.: 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 7745–7754, doi: 10.1109/CVPR.2019.00794.

[25] Peng, X.—Tang, Z.—Yang, F.—Feris, R. S.—Metaxas, D. N.: Jointly Optimize Data Augmentation and Network Training: Adversarial Data Augmentation in Human Pose Estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2226–2234, doi: 10.1109/cvpr.2018.00237.

[26] Qi, C. R.—Su, H.—Mo, K.—Guibas, L. J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.

[27] Qi, C. R.—Yi, L.—Su, H.—Guibas, L. J.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. CoRR, 2017, arXiv: 1706.02413.

[28] Rogez, G.—Weinzaepfel, P.—Schmid, C.: LCR-Net++: Multi-Person 2D and 3D Pose Detection in Natural Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, 2020, No. 5, pp. 1146–1161, doi: 10.1109/TPAMI.2019.2892985.

[29] Ronneberger, O.—Fischer, P.—Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (Eds.): Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, Cham, Lecture Notes in Computer Science, Vol. 9351, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[30] Schnürer, T.—Fuchs, S.—Eisenbach, M.—Gross, H.-M.: Real-Time 3D Pose Estimation from Single Depth Images. Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – Volume 5: VISAPP, Prague, Czech Republic, 2019, pp. 716–724, doi: 10.5220/0007394707160724.

[31] Shafaei, A.—Little, J. J.: Real-Time Human Motion Capture with Multiple Depth Cameras. 2016 13th Conference on Computer and Robot Vision (CRV), Victoria, BC, Canada, 2016, pp. 24–31, doi: 10.1109/CRV.2016.25.

[32] Shotton, J.—Fitzgibbon, A.—Cook, M.—Sharp, T.—Finocchio, M.—Moore, R.—Kipman, A.—Blake, A.: Real-Time Human Pose Recognition in Parts from Single Depth Images. 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, CO, USA, 2011, pp. 1297–1304, doi: 10.1109/CVPR.2011.5995316.

[33] Sun, K.—Xiao, B.—Liu, D.—Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5686–5696, doi: 10.1109/cvpr.2019.00584.

[34] Sun, X.—Shang, J.—Liang, S.—Wei, Y.: Compositional Human Pose Regression. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2621–2630, doi: 10.1109/iccv.2017.284.

[35] Ofli, F.—Chaudhry, R.—Kurillo, G.—Vidal, R.—Bajcsy, R.: Berkeley MHAD: A Comprehensive Multimodal Human Action Database. 2013 IEEE Workshop on Applications of Computer Vision (WACV), 2013, pp. 53–60, Clearwater Beach, FL, USA, doi: 10.1109/WACV.2013.6474999.

[36] Wang, P.—Li, W.—Ogunbona, P.—Wan, J.—Escalera, S.: RGB-D-Based Human Motion Recognition with Deep Learning: A Survey. CoRR, 2017, arXiv: 1711.08362.

[37] Wu, W.—Qi, Z.—Li, F.: PointConv: Deep Convolutional Networks on 3D Point Clouds. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 9613–9622, doi: 10.1109/CVPR.2019.00985.

[38] Xiao, B.—Wu, H.—Wei, Y.: Simple Baselines for Human Pose Estimation and Tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11210, 2018, pp. 472–487, doi: 10.1007/978-3-030-01231-1_29.

[39] Xiong, F.—Zhang, B.—Xiao, Y.—Cao, Z.—Yu, T.—Zhou, J. T.—Yuan, J.: A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation from a Single Depth Image. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 2019, pp. 793–802, doi: 10.1109/ICCV.2019.00088.

[40] Xiu, Y.—Li, J.—Wang, H.—Fang, Y.—Lu, C.: Pose Flow: Efficient Online Pose Tracking. 29th British Machine Vision Conference (BMVC), Newcastle, UK, CoRR, 2018, arXiv: 1802.00977.

[41] Ye, M.—Wang, X.—Yang, R.—Ren, L.—Pollefeys, M.: Accurate 3D Pose Estimation from a Single Depth Image. 2011 International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 731–738, doi: 10.1109/ICCV.2011.6126310.

[42] Yin, B.—Zhang, D.—Li, S.—Hao, A.—Qin, H.: Context-Aware Network for 3D Human Pose Estimation from Monocular RGB Image. 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8852263.

Dana Škorvánková is Ph.D. student at the Faculty of Mathemathics, Physics and Informatics of the Comenius University in Bratislava. Her dissertation thesis is oriented on neural networks applied on skeleton tracking and anthropometric measurements estimation. She finished her bachelor and master degree at the stated faculty, with focus on computer graphics and computer vision, along with machine learning and neural networks. The bachelor thesis "Capturing of Movement During Music Performance" was selected as the best thesis of the Department of Applied Informatics, and the master thesis "Deep Learning-Based Human Pose Estimation from 3D Data" has been awarded by the head of the university.

Martin Madaras received his Ph.D. degree in computer science in 2014 from the Comenius University in Bratislava with the focus on mesh processing and skeleton applications. In 2017 he co-founded a company Skeletex Research, where the main focus has been given on processing data from 3D scanners and cameras. He has been Postdoctoral Researcher at the Comenius University for 6 years and research lead at Skeletex Research for 4 years. Currently, he is also working as Assistant Professor at the Slovak University of Technology. The main research topics in his scope are geometry and point cloud processing, skeleton tracking and 3D model reconstruction.

Computing and Informatics, Vol. 40, 2021, 408–427, doi: 10.31577/cai_2021_2_408

OPTIMAL ALLOCATION OF CHARGING STATIONS FOR ELECTRIC VEHICLES USING PROBABILISTIC ROUTE SELECTION

Henrik Fredriksson, Mattias Dahl

Department of Mathematics and Natural Sciences
Blekinge Institute of Technology
371 79 Karlskrona, Sweden
e-mail: {henrik.fredriksson, mattias.dahl}@bth.se

Johan Holmgren

Department of Computer Science and Media Technology
Malmö University
205 06 Malmö, Sweden
e-mail: johan.holmgren@mau.se

Abstract. Electric vehicles (EVs) are environmentally friendly and are considered to be a promising approach toward a green transportation infrastructure with lower greenhouse gas emissions. However, the limited driving range of EVs demands a strategic allocation of charging facilities, hence providing recharging opportunities that help reduce EV owners' anxiety about their vehicles' range. In this paper, we study a set covering method where self-avoiding walks are utilized to find the most significant locations for charging stations. In the corresponding optimization problem, we derive a lower bound of the number of charging stations in a transportation network to obtain full coverage of the most probable routes. The proposed method is applied to a transportation network of the southern part of Sweden.

Keywords: Charging stations, electric vehicle, transportation network, optimal placement, self-avoiding random walk

Mathematics Subject Classification 2010: 90-C06, 90-B20, 90-B80

1 INTRODUCTION

The sale rate of electric vehicles (EVs) has been growing rapidly over the past ten years, and there is a need to adapt the current transportation infrastructure to meet future recharging demands. Increased use of EVs – including both plug-in hybrid electric vehicles (PHEVs) and battery electric vehicles (BEVs) – has been recognized as a promising, sustainable approach to lowering traffic emissions, including greenhouse gases [1]. However, their limited driving range and the scarcity of public accessible charging stations prevent EVs from gaining widespread market acceptance [2]. Psychological stress caused by the fear that the vehicle will run out of energy and be stranded is referred to as range anxiety [3]. As a consequence of range anxiety, EV owners may use their electric-powered vehicles for short trips exclusively, with the result that they require an additional vehicle for longer trips. In a small survey with 58 participants from 2011, Skippon and Garwood [4] report that consumers might consider an EV as their main car or second car if it had a range of 150 miles (241 km) or 100 miles (161 km), respectively. In a more recent survey, from 2016, Skippon et al. [5] report that consumers' desired driving ranges for EVs have substantially increased. The results show that people who have driven a modern EV would consider having an EV as the main car if the driving range is 200 miles (322 km) and as the second car if the driving range is 150 miles. Also, the study by Jensen et al. [6] confirms that the limited driving range is a concern for the acceptance of EVs. The study reports that the EVs' driving ranges do not match the expectations of consumers, after they use EVs for a trial period. Thus, as decision-makers and infrastructure planners consider to gain market acceptance for EVs, it will be important to determine how to best allocate charging stations to compensate for the current limited capacity of EV batteries. A strategic allocation of accessible charging facilities or battery swap stations may reduce the range anxiety of EV owners. Hence, an important step in addressing the problem of allocating of the charging station is to identify the routes in the transportation network that are most likely to be used to serve as many EV drivers as possible. Additionally, a desirable outcome of a charging station allocation is that every vehicle that drive around in the network, regardless of its position, should be able to reach a charging station before it runs out of energy. A strategic deployment of charging infrastructure may also minimize the initial cost of the installment of new charging facilities and relieve the load on the electrical power system [7].

In the current paper, we propose a novel solution procedure for the set covering problem for the allocation of EV charging stations. The basic formulation of the set covering problem is to minimize the number of charging stations such that each route is covered by at least one charging station [8, 9]. The set covering formulation, among other methods of deploying a public charging infrastructure, has received substantial attention from the research community. In what follows, we give a brief review of some of the methods that have been proposed in the literature. In Wang and Lin [10], a flow-based set covering method is proposed to minimize the cost of

installation of charging facilities such as fast-refueling stations and battery exchange stations. The optimization method for the flow-based set covering method is based on the following vehicle-routing logic: the greater the distance a vehicle is driving, the more likely it is that the vehicle will require refueling. An extension to the problem is to consider a dual objective model to minimize the installation cost and maximize the population coverage, by combining the flow-based set covering model and the traditional set covering model [11]. Wang [11] considers the allocation of charging stations for electric scooters, where the aim is to minimize the total installation cost. The model by Wang [11] is extended by Wang and Lin [12] to consider facility budget constraints, multiple types of recharging stations, and vehicle routing behavior. The types of charging stations include slow- and fast-recharging stations as well as battery swap stations. The case study presented in [12] shows that the results achieved with mixed-type charging facilities are better than those achieved with single-type facilities. The refueling logic requires numerous binary variables, which makes the problem hard to solve. When a flexible expanded network method is used, as proposed by MirHassani and Ebrazi [13], the solution time of the flow-based set covering method is significantly reduced. Wen et al. [14] consider both the problem of how to maximize the flow coverage with a fixed number of available charging stations and the problem of how to minimize the number of charging stations to obtain full coverage. In both models, the limited driving range of EVs is addressed by partitioning routes into sub-routes according to recharging logic. In the above-mentioned model, the general assumption is that vehicles would only consider routes that are the shortest or have the least traveling time between origin and destination. Li and Huang [15], Huang and Zhou [16], and Hosseini et al. [17] use the concept of deviation path. In their models, the shortest-path assumption is relaxed by the assumption that EV users are willing to slightly deviate from their preferred trips to ensure that they can refuel en route to their destinations. Frade et al. [18] present a mixed-integer optimization problem to maximize the coverage of both daytime and nighttime demand within an acceptable level of service for a neighbourhood in Lisbon. Funke et al. [19] propose a framework based on the hitting set problem, which aims to guarantee energy supply for all shortest paths in the network.

Another method commonly used to locate charging facilities and battery swap stations is based in flow capturing models [20, 21]. The Flow Refueling Location Model (FRLM) aims to locate a fixed number of charging stations in the nodes in such a way that the total number of vehicles can be refueled within their limited driving range [22]. Further, if it is possible to locate charging stations both along links and in nodes, the coverage of the network may be substantially improved compared to when nodes are the only candidate sites [23]. A more realistic extension to the FRLM is to consider the charging capacity of the allocated facilities [24]. In the capacitated FRLM, the location variables are not binary but non-negative integers, meaning that multiple charging facilities could be located to serve as many vehicles as possible. Based on a flow-capturing model, Lim and Kuby [25] developed three heuristic algorithms for locating alternative-fuel stations. Solving the

FRLM is usually a two-stage process; the first stage generates combinations of candidate locations, and the second stage uses these combinations to locate charging stations to maximize the number of refueled vehicles. Capar and Kuby [26] developed a method to solve the FRLM in one stage. To plan and design an infrastructure complete with battery swapping stations and battery management, Mak et al. [27] study the robust location problem of battery swapping stations under demand uncertainty.

If available, Global Position System (GPS) data can be utilized to support the allocation of charging stations. GPS data can, for example, be collected from taxis to obtain vehicle travel patterns, which are used to allocate charging stations [28, 29, 30, 31]. GPS travel survey data can also be used to simulate vehicles' driving and charging behavior to optimally locate charging stations such that the number of missed trips is minimized [32]. Additional data, such as initial battery level of the vehicle and charging mode (normal or fast), together with GPS data, may provide useful insights on the charging behavior of EVs [33]. However, the applicability of methodologies based on GPS data is limited, due to the lack of data available for research purposes [34].

To capture the interaction between the availability of charging stations and the route choices of drivers, several studies use traffic assignment models to identify the locations of charging stations. Traffic assignment models are multi-commodity flow problems under some given optimal or equilibrium routing principle. He et al. [35] propose a bi-level traffic assignment model. The upper level allocates a fixed number of charging stations such that the number of vehicles that use a charging station is maximized, while user equilibrium of route choice together with the EV's limited range is considered in the lower level. He et al. [36] propose a framework to capture the interactions between the locations of charging stations, electricity prices, route choices, and recharging time, which is solved by an active-set algorithm. Additionally, He et al. [37] present three different network equilibrium models, where the different flow dependencies and energy consumption are integrated. A similar model considers drivers' spontaneous adjustments and the interactions of travel and recharging decisions [38].

Despite the extensive work in academia, the process of allocating charging stations is still a challenging problem in real-world scenarios due to legal, physical, and financial constraints [34]. Typically, constraints are much stricter and more complex than is assumed in studies, and there is a need for data and knowledge to fully understand the impacts of charging infrastructure concerning location, installation, operation, and future maintenance.

The current paper extends the paper by Fredriksson et al. [39] where an iterative solution procedure to optimally allocate charging stations for EVs is proposed. In particular, the contribution of the current paper is an explicit termination criteria for the iterative method, along with an improved route identification method to capture driving behavior. The studied problem is formulated as a set covering problem where the constraints in the associated integer problem are obtained by self-avoiding random walks. In the random walks, a probabilistic rule determined by link flows

is applied in each node to select the next node in the walk. By iteratively adding constraints and solving sub-problems, we obtain a lower bound approximation of the minimal number of charging stations required to cover a transportation network without route enumeration.

The paper is organized as follows: In Section 2 we describe the model and the problem formulation. The proposed solution algorithm is explained in Section 3. Numerical results are presented in Section 4, and Section 5 concludes this paper and discusses some future research directions.

2 PROBLEM FORMULATION

A transportation network is described by a set of nodes $N = \{1, 2, \ldots, n\}$ and a set of routes R. For each route $r \in R$, let $\delta_{ir} = 1$ if a vehicle is visiting node $i \in N$ while traveling on route r, and $\delta_{ir} = 0$ otherwise. Let x_i be a binary variable where $x_i = 1$ if a charging station is allocated in the node $i \in N$, and otherwise $x_i = 0$. An allocation of charging stations is mathematically defined by a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$. The driving range is the maximal distance a vehicle can drive without recharging and is denoted by $d_{\max}$. Let d_w^E be the Euclidean distance from the start node to the end node in a self-avoiding walk w. A self-avoiding walk w is a route if $d_w^E > d_{\min}^E$ where $d_{\min}^E$ is a positive real number.

The problem studied in this paper is formulated as a set covering problem and is based on covered known routes. A route $r \in R$ is *covered* if at least one charging station is placed in one of its nodes. The route cover criteria for a route r correspond to the inequality

$$\sum_{i \in N} \delta_{ir} x_i \geq 1. \tag{1}$$

The set covering problem for the allocation of charging stations can be described as follows: Given a transportation network, find the set of all routes R and the minimal number of charging stations and their locations such that each route $r \in R$ is covered. The optimization part of the problem corresponds to the optimization program

$$(P) \quad z = \min_{\boldsymbol{x} \in \{0,1\}^n} \left\{ \sum_{i \in N} x_i : \sum_{i \in N} \delta_{ir} x_i \geq 1, \forall r \in R \right\}. \tag{2}$$

Theoretically, even for a rather small transportation network, a complete route enumeration can lead to an unmanageable number of constraints in our set covering problem. Furthermore, if a complete route enumeration is available, we strongly believe that many of these routes may not be used at all by any vehicle. Instead of trying to find the set of all routes, we claim that it is sufficient to identify the most probable sub-routes (sub-routes that are most likely to be used), which, however, is far from a trivial task in large transportation networks. For simplicity, we refer to these sub-routes as routes. We assume that all routes $r \in R$ have a maximum

length $d_{\max}$. Hence, vehicles driving distances shorter or longer than this maximum distance cover fractions of routes or multiple routes, respectively. We describe our route identification method in Section 3.2, where we also discuss how we deal with the limited driving range of EVs.

3 OPTIMAL PLACEMENT OF CHARGING STATIONS

In this section we describe the solution procedure for our set covering problem. We also discuss how we obtain the routes that will be iteratively added to our model.

3.1 Solution Procedure

A solution to the charging station allocation problem is obtained by iteratively solving sub-problems ($P^{(k)}$) of our problem (P), where the sub-problems are based on subsets of routes $R_k \subset R$. The sub-problems yield a sequence of optimal solutions with a monotonically increasing minimum number of charging stations, converging to the optimum value of problem (P). Our aim is to search and select routes in the network to achieve a solution with coverage of the most probable routes. To this end, we exploit probabilistic self-avoiding random walks to identify routes in the transportation network. The identified routes are iteratively generated and added as constraints to the integer problem under consideration. The approximation technique proposed to solve our problem is based on integer programming, and the outline of the method can be described with the following basic steps:

0. Initialization: Set iteration counter $k = 0$, and $R_k = \emptyset$. Fix $x_i = 1$ if a charging station is already allocated in node $i \in N$.
1. Given a reference set $R_k \subset R$, solve the sub-problem

$$\left(P^{(k)}\right) \quad z^{(k)} = \min_{\boldsymbol{x}\in\{0,1\}^n} \left\{ \sum_{i\in N} x_i : \sum_{i\in N} \delta_{ir} x_i \geq 1, \forall r \in R_k \right\} \tag{3}$$

   yielding the solution vector $\boldsymbol{x}^{(k)}$.
2. Define the entering index by a route $r_e \in R$ satisfying $\sum_{i\in N} \delta_{ir_e} x_i^{(k)} = 0$. Stop with approximate optimal reference set $R_k \subset R$, if a new r_e cannot be found according to some termination criteria.
3. Define the new reference set by $R_{k+1} = R_k \cup \{r_e\}$, set $k = k + 1$, and go to step 1.

A flowchart of the solution procedure is depicted in Figure 1. At iteration k, the solution vector $\boldsymbol{x}^{(k)}$ is infeasible for the new reference set R_{k+1} since the new entering index r_e is uncovered. We can easily provide a feasible solution to sub-problem $P^{(k+1)}$ by setting $x_i = 1$ where $i = \min\{k \,:\, \delta_{kr_e} = 1\}$.

Without any prior knowledge about the routes in the network, it is computationally difficult to estimate the number of routes in a transportation network, and thereby to find a suitable, explicit termination criteria. For instance, the iterative method could be terminated when a sufficient number of constraints has been collected, when some search time for a new uncovered route is exceeded, or when a new uncovered route cannot be found within a certain number of iterations. In our numerical examples in Section 4 we apply the latter termination criteria. If a new uncovered route cannot be found after 10 000 random walk attempts, we assume that the most commonly used routes in the network have been taken into consideration.

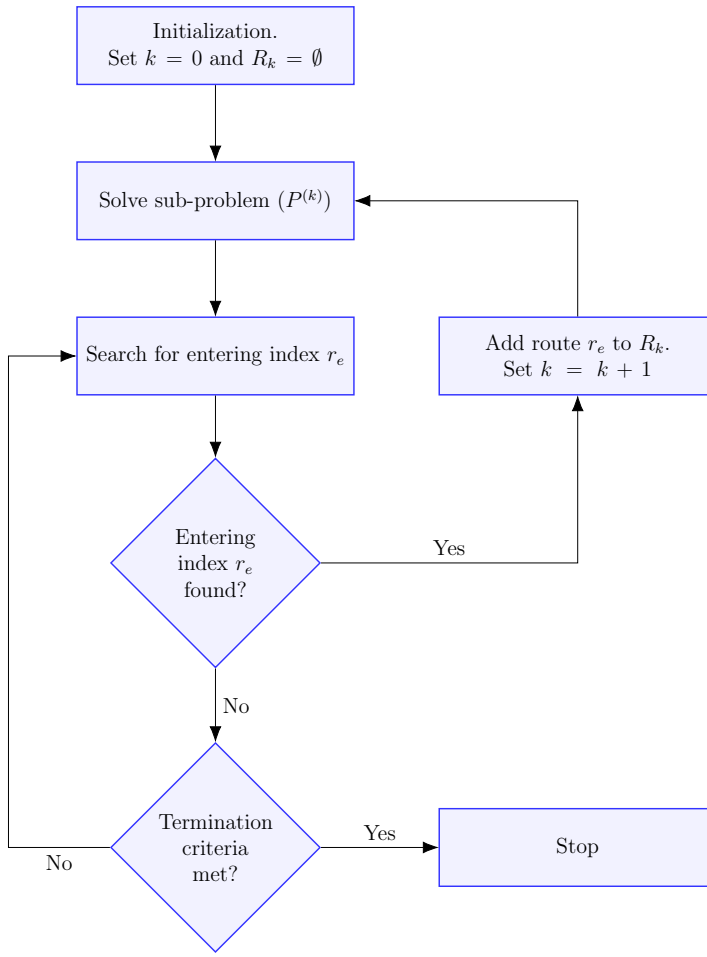


Figure 1. Flowchart for the solution procedure for our set covering problem

3.2 Self-Avoiding Random Walks

As mentioned above, the constraint index r_e, which is chosen to enter the basis R_k is defined by a route that does not satisfy the current solution, i.e., a constraint violation. By using the most probable routes in the network, we claim that when no entering index r_e can be found according to some termination criteria, the most commonly used (probable) routes in the network will be covered by at least one charging station.

In practice, finding the complete route set R is computationally difficult, and several models use a predefined set of routes. We use self-avoiding random walks as a search method to simulate the driving behavior of EV drivers to find new uncovered routes. A self-avoiding random walk is defined as a random walk with the restriction that it cannot revisit a node. Self-avoiding random walks can be more effective when exploring a transportation network and may model driving behavior more realistically than unrestricted random walks, since they cannot return to already visited nodes [40]. A probabilistic rule determined by collected link flows is applied at each node to select the next node among the neighbors of the current node, yielding a probabilistic self-avoiding random walk. Let $i \in N$ be the current node in the walk; then the probability of choosing the next neighboring node j is

$$p = \frac{\phi(i,j)}{\sum_{x \in N_i} \phi(i,x)}. \tag{4}$$

Here, $N_i \subset N$ is the set of adjacent and unvisited nodes of the current node i, in the current walk under construction, and $\phi(i,j)$ is the flow on link (i,j).

Link flows can be collected using temporary or permanent sensors, including pneumatic tubes, that are moved around in the network according to a periodic scheme. Obviously, the link flows do not provide as much information as other flows – e.g., origin–destination flows and route flows – but they are easier to obtain. Obviously, the link flows vary over time, but on an aggregate level, static link flows provide a sufficient approximation to generate the most probable routes in a transportation network. The collected link flows are utilized to model driving behavior in the network. We argue that this approach provides a reference for planning and charging infrastructure for the current transportation situation. Due to the current vehicle market, the collected link flow data are mainly constituted by vehicles with internal combustion engines fueled by gasoline or diesel. Although changes may occur depending on vehicle type, in this study we assume that travel patterns and driving behavior are the same for EVs and for gasoline or diesel-fueled vehicles.

Depending on the scenario under investigation, in our route generation process, we restrict our selection to only consider routes with maximum length $d_{\max}$, which is assumed to be the maximal distance a vehicle can to travel without passing a node equipped with a charging station. Thus, a partition of the full route into road segments, of appropriate lengths to handle the limited driving range of EVs is

identified directly. Since the program optimally allocates charging stations such that each route is covered, a solution will ensure coverage of the most probable routes of length $d_{\max}$. In other words, an EV cannot travel longer than $d_{\max}$ without passing a charging facility if it travels on the most probable route. In this way, we can adjust the parameter $d_{\max}$ to respect the driving range of an EV.

Since we are demonstrating our proposed method on a government-controlled transportation network with roads between cities and metropolitan areas, we are especially interested in covering routes that are included in trips between communities. To avoid peculiar zigzag behaviors or U-turns, we only consider routes where the Euclidean distance is at least $d^E_{\min}$ between the start- and end-nodes of the route.

Each of the random walks begins in a randomly chosen node. If a walk reaches a node already equipped with a charging station, i.e., the route is already covered by the current solution, a new random walk is restarted in a randomly chosen node until a new unique entering index r_e, violating the constraints in the current sub-problem, is found. The search method for the entering index r_e is done by the following steps.

1. Start in a random node.
2. With probabilities according to Equation (4), randomly choose a neighbouring unvisited node (in the current walk).
   (a) If the chosen node has a charging station allocated on it, discard the walk and go to Step 1.
   (b) If the total length of the walk under construction exceeds $d_{\max}$, and the Euclidean distance between the first and current node exceeds $d^E_{\min}$, add the walk to the problem as a constraint, solve the sub-problem, and go to Step 1.
   (c) Otherwise, repeat Step 2.

Depending on the network under study, there is a margin of error concerning the limited driving range, since the generated walks are presumed to be slightly longer than $d_{\max}$. This error, however, can be alleviated by lowering $d_{\max}$.

As the number of constraints increases, it will be more and more difficult to find new uncovered routes, since vehicles are more likely to traverse a node already equipped with a charging station. Hence, our iterative process converges towards an optimal solution where the most probable routes are covered. Because the iterative solution procedure converges to an optimal solution with full coverage of the most probable routes in a transportation network, we argue that a solution using our proposed method is worthy of consideration since it does not contain routes that are less likely to be used.

3.3 Outline of the Convergence

The main idea of the optimization procedure is to iteratively solve sub-problems of our set covering problem and to continuously further extend the current sub-

problem into a slightly larger problem. Since the problem is formulated as a set covering problem, a solution to the problem is not necessarily unique, but the iterative procedure is continuously striving for a solution with a minimal number of charging stations. The procedure ensures coverage and localizes common junctions of interest within the given transportation network. The main purpose of this procedure is to ensure that each charging station is placed in an environment that is of mutual interest to several of the found routes.

The behavior of convergence can be briefly outlined, and the emphasis is on an intuitive understanding for the optimization procedure. Assume that we are given the solution $\boldsymbol{x}^k$ to any sub-problem $P^{(k)}$. If we now add one single constraint where $\boldsymbol{x}^k$ violates the constraint, we can easily provide a feasible solution to iteration $k+1$ as described in Section 3.1. We now establish the following important inequality relations

$$z^{(1)} \leq \ldots \leq z^{(k)} \leq z^{(k+1)} \leq \ldots \leq z. \tag{5}$$

The first inequality is due to the fact that the optimal solution for problem $(P^{(k+1)})$ is a feasible solution for problem $(P^{(k)})$ in integer problem formulation (3) and provides an upper bound for $(P^{(k)})$. The sequence $(z^{(k)})$ is thus monotonically increasing and upper bounded. Convergence to the optimum value of $z^{(k)}$ is motivated by the last equality where the existence of a finite optimal reference set R is established. We note, in this context, that the convergence of the algorithm is based on the assumption that numerical difficulties (such as randomness, size of network, etc.) are avoided due to the presence of a reliable software for the solution. In practice, the optimal value of z will never be reached for large-scale networks. However, the system yields an appropriate approximation, and therefore it is subject to continuous improvement with the aim of reaching the optimum.

In this presentation, we emphasize the simplicity and efficiency of this approach to finding a minimal number of charging stations and allocating them so that every route is covered. Since the procedure is based on a controlled selection of constraints, there are opportunities to add and fulfill requirements that make the procedure of selecting routes to model the transportation network even more realistic.

4 COMPUTATIONAL STUDY

To demonstrate our proposed method's effectiveness, we consider the network of the southernmost part of Sweden. This network is one of six traffic regions maintained by the Swedish Transport Administration. The network consists of 14 500 nodes and 34 500 links distributed over an area of approximately 44 500 square kilometers, spread across five counties, and it is shown in Figure 2. The mean length of the links is 1.29 km with a standard deviation of 1.98 km. The distribution of the link lengths in the network under study is shown in Figure 3. Since the majority of the links are quite short, routes in our scenarios ranges 60 km to 100 km in length, and

will consist of several nodes. Although, a random walk may have choose a node with low probability, according to the probabilistic rule (4), it will most likely reassemble with more probable links in a later stage. The link flows used in our study are based on real-world data collected from traffic counts in July 2018. The data was obtained from *The Swedish National Road Database* (NVDB) provided by the Swedish Transport Administration. In the optimization procedure, neighboring regions have not been taken into account.

Figure 2. The national transportation main road map for the southern Sweden (area of interest is beneath the black border). The transportation network (all roads) consists of 14 500 nodes 34 500 links distributed over an area of 44 500 square kilometers.

We implemented our problem-solving method to find allocations of charging stations in the numerical computing environment MATLAB® [41] and Gurobi Optimizer [42].

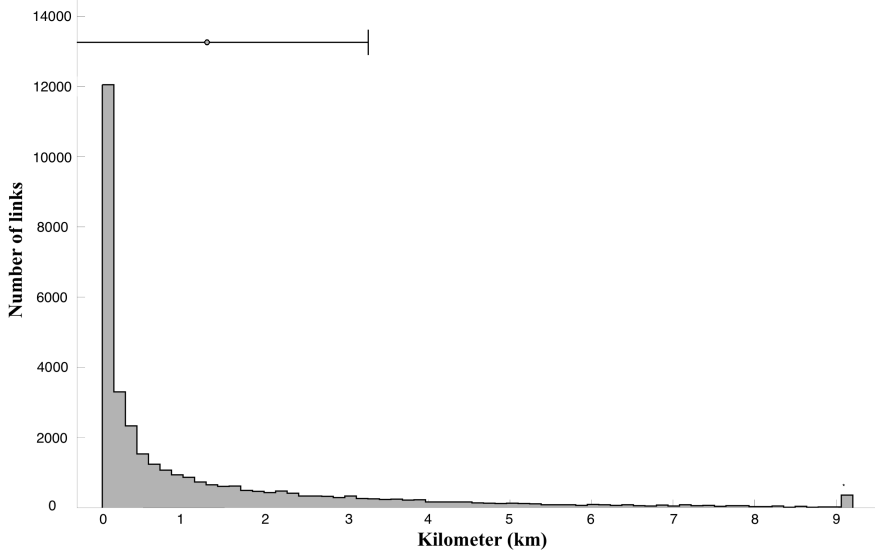


Figure 3. Distribution of link lengths in the network of southern Sweden, where the mean length of the links is 1.29 km, with a standard deviation of 1.98 km. The 339 points counted in the rightmost bin are links with length greater than 9.06 km.

4.1 Scenario Descriptions

In our numerical evaluations we consider 3 different scenarios, each with varying maximal driving range $d_{\max}$, and minimal distance travelled $d^E_{\min}$. We study the number of charging stations required to obtain full coverage of all found routes. In all scenarios we used a maximum number of search iterations as termination criteria. If no new uncovered route r_e could be found within 10 000 iterations, the program was terminated. The number of constraints in the final iteration and the number of allocated charging stations for each scenario are shown in Table 1. As mentioned in Section 3.3, a solution to our set covering problem might not be unique, but it provides a lower bound of the number of charging stations required to cover each of a number of considered routes. Examples of solutions for each of the considered scenarios are illustrated in Figures 4, 5, and 6.

We emphasize that, since we use a network in the southernmost part of Sweden, the studied scenarios have a macroscopic approach. The allocation of charging stations has a focus outside urban and metropolitan areas, in this regard, the process of its optimization has a focus on the government-controlled transportation network.

As the numerical results show, the number of required charging stations decreases as $d_{\max}$ increases. For illustrative purposes, Figure 7 shows how the optimal placements of charging stations have evolved during the iterative process for Sce-

	$d_{\max}$ (km)	$d^E_{\min}$ (km)	Constraints	Allocated Charging Stations
Scenario 1	60	30	1 177	203
Scenario 2	80	40	1 075	174
Scenario 3	100	50	569	104

Table 1. Number of constraints and number of allocated charging stations of the iterative set covering model for the three scenarios generated by different parameter settings of $d_{\max}$ and $d^E_{\min}$

nario 3 with $d_{\max} = 100\,\text{km}$ and $d^E_{\min} = 50\,\text{km}$. The figure shows all nodes that have been included in an optimal solution $\boldsymbol{x}^{(k)}$ for some iteration index k. As seen in the figure, numerous nodes have been considered as candidate sites for optimal placement to achieve coverage. The final allocation of charging stations for this scenario is depicted in Figure 6.

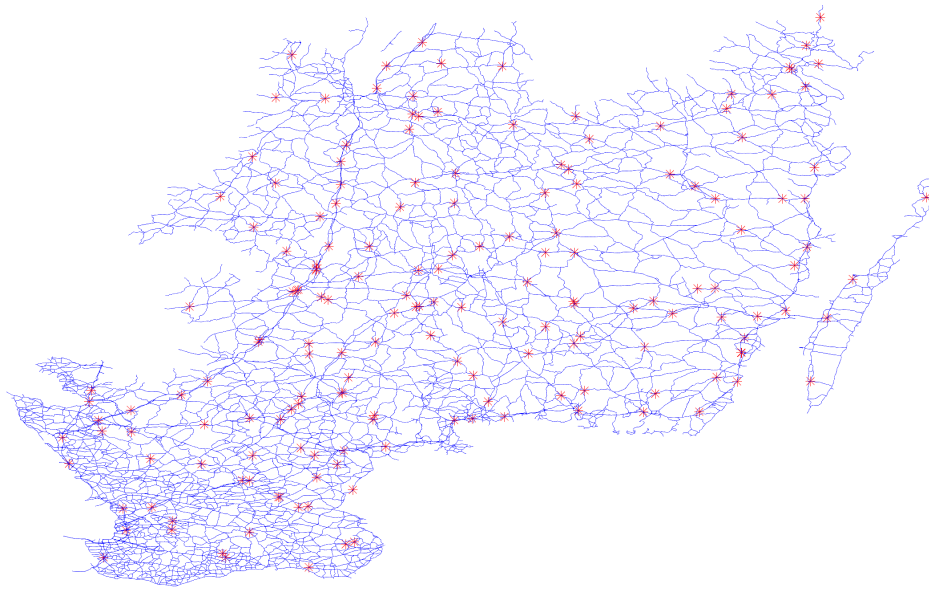


Figure 4. The allocation of charging stations for $d_{\max} = 60\,\text{km}$ and $d^E_{\min} = 30\,\text{km}$

5 CONCLUSIONS AND FUTURE WORK

EVs for both public and private transport seem to be a promising solution to reduce greenhouse gas emissions. However, the scarcity of available charging stations and the limited driving range of EVs, are two major barriers for the widespread adoption of EVs. In the current paper, we propose a node-based formulation of a set covering method to optimally allocate charging stations for EVs. The constraints in the

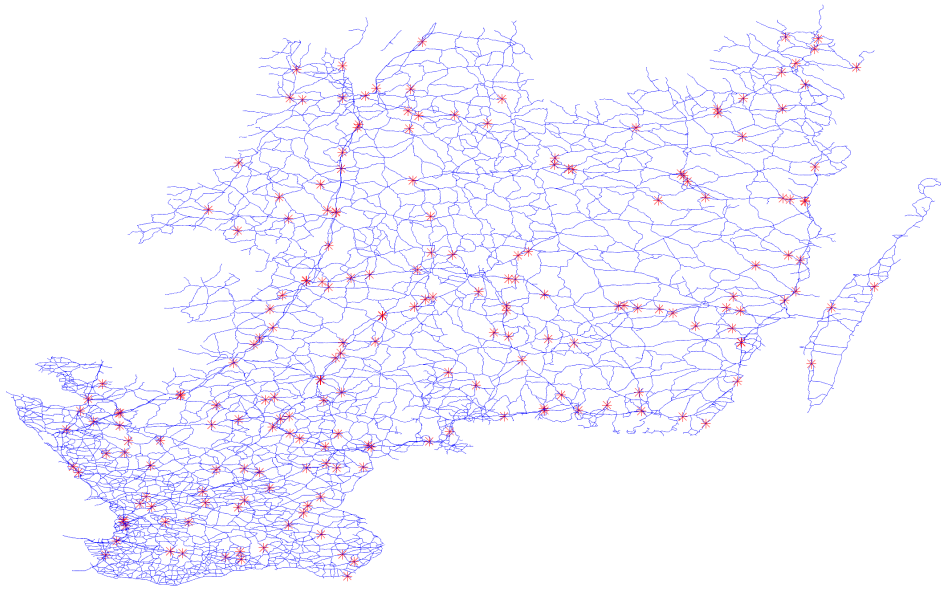


Figure 5. The allocation of charging stations for $d_{\max} = 80\,\text{km}$ and $d^E_{\min} = 40\,\text{km}$

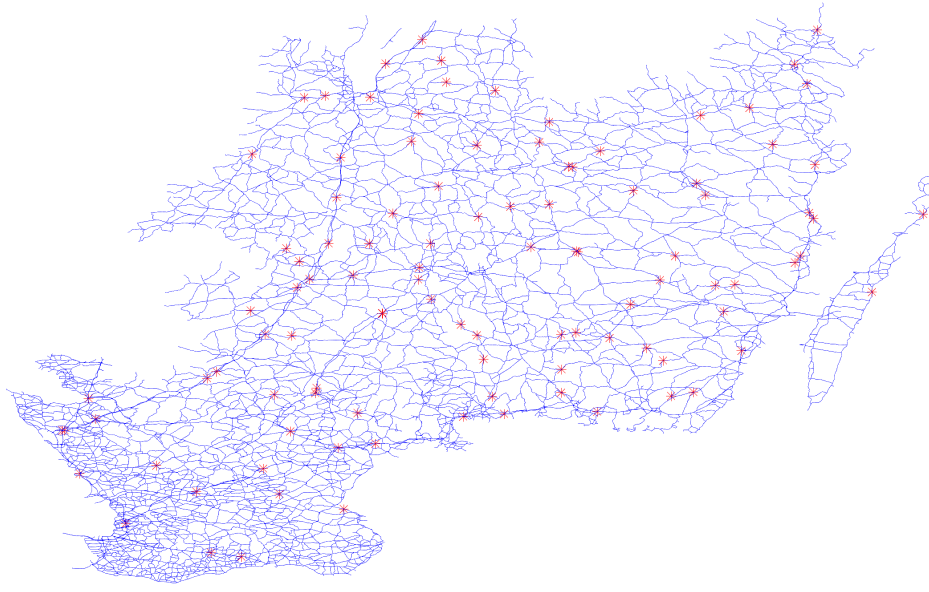


Figure 6. The allocation of charging stations for $d_{\max} = 100\,\text{km}$ and $d^E_{\min} = 50\,\text{km}$

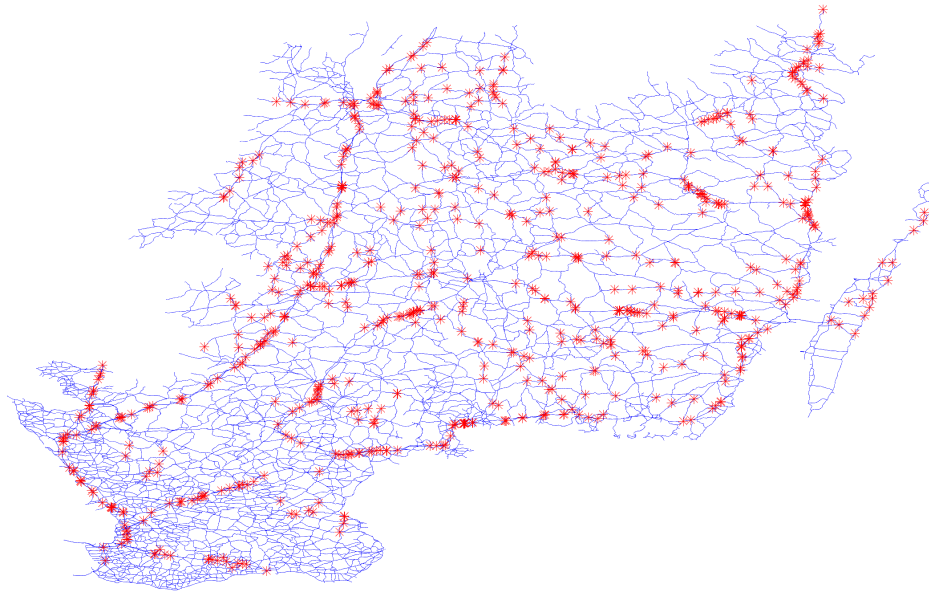


Figure 7. All candidate nodes for optimal allocation of charging stations for $d_{\max} = 100\,\text{km}$ and $d^E_{\min} = 50\,\text{km}$

corresponding optimization problem are based on self-avoiding random walks along the links in the network and the problem is solved by a pruned integer program that can take the existing infrastructure into consideration. The iterative optimization procedure and the probabilistic route selection provide an approximation of the optimal allocation to obtain full coverage.

The computational results of the proposed iterative method, in the case study of the Southern Sweden transportation network, shows that the method is able to allocate charging stations optimally without numerical difficulties. The results indicate how charging facilities can be located strategically to cover the most probable routes for an EV fleet in the studied area, for the several different scenarios. In particular, the method provides a lower bound to obtain full coverage and localizes nodes that are of common interest for the routes in the transportation network. From the results of our computational study, in which we compared maximal driving distances of 60, 80, and 100 kilometers, respectively, we observe that the ability to extend the driving distance of EVs significantly reduces the need for EV charging stations. In particular, the number of allocated charging stations in our considered network dropped from 203 to 88 when d_{max} was increased from 60 to 100 kilometers. This means that the number of charging stations dropped by 56.7 % when we increased d_{max} by 66.7 %. Hence, in addition to helping drivers to overcome the range anxiety, increasing the driving range of EVs also significantly reduces the need to build charging infrastructure. Further research based on

the current work can be conducted in several directions. One direction could consider the inclusion of budgetary constraints and the recharging capacity of charging stations. Another research could be focused on the search method for uncovered routes. A third direction of further research includes the optimal allocation considering that charging will most likely occur in the beginning or the end of a trip.

REFERENCES

[1] Ohnishi, H.: Greenhouse Gas Reduction Strategies in the Transport Sector: Preliminary Report. Technical report, OECD/ITF Joint Transport Research Centre Working Group on GHG Reduction Strategies in the Transport Sector, OECD/ITF, Paris, 2008.

[2] Romm, J.: The Car and Fuel of the Future. Energy Policy, Vol. 34, 2006, No. 17, pp. 2609–2614, doi: 10.1016/j.enpol.2005.06.025.

[3] Neubauer, J.—Wood, E.: The Impact of Range Anxiety and Home, Workplace, and Public Charging Infrastructure on Simulated Battery Electric Vehicle Lifetime Utility. Journal of Power Sources, Vol. 257, 2014, pp. 12–20, doi: 10.1016/j.jpowsour.2014.01.075.

[4] Skippon, S.—Garwood, M.: Responses to Battery Electric Vehicles: UK Consumer Attitudes and Attributions of Symbolic Meaning Following Direct Experience to Reduce Psychological Distance. Transportation Research Part D: Transport and Environment, Vol. 16, 2011, No. 7, pp. 525–531, doi: 10.1016/j.trd.2011.05.005.

[5] Skippon, S. M.—Kinnear, N.—Lloyd, L.—Stannard, J.: How Experience of Use Influences Mass-Market Drivers' Willingness to Consider a Battery Electric Vehicle: A Randomised Controlled Trial. Transportation Research Part A: Policy and Practice, Vol. 92, 2016, pp. 26–42, doi: 10.1016/j.tra.2016.06.034.

[6] Fjendbo Jensen, A.—Cherchi, E.—Lindhard Mabit, S.: On the Stability of Preferences and Attitudes Before and After Experiencing an Electric Vehicle. Transportation Research Part D: Transport and Environment, Vol. 25, 2013, pp. 24–32, doi: 10.1016/j.trd.2013.07.006.

[7] Hajimiragha, A.—Canizares, C. A.—Fowler, M. W.—Elkamel, A.: Optimal Transition to Plug-In Hybrid Electric Vehicles in Ontario, Canada, Considering the Electricity-Grid Limitations. IEEE Transactions on Industrial Electronics, Vol. 57, 2010, No. 2, pp. 690–701, doi: 10.1109/TIE.2009.2025711.

[8] Toregas, C.—Swain, R.—ReVelle, C.—Bergman, L.: The Location of Emergency Service Facilities. Operations Research, Vol. 19, 1971, No. 6, pp. 1363–1373, doi: 10.1287/opre.19.6.1363.

[9] Daskin, M. S.: Network and Discrete Location: Models, Algorithms, and Applications. John Wiley & Sons, Hoboken, New Jersey, 2013, doi: 10.1002/9781118537015.

[10] Wang, Y. W.—Lin, C. C.: Locating Road-Vehicle Refueling Stations. Transportation Research Part E: Logistics and Transportation Review, Vol. 45, 2009, No. 5, pp. 821–829, doi: 10.1016/j.tre.2009.03.002.

[11] Wang, Y. W.: Locating Flow-Recharging Stations at Tourist Destinations to Serve Recreational Travelers. International Journal of Sustainable Transportation, Vol. 5, 2011, No. 3, pp. 153–171, doi: 10.1080/15568311003717199.

[12] Wang, Y. W.—Lin, C. C.: Locating Multiple Types of Recharging Stations for Battery-Powered Electric Vehicle Transport. Transportation Research Part E: Logistics and Transportation Review, Vol. 58, 2013, pp. 76–87, doi: 10.1016/j.tre.2013.07.003.

[13] MirHassani, S. A.—Ebrazi, R.: A Flexible Reformulation of the Refueling Station Location Problem. Transportation Science, Vol. 47, 2013, No. 4, pp. 617–628, doi: 10.1287/trsc.1120.0430.

[14] Wen, M.—Laporte, G.—Madsen, O. B. G.—Nørrelund, A. V.—Olsen, A.: Locating Replenishment Stations for Electric Vehicles: Application to Danish Traffic Data. Journal of the Operational Research Society, Vol. 65, 2014, No. 10, pp. 1555–1561, doi: 10.1057/jors.2013.100.

[15] Li, S.—Huang, Y.: Heuristic Approaches for the Flow-Based Set Covering Problem with Deviation Paths. Transportation Research Part E: Logistics and Transportation Review, Vol. 72, 2014, pp. 144–158, doi: 10.1016/j.tre.2014.10.013.

[16] Huang, Y.—Zhou, Y.: An Optimization Framework for Workplace Charging Strategies. Transportation Research Part C: Emerging Technologies, Vol. 52, 2015, pp. 144–155, doi: 10.1016/j.trc.2015.01.022.

[17] Hosseini, M.—MirHassani, S. A.—Hooshmand, F.: Deviation-Flow Refueling Location Problem with Capacitated Facilities: Model and Algorithm. Transportation Research Part D: Transport and Environment, Vol. 54, 2017, pp. 269–281, doi: 10.1016/j.trd.2017.05.015.

[18] Frade, I.—Ribeiro, A.—Gonçalves, G.—Antunes, A. P.: Optimal Location of Charging Stations for Electric Vehicles in a Neighborhood in Lisbon, Portugal. Transportation Research Record, Vol. 2252, 2011, No. 1, pp. 91–98, doi: 10.3141/2252-12.

[19] Funke, S.—Nusser, A.—Storandt, S.: Placement of Loading Stations for Electric Vehicles: No Detours Necessary! Journal of Artificial Intelligence Research, Vol. 53, 2015, pp. 633–658, doi: 10.1613/jair.4688.

[20] Hodgson, M. J.: A Flow-Capturing Location-Allocation Model. Geographical Analysis, Vol. 22, 1990, No. 3, pp. 270–279, doi: 10.1111/j.1538-4632.1990.tb00210.x.

[21] Berman, O.—Larson, R. C.—Fouska, N.: Optimal Location of Discretionary Service Facilities. Transportation Science, Vol. 26, 1992, No. 3, pp. 201–211, doi: 10.1287/trsc.26.3.201.

[22] Kuby, M.—Lim, S.: The Flow-Refueling Location Problem for Alternative-Fuel Vehicles. Socio-Economic Planning Sciences, Vol. 39, 2005, No. 2, pp. 125–145, doi: 10.1016/j.seps.2004.03.001.

[23] Kuby, M.—Lim, S.: Location of Alternative-Fuel Stations Using the Flow-Refueling Location Model and Dispersion of Candidate Sites on Arcs. Networks and Spatial Economics, Vol. 7, 2007, No. 2, pp. 129–152, doi: 10.1007/s11067-006-9003-6.

[24] Upchurch, C.—Kuby, M.—Lim, S.: A Model for Location of Capacitated Alternative-Fuel Stations. Geographical Analysis, Vol. 41, 2009, No. 1, pp. 85–106, doi: 10.1111/j.1538-4632.2009.00744.x.

[25] Lim, S.—Kuby, M.: Heuristic Algorithms for Siting Alternative-Fuel Stations Using the Flow-Refueling Location Model. European Journal of Operational Research, Vol. 204, 2010, No 1. pp. 51–61, doi: 10.1016/j.ejor.2009.09.032.

[26] Capar, I.—Kuby, M.: An Efficient Formulation of the Flow Refueling Location Model for Alternative-Fuel Stations. IIE Transactions, Vol. 44, 2012, No. 8, pp. 622–636, doi: 10.1080/0740817X.2011.635175.

[27] Mak, H. Y.—Rong, Y.—Shen, Z. J. M.: Infrastructure Planning for Electric Vehicles with Battery Swapping. Management Science, Vol. 59, 2013, No. 7, pp. 1557–1575, doi: 10.1287/mnsc.1120.1672.

[28] Tu, W.— Li, Q.—Fang, Z.—Shaw, S.—Zhou, B.—Chang, X.: Optimizing the Locations of Electric Taxi Charging Stations: A Spatial-Temporal Demand Coverage Approach. Transportation Research Part C: Emerging Technologies, Vol. 65, 2016, pp. 172–189, doi: 10.1016/j.trc.2015.10.004.

[29] Shahraki, N.—Cai, H.—Turkay, M.—Xu, M.: Optimal Locations of Electric Public Charging Stations Using Real World Vehicle Travel Patterns. Transportation Research Part D: Transport and Environment, Vol. 41, 2015, pp. 165–176, doi: 10.1016/j.trd.2015.09.011.

[30] Cai, H.—Jia, X.—Chiu, A. S. F.—Hu, X.—Xu, M.: Siting Public Electric Vehicle Charging Stations in Beijing Using Big-Data Informed Travel Patterns of the Taxi Fleet. Transportation Research Part D: Transport and Environment, Vol. 33, 2014, pp. 39–46, doi: 10.1016/j.trd.2014.09.003.

[31] Yang, J.—Dong, J.—Hu, L.: A Data-Driven Optimization-Based Approach for Siting and Sizing of Electric Taxi Charging Stations. Transportation Research Part C: Emerging Technologies, Vol. 77, 2017, pp. 462–477, doi: 10.1016/j.trc.2017.02.014.

[32] Dong, J.—Liu, C.—Lin, Z.: Charging Infrastructure Planning for Promoting Battery Electric Vehicles: An Activity-Based Approach Using Multiday Travel Data. Transportation Research Part C: Emerging Technologies, Vol. 38, 2014, pp. 44–55, doi: 10.1016/j.trc.2013.11.001.

[33] Xu, M.—Meng, Q.—Liu, K.—Yamamoto, T.: Joint Charging Mode and Location Choice Model for Battery Electric Vehicle Users. Transportation Research Part B: Methodological, Vol. 103, 2017, pp. 68–86, doi: 10.1016/j.trb.2017.03.004.

[34] Motoaki, Y.: Location-Allocation of Electric Vehicle Fast Chargers—Research and Practice. World Electric Vehicle Journal, Vol. 10, 2019, No. 1, Art. No. 12, doi: 10.3390/wevj10010012.

[35] He, J.—Yang, H.—Tang, T. Q.—Huang, H. J.: An Optimal Charging Station Location Model with the Consideration of Electric Vehicle's Driving Range. Transportation Research Part C: Emerging Technologies, Vol. 86, 2018, pp. 641–654, doi: 10.1016/j.trc.2017.11.026.

[36] He, F.—Wu, D.—Yin, Y.—Guan, Y.: Optimal Deployment of Public Charging Stations for Plug-In Hybrid Electric Vehicles. Transportation Research Part B: Methodological, Vol. 47, 2013, pp. 87–101, doi: 10.1016/j.trb.2012.09.007.

[37] He, F.—Yin, Y.—Lawphongpanich, S.: Network Equilibrium Models with Battery Electric Vehicles. Transportation Research Part B: Methodological, Vol. 67, 2014, pp. 306–319, doi: 10.1016/j.trb.2014.05.010.

[38] He, F.—Yin, Y.—Zhou, J.: Deploying Public Charging Stations for Electric Vehicles on Urban Road Networks. Transportation Research Part C: Emerging Technologies, Vol. 60, 2015, pp. 227–240, doi: 10.1016/j.trc.2015.08.018.

[39] Fredriksson, H.—Dahl, M.—Holmgren J.: Optimal Placement of Charging Stations for Electric Vehicles in Large-Scale Transportation Networks. Procedia Computer Science, Vol. 160, 2019, pp. 77–84, doi: 10.1016/j.procs.2019.09.446.

[40] López Millán, V. M.—Cholvi, V.—López, L.—Fernández Anta, A.: A Model of Self-Avoiding Random Walks for Searching Complex Networks. Networks, Vol. 60, 2012, No. 2, pp. 71–85, doi: 10.1002/net.20461.

[41] MATLAB, Version 9.4.0813654 (R2018a). The MathWorks Inc., Natick, Massachusetts, 2018.

[42] Inc. Gurobi Optimization. Gurobi Optimizer Reference Manual, 2018. Available at: `http://www.gurobi.com`.

Henrik Fredriksson received his B.Sc. in media technology from Blekinge Institute of Technology in 2011, and B.Sc. and M.Sc. in mathematics from Linnaeus University in 2012 and 2013, respectively. He is currently Ph.D. student in applied mathematics at the Department of Mathematics and Natural Sciences, Blekinge Institute of Technology. His scientific research is focused on mathematical models within the areas of transportation and traffic.

Mattias Dahl received his M.Sc. in computer engineering from Luleå Institute of Technology, 1993, Licentiate in Engineering, Lund University, 1997, and Ph.D. in applied signal processing, Blekinge Institute of Technology (BTH), 2000. Since 2005, he has been with the Department of Mathematics and Natural Sciences, BTH, where he is currently Professor of Systems Engineering. He has authored about 100 scientific publications and patents and has received several awards from the Swedish Innovation Agency and the Swedish Foundation of Technology Transfer.

Johan Holmgren is Associate Professor of computer science at Malmö University, Sweden. He received his Ph.D. degree in computer science at Blekinge Institute of Technology in 2010. Currently, he is employed at the Department of Computer Science and Media Technology at Malmö University. His research interests include agent–based simulation and modeling, mathematical optimization, simulation, and machine learning, and his application areas include freight transport and traffic modeling, public transport, and health-care logistics.

Computing and Informatics, Vol. 40, 2021, 428–445, doi: 10.31577/cai_2021_2_428

VERIFICATION OF LOCALIZATION VIA BLOCKCHAIN TECHNOLOGY ON UNMANNED AERIAL VEHICLE SWARM

Mustafa Cosar, Harun Emre Kiran

Department of Computer Engineering
Hitit University
North Campus
19030, Corum, Turkey
e-mail: {mustafacosar, harunemrekiran}@hitit.edu.tr

Abstract. Verification of the geographic location of a moving device is vital. This verification is important in terms of ensuring that the flying systems moving in the swarm are in orbit and that they are able to task completion and manage their energy efficiency. Cyber-attacks on unmanned aerial vehicles (UAV) in a swarm can affect their position and cause various damages. In order to avoid this challenge, it is necessary to share with each other the positions of UAV in the swarm and to increase their accuracy. In this study, it is aimed to increase position accuracy and data integrity of UAV by using blockchain technology in swarm. Experiments were conducted on a virtual UAV network (UAVNet). Successful results were obtained from this proposed study.

Keywords: Distributed network, UAVNet, localization, blockchain, Merkle algorithm, energy efficiency

1 INTRODUCTION

In recent years, the concept of blockchain in information technologies has started to come to the fore in areas such as finance, wireless sensor networks [1], medical applications [2], unmanned aerial vehicles [3], cyber security [4] and finance [5]. It is obvious that the majority of these areas focus on communication security. The starting point of the research is based on taking measures in order to increase

the coefficient of confidence during the communication of the devices used and to establish a healthier communication.

While blockchain is used in various fields such as cryptocurrencies, supply chain management and e-voting systems, it is also used to provide secure communication in autonomous systems. UAV Swarm Robot autonomy, decentralized control, collective decision making ability, high fault tolerance etc. It has some features such as Blockchain, a decentralized ledger managed by a peer-to-peer network with cryptographic algorithms, provides a platform for the secure execution of different transactions [6]. The decentralized nature of swarm robotics pushes it to think together with blockchain technology. It also allows it to implement different decentralized decision making, behavior differentiation and other business models. Blockchain technology has become an increasingly popular technology in recent years to provide decentralized trust and security in many digital systems, following its success with Bitcoin. Blockchain is used in many areas, which are listed below, as it eliminates the need for a trusted third party.

- Access management [7],
- Digital content distribution [8],
- Applied in areas such as supply chain management [9],
- Smart contracts in the Internet of Things (IoT) field [10, 11],
- Distribution and verification of sensitive business documents [12],
- Increasing privacy in healthcare [13, 14],
- Firmware update of embedded devices [15],
- Security of military autonomous systems [16],
- Swarm management, organization [11].

UAV uses many sensors such as Inertial Measurement Units (IMU), Laser, Global Positioning System (GPS) and Cameras to solve the positioning problem [17]. The UAV is only equipped with GPS equipment for location information due to various constraints such as cost, weight and range. The GPS signal can be easily affected by external interference, noise, receiving equipment failure [18] and cyber attacks. In real experiments for UAVs equipped with GPS only, it has been observed that some may temporarily lose their GPS connection for an extended period of time. In such a case, UAVs that lose their GPS connection have to be downloaded and their missions are canceled due to security concerns. To solve this problem, UAVs need location verification from within the swarm.

The addition of location information to the blockchain as described above provides control of the pre-recorded position evidence with distributed architecture. In this way, it can be stored and protected against attack. Furthermore, the blockchain as a distributed control and security system scenario, can provide the autonomy of UAV when communication channels from other components of UAVNet are lost [19].

Data packets are encrypted using cryptology to ensure secure data transmission in network technologies. However, in some cases, cryptology may be insufficient when a group of network nodes communicate among themselves. In the transmission between nodes, security measures are taken from the center according to the traditional architecture. In these centralized security methods, some major problems arise when there is no communication with the center at the time of the attack. As a solution to these problems, a new technology called blockchain is suggested when it is required to communicate securely between complex nodes.

The use of UAV for civil and military purposes is increasing day by day [20]. These vehicles have some disadvantages. For example, battery life, energy consumption values, physical and cyber-attacks [21]. There are many types of cyber-attacks on UAV networks. Examples of these cyber-attacks include disrupting communication broadcast, DoS / DDoS, buffer overflow, flooding can be shown as attacks [22]. Such attacks cause of UAVs to remain out of service, to fail the task, and even to crash. Another attack on UAV swarms is GPS attacks to change and disrupt their location information [23]. While the GPS signals of the UAVs used for military purposes are kept in an encrypted manner, this information of the civilian UAVs is transmitted directly in an unencrypted form [24].

The process of protecting the connection of devices in UAVNet is largely provided manually. This protection, which is done by encryption methods, is deprived of operational agility and transparency once applied. In the future, a kind of cryptographic infrastructure architecture should be simplified to distribute key data to the desired operating positions against the Man-In-The-Middle attack. In addition, some improvements in transmission methods should ensure that the network is protected against intrusion or interference, while the key information must be able to verify operational changes and users on the route, minimizing the burden on each UAV [19].

In this study, blockchain technology was applied in order to increase the accuracy and data integrity of geographic location information of different numbers of UAV swarms. In this way, against the attacks on the GPS position was tried to provide protection against the swarms. In addition, the energy consumption data of the UAVs were measured as a result of the application and another advantage of this technology was revealed.

In the next part of the study we provided the literature research, and in Section 3 the basic information was given. Section 4 introduces the proposed method and the results are given in Section 5. Conclusions are summarized in Section 6.

2 RELATED WORK

When the blockchain distributes directly to each UAV, it can significantly prevent the implementation of integrity and usability threats. Since each UAV has a copy of a blockchain, it can autonomously complete its course regardless of other ele-

ments of UAVNet. Knowing the location of neighbouring UAVs will prevent air collision [19].

Nowadays, the increased exposure of UAVs to cyber-attack due to their communication with wireless technologies leads to an increase in their work on their security [22]. In these studies, attentions are focused to position accuracy. In the literature, there are many studies on location accuracy of UAVs. The authors in [25] analyze the behaviour of the GPS deception attack targeting the GPS coordinates of the UAVs from the satellite. This study, although the author says that only the distribution of signal strength requires monitoring, does not give much detail on how to determine it. In [26], the author aimed at taking a precaution against GPS attack by means of multiple antennas. The solution made with this multi-antenna is an effective solution to the deterioration when used with the physical security function. This solution, however, is not cryptographic, and brings more weight and cost to the buyer. In many studies similar to these studies, a centralist security approach is recommended.

Since the blockchain is still a new technology, there are very few studies in the UAVNet area. In this area, there are studies such as data collection with data chain, protocol architecture [3] and data transfer [27]. A UAV that wants to ensure the position accuracy can verify the location via its short distance technology with the help of the UAVs in its neighbouring area [28]. In addition to neighbouring verification, we also recommend that the current location of the requesting UAV should not be changed with the previous location records thanks to the blockchain records.

As the use of blockchain technology increases in different areas, it is possible to use it in networks where devices such as UAVNet need to make geo-location accuracy. The fact that this kind of practice is not included in the literature has been evaluated as a motivation for this study. In this study, a model proposal has been made to verify the geographical location information of UAVs. In this model, it is based on a safe publication of adjacent UAVs by adding blockchain switch in a data packet for the location of a selected node of the selected UAVs.

Blockchain is a structure in which transactions and messages are electronically stored in blocks. It is also known that these messages and transactions are recorded by sending them to the entire network. If the messages pass the authentication test, one more block is added to the chain. Verifying any message on the blockchain is extremely simple and only requires a single hashing. In literature, numerous methods have been proposed in the literature to filter false reports from networks. All these schemes are based on collaborative report approval and hop-by-hop report verification [29].

3 CONCEPTUAL FRAMEWORK

The UAV requesting location verification from neighbouring UAVs in blockchain technology is protected against attack because they have signed the information in

the data packets they use with their private keys. However, due to GPS attacks, the UAV cannot determine its correct position. Thanks to the close distance technology [28] used in this study, it can determine its actual position with evidence from neighbouring GPSs. However, if the neighbouring UAVs are also attacked after the control center publishes these packages:

- If the UAV which requests location verification, has position information already registered in the blockchain, the maximum distance that this UAV can go with the old records is calculated. If the new location info is greater than the maximum destination, this package is not used in the blockchain. In such a case, if the attacker wants to accept the package, he either has to change the blockchain or must capture more than 50 % of the system. The capture of such a system is almost impossible [28].
- If the UAVs requesting proof of location have two or more different location proofs in their neighbour UAVs, the high number of approved proof packages will be approved.

The addition of location information to the blockchain as described above provides control of the pre-recorded position evidence with distributed architecture. In this way, it can be stored and protected against attack. Furthermore, the blockchain as a distributed control and security system scenario, can provide the autonomy of UAV when communication channels from other components of UAVNet are lost [19].

3.1 Distributed Consensus

The concept of blockchain was first in the economy sector with the crypto currency called bitcoin [5]. The blockchain consists of each block and all blocks are connected to each other. Each block is created according to the consensus mechanism [30].

Consensus mechanisms allow nodes in the network to trust others. The four most popular consensus mechanisms, according to [31], are Proof of Work (PoW) [32], Proof of Stake (PoS) [33], Practical Byzantine Fault Tolerance (PBFT) and Delegated Proof of Stake (DPoS), with other significant approaches including Proof of Authority (PoA), Proof of Elapsed Time (PoET) or Proof of Bandwidth (PoB). Of these PoW is considered to be a disadvantage because of the resource requirements of the systems that will produce the block [34].

The model we propose in this study is closer to the PoS algorithm in terms of block creation process. In UAVNet, as in this algorithm, it will be tasked to create a block to one of the UAVs whose neighbours receive the highest approval over the limit value. The UAVs exceeding the maximum limit shall have the right to be elected in proportion to the number of approved UAVs. Any of these UAVs will be randomly selected. This UAV is assigned by the control center to be selected in a random time and the task of creating new blocks. The first task of the UAV

that has taken the task of creating a block is to summarize the location evidence packages of the UAVs with the current blockchain summary and combine it with the Merkle algorithm [35].

Then, to add new blocks to this UAV; it creates a data packet by adding the credential, location records, summary of the previous block, summary of the candidate block, and block creation time. Finally, it signs it with its own private key and transmits this package to other UAVs through the control center. If the majority of UAVs accepts the request to create this new block, the data packet is added to the blockchain, as shown in Figure 1.

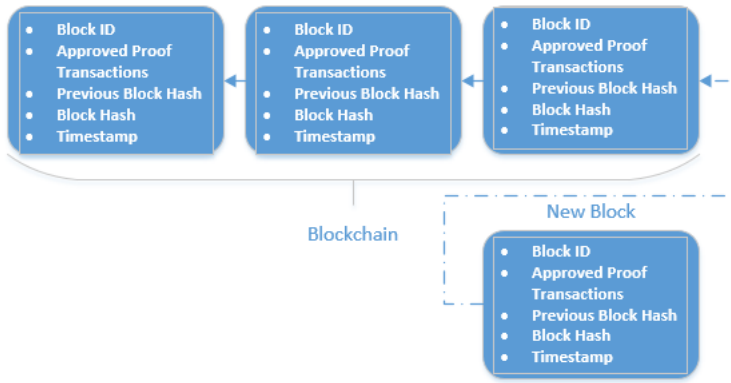


Figure 1. Data package for adding new block to blockchain

3.2 Merkle Tree

A Merkle Tree is a binary tree, a way of iteratively repeating to combine the two hashes in each transaction of a block, then hashing the two hashed transactions again and concatenating them two by one until they become one. The Summary Syntax Tree is a way to bind each function until all of the program's dependencies have been matched. A general structure of the Merkle Tree is shown in Figure 2. A Merkle tree is a secure and efficient [36] chain structure used to verify the consistency of a large set of data records. This structure improves transaction validation performance on blocks.

Each parent node derives its hash value from the value of its children that are recursively dependent on all values in its subtree. Figure 2 shows an example of a Merkle tree, each leaf (H1–H4) gets its value by calculating the imported value (D1–D4) and parents (H5–H6) get values from their children (H1–H4) and finally the root. The value of this Merkle tree (H7) corresponding to each value in the

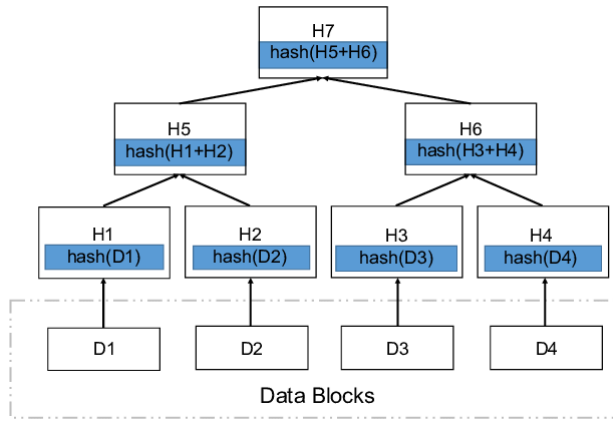


Figure 2. A general structure of the Merkle Tree [37]

tree is obtained. For example, H4 and H5 are needed to verify if H3 is in the tree. A root (H8) can be calculated using H3, H4, and H5. By comparing H7 and H8 we can confirm that H3 is in the tree if the two roots are the same, or that H3 is not in the tree if the two roots are different [37]. The top node named Top Hash [38] represents the Merkle root. All child nodes are leaf nodes and intermediate hash nodes are branches. The leaf nodes of the Merkle tree calculate hashes of numbers proportional to the logarithm, while the number of proportional leaf nodes has lists of hashes.

In the blockchain, a public key is used as the identity of the user and a private key is used to authenticate the data [39]. Also, the blockchain uses a Merkle tree in 1 blocks. In a Merkle tree, changing a value also changes the hash of the entire tree. However, before adding data to the blockchain, each miner must reach an agreement on the validity of the data [40]. Once accepted by everyone, the data is included in the blockchain. After adding data to the blockchain, no changes can be made. If someone tries to make a change to the block, the hash of the block also changes and breaks the blockchain. All validators must agree on this change to reconfigure the chain. Thus, the data on the blockchain remains secure. These blockchain features can be a potential solution to the above mentioned security threats (i.e. cyber attacks, data integrity issue).

3.3 UAV and Swarm Organization

Although UAVs were originally designed for military purposes, they are also preferred in civilian applications due to their promising functions such as ease of deployment, low maintenance cost and usability [41, 42]. Also, a drone swarm can power IoTs by acting as a relay to transmit data. The use of autonomous systems such as unmanned aerial vehicles has greatly facilitated military operations and is successful

in collecting sensitive data and transmitting it to the command center. However, due to the hardware and software components it contains, it has become the target of cyber attacks. Examples of these attacks are manipulating the content of critical messages used in the decision-making of autonomous systems, changing the flight path and changing the current location information. To ensure the successful operation of autonomous military systems [15], it is necessary to develop mechanisms that will strictly protect the integrity of data and messages collected/exchanged, and to provide an immutable record of each message.

Particle swarm optimization and ant colony optimization are the two main techniques in the swarm intelligence family. In particle swarm optimization (PSO), a swarm of particles is placed in a hypothetical solution space with multiple constraints to be met. The particle's global best position and velocity guide the swarm to reach the optimum position and velocity. PSO is a population-based technique and can be effectively used for route optimization problem [43].

3.4 Location Verification Process of a UAV

1. A UAV wanting to verify location, as shown in Figure 3, signs its own identity, the current location information, the summary of the blockchain, the creation time of the last block added to the chain, the time to create the package and the public key in a package and sign it with its private key. Then, the UAV issues this signed data package with the original package to the neighbour nodes and waits for a while. This UAV makes a publication covering the locations of neighbouring UAVs close to it, taking into account the energy consumption when broadcasting the package. If a certain number of neighbouring UAVs have not received approval during this period, they will repeat the broadcast by increasing the area to send the package. This process continues in the form of iterations until the confirmation of the determined number of neighbours is received.

Figure 3. The contents of the location verification in data package that the UAV sends to neighbouring UAVs

2. Neighbouring UAVs compare the contents of the data packet from the UAV that wants to verify its position with a summary of the current blockchain available in them. Because of the comparison, if the summaries of the blockchains are equalized, the neighbouring UAV approves its position information in the chain.

Then, comparing the requesting UAV's position information with the information in its chain, it summarizes its answer with its own private key, including the information in Figure 4, adds it to the data package and sends it to the control center.

Figure 4. Content of the data package prepared by the neighbouring UAV for the UAV that wants to verify its position

3. The control center broadcasts this package to the other UAVs without making any changes. In fact, the neighbouring UAVs proving the accuracy of the location can broadcast this package directly to all UAVs. However, this publication is made by the center because the UAVs in the swarm may be scattered over a wide area and their battery capacity may have decreased.

3.5 Security

Information security is a concept that has been studied since the beginning of computing. Also, some specialized fields such as cryptography have been explored earlier than this. The main objectives of security requirements are: confidentiality, authentication, availability, integrity, and non-repudiation [42]. Cyber security comes to the fore when computers are connected to each other.

Wireless sensor networks are an easy target for report generation attacks where compromised sensor nodes can be used by an attacker to flood the network with fake/false reports. Pathway filtering is a mechanism where intermediate forwarding nodes identify and drop false reports as they are routed to the pool. Current path-through filtering schemes have either high storage overhead or low filtering efficiency [44]. As it is known, DoS, DDoS, MITM ataks, non-repudiation, content poisoning [45] are cyber attacks on UAVs. In addition, attackers UAVs can launch alteration attack to inject, delete or modify any message. Therefore, they may maliciously respond with data packets modified to meet the consumer interests, resulting in cache poisoning.

4 METHOD

The communication architecture with the neighbours for the location verification of a UAV within the UAV swarm is shown in Figure 4. The most remarkable innovation

in this model is the use of blockchain technology. When a communication started, it was hashed into the code sent by a block and then broadcast to each node. Since thousands of transaction records can be processed in each node's block, the blockchain uses the Merkle Tree function to generate a final hash, which is the Merkle tree root.

The reason we included the Merkle Tree method in the formation of the blockchain is that it leads to a decrease in the block propagation speed between points. In [36], performing an application with the blockchain and Merkle tree simulator, showed that block transactions have a high effect of reducing the verification time by up to 30 times, with no effect on the block propagation delay. This latest hash value will be saved in the block header (the hash of the current block), thus greatly reducing data transmission and system resources using the Merkle Tree function.

As shown in Figure 5, UAV, which is orange in its background, transmits a broadcast to the neighbour UAVs to ensure location verification. The vector position deviation that occurs during the attack of the selected UAV is planned to be equal to the position deviations of the neighbouring UAVs.

The location verification process we have tried to summarize in 3 items above in Section 3, we experimented with a lot of 100 UAVs in the simulation environment. In this study, 10 different flight plans from 500 meters to 5 000 meters were conducted with UAVs. The energy consumption and position verification information of the swarm which is distributed randomly is tried to be calculated. Experiments were repeated 10 000 times in order not to be affected by different parameters, then an average value was calculated.

The information obtained using GPS trajectory data is becoming more comprehensive, detailed and accurate [46]. UAVs need accurate location information for a variety of purposes, including route planning, operations, control and mission completion [17]. Most UAVs use location information; global positioning system (GPS), inertial navigation system, or a combination of both [47].

5 RESULTS

5.1 Location Verification During Attack

During the location verification process, a simulation was made with 100 UAVs in the 5 000 m × 5 000 m area. In order to determine how the location verification of the UAVs was affected during the attack, one of the UAVs with randomly selected at least 6 neighbours was attacked with GPS Spoofing. The resulting position deviation of the attacked UAV was applied to the neighbours UAVs as the position deviation vector of the same value. This experiment was repeated 10 000 times in order to minimize the effects of the parameters that could not be taken into account.

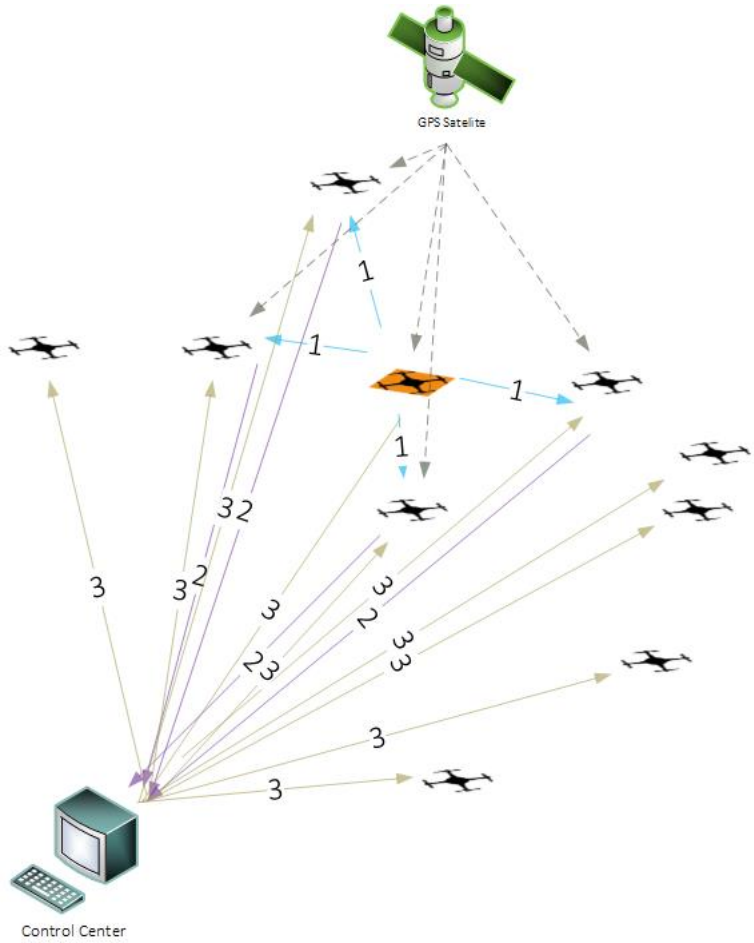


Figure 5. Location verification model among neighbours with blockchain technology in UAV swarm

As shown in Table 1, the number of attacked neighbouring UAVs increased as a result of a decrease in the position verification performance of UAVs. However, even with a decrease in performance rate, even if the attack rate is the highest, it is determined that the UAVs in the network, such as 84 %, have a high position accuracy.

When GPS is the Spoofing attack, a certain period of time is expected as the UAVs cannot be contacted immediately. In this study, since the UAVs are assumed to be attacked as soon as they start flying, the initial position information is not added to the blockchain. If the instant location information can be added to the blockchain in the first time interval, it is thought that these performance rates will

Number of attacked UAVs	Position verification performance of UAVs [%]
0	99.9
20	99.7
40	98.5
60	92.2
80	87.4
100	84.3

Table 1. Position accuracy performance of UAV, which will make position verification, based on the increase in the number of attacked UAVs

increase to close to 100 %. Because, even if an attack occurs, the positions stored in the blockchain will not be affected and the UAVs will be able to broadcast the correct position information.

5.2 Energy Consumption

As a result of the experiments, the second calculated value was energy consumption data. This value was calculated by taking the average of all experiments. It is known that the energy consumption values of UAVs increase as the flight range increases. As shown in Figure 6, it was determined that the UAVs communicating via the control center consumed more energy than the UAVs that communicated directly among them.

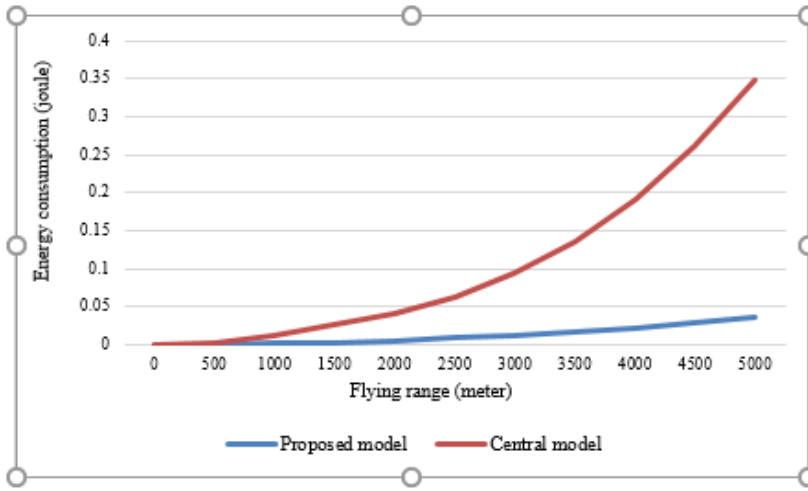


Figure 6. Energy consumption between models

6 CONCLUSIONS

In the proposed model, a hybrid method was used by combining UAV-Neighboring communication and UAV-Control center communication. In particular, in the position verification phase, UAV-Neighbour communication was made according to the possibility of attack of the control center. Blockchain technology was used to increase the reliability coefficient of this communication. Evidence validation was performed with special key in the chain formed by the merkle algorithm. With this model, a UAV deviating from orbit at the time of an attack may request verification from its neighbours. In addition, it is thought that performing location verification by block broadcasting between neighbors will increase energy efficiency, with the thought that broadcasting new locations to the entire swarm by the control center at the end of the location verification process of UAVs will increase energy consumption. Since the control center is included in the chain here, the new location information will also reach it. It can be said that it is unnecessary to verify by contacting the control center again.

In this study, it is aimed to increase the coefficient of trust by the blockchain technology of the location verification process of the UAV flock under attack. During CPS-attack in a UAV swarm, a UAV can verify a position information added to the blockchain, at a rate close to 100 % when from neighboring UAVs require verification. Even if the number of attacked UAVs increased, the verification rate did not fall below 80 %. When the energy consumption values with our model are examined, it is seen that there is a decrease in the rate of 8 times.

An attacker could compromise multiple sensor nodes to inject false reports into the network. These false reports claim events that do not exist at random locations on the network, causing the pool to make incorrect decisions. Therefore, such attacks can cause mission-critical networks to fail. Thanks to this blockchain developed for UAVs,

- Avoiding excessive signature verification for each Data packet,
- Refrain from passing on Interests to potential attackers, provided.

These advantages ensured energy efficiency. Blockchain technology is recommended against some attacks such as content poisoning. However, the method of moving block data to the cache may have a degrading effect on system performance.

REFERENCES

[1] Dorri, A.—Kanhere, S. S.—Jurdak, R.—Gauravaram, P.: Blockchain for IoT Security and Privacy: The Case Study of a Smart Home. 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2017, pp. 618–623, doi: 10.1109/PERCOMW.2017.7917634.

[2] Azaria, A.—Ekblaw, A.—Vieira, T.—Lippman, A.: MedRec: Using Blockchain for Medical Data Access and Permission Management. 2016 2nd In-

ternational Conference on Open and Big Data (OBD), 2016, pp. 25–30, doi: 10.1109/obd.2016.11.

[3] Kapitonov, A.—Lonshakov, S.—Krupenkin, A.—Berman, I.: Blockchain-Based Protocol of Autonomous Business Activity for Multi-Agent Systems Consisting of UAVs. 2017 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS), 2017, pp. 84–89, doi: 10.1109/red-uas.2017.8101648.

[4] Liang, G.—Weller, S. R.—Luo, F.—Zhao, J.—Dong, Z. Y.: Distributed Blockchain-Based Data Protection Framework for Modern Power Systems Against Cyber Attacks. IEEE Transactions on Smart Grid, Vol. 10, May 2019, No. 3, pp. 3162–3173, doi: 10.1109/TSG.2018.2819663.

[5] Nguyen T. T.—Hatua, A.—Sung A. H.: Blockchain Approach to Solve Collective Decision Making Problems for Swarm Robotics. In: Prieto, J., Das, A., Ferretti, S., Pinto, A., Corchado, J. (Eds.): Blockchain and Applications (BLOCKCHAIN 2019). Springer, Cham, Advances in Intelligent Systems and Computing, Vol. 1010, 2020, pp. 118–125, doi: 10.1007/978-3-030-23813-1_15.

[6] Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System. Self-Publication, available at: `https://bitcoin.org/bitcoin.pdf`, 2009.

[7] Di Pietro, R.—Salleras, X.—Signorini, M.—Waisbard, E.: A Blockchain-Based Trust System for the Internet of Things. Proceedings of the 23rd ACM on Symposium on Access Control Models and Technologies (SACMAT '18), 2018, pp. 77–83, doi: 10.1145/3205977.3205993.

[8] Kishigami, J.—Fujimura, S.—Watanabe, H.—Nakadaira, A.—Akutsu, A.: The Blockchain-Based Digital Content Distribution System. 2015 IEEE 5th International Conference on Big Data and Cloud Computing, Dalian, China, 2015, pp. 187–190, doi: 10.1109/BDCloud.2015.60.

[9] Hofmann, E.—Johnson, M.: Supply Chain Finance – Some Conceptual Thoughts Reloaded. International Journal of Physical Distribution and Logistics Management, Vol. 46, 2016, No. 4, pp. 1–8, doi: 10.1108/IJPDLM-01-2016-0025.

[10] Christidis, K.—Devetsikiotis, M.: Blockchains and Smart Contracts for the Internet of Things. IEEE Access, Vol. 4, 2016, pp. 2292–2303, doi: 10.1109/ACCESS.2016.2566339.

[11] Islam, A.—Shin, S. Y.: BUS: A Blockchain-Enabled Data Acquisition Scheme with the Assistance of UAV Swarm in Internet of Things. IEEE Access, Vol. 7, 2019, pp. 103231–103249, doi: 10.1109/ACCESS.2019.2930774.

[12] Aitzhan, N. Z.—Svetinovic, D.: Security and Privacy in Decentralized Energy Trading Through Multi-Signatures, Blockchain and Anonymous Messaging Streams. IEEE Transactions on Dependable and Secure Computing, Vol. 15, 2018, No. 5, pp. 840–852, doi: 10.1109/TDSC.2016.2616861.

[13] Aydar, M.—Çetin, S.: Blockchain for Health Information Systems. European Journal of Science and Technology, 2020, No. 19, pp. 533–538, doi: 10.31590/ejosat.735052 (in Turkish).

[14] Yue, X.—Wang, H.—Jin, D.—Li, M.—Jiang, W.: Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control. Jour-

nal of Medical Systems, Vol. 40, 2016, No. 10, Art. No. 218, doi: 10.1007/s10916-016-0574-6.

[15] Lee, B.—Lee, J. H.: Blockchain-Based Secure Firmware Update for Embedded Devices in an Internet of Things Environment. The Journal of Supercomputing, Vol. 73, 2017, No. 3, pp. 1152–1167, doi: 10.1007/s11227-016-1870-0.

[16] Angin, P.: Blockchain-Based Data Security in Military Autonomous Systems. European Journal of Science and Technology, Special Issue, 2020, pp. 362–368, doi: 10.31590/ejosat.824196.

[17] Abdelkrim, N.—Aouf, N.—Tsourdos, A.—White, B.: Robust Nonlinear Filtering for INS/GPS UAV Localization. 2008 16th Mediterranean Conference on Control and Automation, 2008, pp. 695–702, doi: 10.1109/MED.2008.4602149.

[18] Mao, G.—Drake, S.—Anderson, B. D. O.: Design of an Extended Kalman Filter for UAV Localization. Information, Decision and Control, Adelaide, SA, Australia, 2007, pp. 224–229, doi: 10.1109/IDC.2007.374554.

[19] Kuzmin, A.—Znak, E.: Blockchain-Base Structures for a Secure and Operate Network of Semi-Autonomous Unmanned Aerial Vehicles. 2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), 2018, pp. 32–37, doi: 10.1109/SOLI.2018.8476785.

[20] Goh, G. D.—Agarwala, S.—Goh, G. L.—Dikshit, V.—Sing, S. L.—Yeong, W. Y.: Additive Manufacturing in Unmanned Aerial Vehicles (UAVs): Challenges and Potential. Aerospace Science and Technology, Vol. 63, 2017, pp. 140–151, doi: 10.1016/j.ast.2016.12.019.

[21] Mansfield, K.—Eveleigh, T.—Holzer, T. H.—Sarkani, S.: Unmanned Aerial Vehicle Smart Device Ground Control Station Cyber Security Threat Model. 2013 IEEE International Conference on Technologies for Homeland Security (HST), 2013, pp. 722–728, doi: 10.1109/THS.2013.6699093.

[22] Javaid, A. Y.—Sun, W.—Devabhaktuni, V. K.—Alam, M.: Cyber Security Threat Analysis and Modeling of an Unmanned Aerial Vehicle System. 2012 IEEE Conference on Technologies for Homeland Security (HST), 2012, pp. 585–590, doi: 10.1109/ths.2012.6459914.

[23] He, L.—Li, W.—Guo, C.—Niu, R.: Civilian Unmanned Aerial Vehicle Vulnerability to GPS Spoofing Attacks. 2014 Seventh International Symposium on Computational Intelligence and Design, 2014, pp. 212–215, doi: 10.1109/iscid.2014.131.

[24] He, D.—Chan, S.—Guizani, M.: Communication Security of Unmanned Aerial Vehicles. IEEE Wireless Communications, Vol. 24, 2017, No. 4, pp. 134–139, doi: 10.1109/mwc.2016.1600073wc.

[25] Shepard, D. P—Bhatti, J. A.—Humphreys, T. E.—Fansler, A. A.: Evaluation of Smart Grid and Civilian UAV Vulnerability to GPS Spoofing Attacks. ION GNSS Conference, Nashville, TN, 2012.

[26] Magiera, J.—Katulski, R.: Detection and Mitigation of GPS Spoofing Based on Antenna Array Processing. Journal of Applied Research and Technology, Vol. 13, 2015, No. 1, pp. 45–57, doi: 10.1016/S1665-6423(15)30004-3.

[27] Liang, X.—Zhao, J.—Shetty, S.—Li, D.: Towards Data Assurance and Resilience in IoT Using Blockchain. 2017 IEEE Military Communications Conference (MILCOM), 2017, pp. 261–266, doi: 10.1109/MILCOM.2017.8170858.

[28] Zhu, Z.—Cao, G.: Toward Privacy Preserving and Collusion Resistance in a Location Proof Updating System. IEEE Transactions on Mobile Computing, Vol. 12, 2013, No. 1, pp. 51–64, doi: 10.1109/tmc.2011.237.

[29] Kumar, A.—Pais, A. R.: Blockchain Based En-Route Filtering of False Data in Wireless Sensor Networks. 2019 11th International Conference on Communication Systems and Networks (COMSNETS), 2019, pp. 1–6, doi: 10.1109/COMSNETS.2019.8711352.

[30] Dey, S.: A Proof of Work: Securing Majority-Attack in Blockchain Using Machine Learning and Algorithmic Game Theory. International Journal of Wireless and Microwave Technologies, Vol. 8, 2018, No. 5, pp. 1–9, doi: 10.5815/ijwmt.2018.05.01.

[31] Vasin, P.—Co, B.: BlackCoin's Proof-of-Stake Protocol v2. Self-Publication, available at: `http://blackcoin.co/blackcoin-pos-protocol-v2-whitepaper.pdf`, 2014.

[32] Queralta, J. P.—Westerlund, T.: Blockchain-Powered Collaboration in Heterogeneous Swarms of Robots. 2019 Symposium on Blockchain for Robotics and AI Systems, 16 pp., 2020, arXiv: arXiv:1912.01711v3.

[33] Conoscenti, M.—Vetrò, A.—De Martin, J. C.: Blockchain for the Internet of Things: A Systematic Literature Review. IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), 2016, pp. 1–6, doi: 10.1109/AICCSA.2016.7945805.

[34] Dorri, A.—Kanhere, S. S.—Jurdak, R.: Towards an Optimized BlockChain for IoT. 2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI), 2017, pp. 173–178.

[35] King, S.—Nadal, S.: PPCoin: Peer-to-Peer Crypto-Currency with Proof-of-Stake. Self-Publication, available at: `https://archive.org/details/PPCoinPaper`, 2012.

[36] Fattahi, S. M.—Makanju, A.—Milani Fard, A.: SIMBA: An Efficient Simulator for Blockchain Applications. 2020 50th Annual IEEE-IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S), 2020, pp. 51–52, doi: 10.1109/DSN-S50200.2020.00028.

[37] Huang, H.—Lin, J.—Zheng, B.—Zheng, Z.—Bian, J.: When Blockchain Meets Distributed File Systems: An Overview, Challenges, and Open Issues. IEEE Access, Vol. 8, 2020, pp. 50574–50586, doi: 10.1109/ACCESS.2020.2979881.

[38] Bosamia, M.—Patel, D.: Current Trends and Future Implementation Possibilities of the Merkle Tree. International Journal of Computer Sciences and Engineering, Vol. 6, Aug 2018, No. 8, pp. 294–301, doi: 10.26438/ijcse/v6i8.294301.

[39] Islam, A.—Uddin, M. B.—Kader, M. F.—Shin, S. Y.: Blockchain Based Secure Data Handover Scheme in Non-Orthogonal Multiple Access. Proceedings of the 4th International Conference on Wireless Telematics (ICWT), 2018, pp. 1–5, doi: 10.1109/ICWT.2018.8527732.

[40] Dinh, T. T. A.—Liu, R.—Zhang, M.—Chen, G.—Ooi, B. C.—Wang, J.: Untangling Blockchain: A Data Processing View of Blockchain Systems. IEEE Transactions on Knowledge and Data Engineering, Vol. 30, 2018, No. 7, pp. 1366–1385, doi: 10.1109/TKDE.2017.2781227.

[41] Merkle, R. C.: Protocols for Public Key Cryptosystems. IEEE Symposium on Security and Privacy, 1980, pp. 122–134, doi: 10.1109/sp.1980.10006.

[42] Hayat, S.—Yanmaz, E.—Muzaffar, R.: Survey on Unmanned Aerial Vehicle Networks for Civil Applications: A Communications Viewpoint. IEEE Communications Surveys and Tutorials, Vol. 18, 2016, No. 4, pp. 2624–2661, doi: 10.1109/COMST.2016.2560343.

[43] Gupta, L.—Jain, R.—Vaszkun, G.: Survey of Important Issues in UAV Communication Networks. IEEE Communications Surveys and Tutorials, Vol. 18, 2016, No. 2, pp. 1123–1152, doi: 10.1109/COMST.2015.2495297.

[44] Cizmar, A.—Papaj, J.—Dobos, L.: Security and QoS Integration Model for MANETs. Computing and Informatics, Vol. 31, 2012, No. 5, pp. 1025–1044, retrieved from http://www.cai.sk/ojs/index.php/cai/article/view/1187.

[45] Lei, K.—Zhang, Q.—Lou, J.—Bai, B.—Xu, K.: Securing ICN-Based UAV Ad Hoc Networks with Blockchain. IEEE Communications Magazine, Vol. 57, 2019, No. 6, pp. 26–32, doi: 10.1109/MCOM.2019.1800722.

[46] Zhu, S.—Sun, H.—Duan, Y.—Dai, X.—Saha, S.: Travel Mode Recognition from GPS Data Based on LSTM. Computing and Informatics, Vol. 39, 2020, No. 1-2, pp. 298–317, doi: 10.31577/cai_2020_1-2_298.

[47] Sasiadek, J. Z.—Hartana, P.: Sensor Fusion for Navigation of an Autonomous Unmanned Aerial Vehicle. IEEE International Conference on Robotics and Automation (ICRA '04), Vol. 4, 2004, pp. 4029–4034, doi: 10.1109/robot.2004.1308901.

Mustafa Cosar received his B.Sc. degree in computer science from the Karadeniz Technical University. He received his Ph.D. degree in computer education from Gazi University, Ankara, Turkey, in 2013. Since 2013, he has been Assistant Professor with the Computer Engineering Department, Hitit University. At the same time, he worked as an IT manager at Hitit University between 2007 and 2017. He is the author of more than 10 articles, and more than 45 conference papers. His research interests include computer networks and cybersecurity, localization, computer forensics and IDS/IPS architecture.

Harun Emre Kiran received his B.Sc. in computer engineering from the Kırıkkale University, Kırıkkale, Turkey. Also, he received his M.Sc. degree in computer engineering from Karadeniz Technical University in 2019. He is currently pursuing the Ph.D. degree in the Deparment of Computer Engineering, Gazi University, Ankara, Turkey. His research interests are mainly in the areas of wireless sensor network and network security.

Computing and Informatics, Vol. 40, 2021, 446–468, doi: 10.31577/cai_2021_2_446

METHOD FOR REPAIRING PROCESS MODELS WITH SELECTION STRUCTURES BASED ON TOKEN REPLAY

Erjing Bai, Na Su

Qingdao Huanghai University
Qingdao 266427, China

Yu Liang

College of Electronics and Information Engineering
Tongji University
Shanghai 201804, China

Liang Qi*, Yuyue Du

College of Computer Science and Engineering
Shandong University of Science and Technology
Qingdao 266590, China
e-mail: 1832678460@qq.com, yydu001@163.com

Abstract. Enterprise information systems (EIS) play an important role in business process management. Process mining techniques that can mine a large number of event logs generated in EIS become a very hot topic. There always exist some deviations between a process model of EIS and event logs. Therefore, a process model needs to be repaired. For the process model with selection structures, the mining accuracy of the existing methods is reduced because of the additional self-loops and invisible transitions. In this paper, a method for repairing Logical-Petri-nets-based process models with selection structures is proposed. According to the relationship between the input and output places of a sub-model, the de-

* Corresponding author

viation position is determined by a token replay method. Then, some algorithms are designed to repair the process models based on logical Petri nets. Finally, the effectiveness of the proposed method is illustrated by some experiments, and the proposed method has relatively high fitness and precision compared with its peers.

Keywords: Logic Petri net, model repair, token replay, choice structures, process model

1 INTRODUCTION

Business process management has significantly promoted the development of company business processes with the help of advanced enterprise information systems (EIS). Meanwhile, a large number of event logs are generated every day [1]. These event logs can be mined which in turn improve EIS [2] and further the competitiveness of enterprises or organizations. Process mining can extract valuable information from the event logs and improve the actual process models [3]. Process mining techniques mainly include process discovery, conformance checking, and process enhancement [1, 2, 3, 4]. For process discovery, a process model can be built by mining the existing event logs. Conformance checking can find the deviations between a process model and the event logs [4]. For process enhancement, a process model can be expanded and improved by further studying event logs [5]. At present, many process discovery algorithms have been proposed by scholars. A reasonable workflow model with complete logs can be mined based on α algorithm [6], but non-free-choice structures and invisible transitions cannot be well handled. Some extension methods of α algorithm are proposed in [7, 8] to deal with the above problems. A proposed approach based on the genetic algorithm [9, 10] can guarantee a certain quality standard, but it restricts the accurate discovery of block-structured process models and has high computational complexity. A repairing method proposed by Fahland et al. has high fitness [5], but the precision is low because of self-loops and invisible transitions. A single activity with self-loops is inserted into the original models based on Goldratt's and Knapsack's method [11], but the precision is still not high.

A process model is described by Petri nets in [12], since Petri nets have the advantages of rigorous mathematical definitions and powerful graphic display. The static and dynamic states of business processes can be described in Petri nets. However, the existing process models cannot completely replay event logs when business processes or environments are changed. The repaired can replay most of the logs without breaking the main structure of the original model [13]. In this paper, a method for repairing process models with selection structures based on logical Petri nets is proposed. First, model deviations are determined by calculating missing and remaining-tokens; then, the process model is repaired according to the deviations.

The proposed method has higher fitness and precision than Fahland's and Goldratt's methods [5, 11].

The rest of the paper is organized as follows. Section 2 presents some preliminaries and briefly reviews some important concepts. Section 3 presents an approach to repair models with selection structures based on a token replay method via logical Petri nets. The results and performance analysis of simulation experiments are given in Section 4. Section 5 concludes this paper and discusses the future work.

2 PRELIMINARIES

Some basic concepts are introduced in this section including multi-sets [3], tuple [14], event logs, projection, Petri nets [14], logical Petri nets [15], process trees, and workflow nets.

Definition 1 (Multi-sets [3])**.** S is a set. A multi-set Z over S is denoted by $Z : S \to N^+$, N^+ represents a set of positive integers. $\beta(S)$ represents all multi-sets over S.

Definition 2 (Tuple [14])**.** A tuple consisting of n elements is denoted by $x = (a_1, a_2, \ldots, a_n) \in S \times \cdots \times S$, where S is a set. $\pi_i(x)$ is the i^{th} element of x, where $i \in (1, 2, \ldots, n)$.

Definition 3 (Trace and event log [16])**.** Let A be a set of actives. $\sigma \in A^*$ is a trace if $1 \leq i < j \leq \sigma : \sigma[i] \neq \sigma[j]$, and it is a queue of actives. $(\&\sigma)$ represents the set of all activities in trace σ. An event log is a finite nonempty multi-set of trace σ, denoted as $L \in \beta(A^*)$.

For example, given an activity set $A = \{t_1, t_2, t_3, t_4\}$, $\sigma = \langle t_2 t_3 t_1 t_4 \rangle$ is a trace and $(\&\sigma) = \{t_2, t_3, t_1, t_4\}$.

Definition 4 (Projection)**.** Let β be a multi-set over A, $Q \subseteq A$, and $\sigma \in A^*$. $\sigma|Q$ denotes the projection of σ on Q, and $\beta|Q$ denotes the projection of β on Q.

For example, if $\sigma = \langle aabc \rangle$, $Q = \{a, c\}$, and $\beta = [a^3, b, c^2]$, then $\sigma|Q = \langle aac \rangle$ and $\beta|Q = [a^3, c^2]$.

Definition 5 (Petri net)**.** A Petri net is a four-tuple $PN = (P, T; F, M)$, where P is a finite place set, T is a finite transition set, and $F \subseteq (P \times T) \cup (T \times P)$ is a finite arc set, where

1. $N = (P, T; F)$ is a net;
2. $M : P \to \{0, 1, 2, \ldots\}$ is a marking of N; and
3. The transition firing rules of Petri nets are as follows:
   (a) For transition $t \in T$, if $\forall p \in^\bullet t : M(p) \geq 1$, then t is enabled at M, denoted as $M[t >$.

(b) If $M[t>$, then transition t can fire at M, and it generates a new marking M', denoted as $M[t > M'$. For $\forall p \in P$, we have

$$M'(P) = \begin{cases} M(P) - 1, & p \in {}^\bullet t - t^\bullet; \\ M(P) + 1, & p \in t^\bullet - {}^\bullet t; \\ M(P), & \text{otherwise.} \end{cases}$$

Definition 6 (Pre-set and post-set [15]). Let $N = (P, T; F)$ be a net. For $x \in P \cup T$, ${}^\bullet x$ is the pre-set of x if ${}^\bullet x = \{y | y \in P \cup T \wedge (y, x) \in F\}$. $x^\bullet$ is the post-set of x if $x^\bullet = \{y | y \in P \cup T \wedge (x, y) \in F\}$.

Definition 7 (Workflow net [17]). $WFN = (P, T; F, M, i, o)$ is a workflow net, where P, T, F and M can constitute a Petri net; i represents an input place and o represents an output place, where

1. There is an input place $i \in P$, ${}^\bullet i = \phi$ and M_i is an initial marking;
2. There is an output place $o \in P$, $o^\bullet = \phi$, and M_o is a final marking; and
3. $x \in P \cup T$ is always on the path from i to o.

Definition 8 (Logic Petri net). A logic Petri net is a six-tuple denoted by $LPN = (P, T; F, I, O, M)$, where

1. P represents a finite set of places;
2. $T = T_D \cup T_I \cup T_O$ represents a finite set of transitions, and $T \cap P = \phi$. If $t \in T_I \cap T_O$, then ${}^\bullet t \cap t^\bullet = \phi$. T_D represents a set of traditional transitions in Petri nets. T_I represents a set of logic input transitions. For $\forall t \in T_I$, ${}^\bullet t$ is restricted by a logical input expression $f_I(t)$. T_O represents a logic output transitions set. For $\forall t \in T_O$, $t^\bullet$ is restricted by a logical output expression $f_O(t)$;
3. $F = (P \times T)(T \times P)$ is a finite set of arcs;
4. I is a mapping from logic input transitions to logic input functions, and for $\forall t \in T_I$, $I(t) = f_I(t)$;
5. O is a mapping from logic output transitions to logic output functions, and for $\forall t \in T_O$, $O(t) = f_O(t)$;
6. $M : P \to \{0, 1, 2, \dots\}$ is the marking function; and
7. The transition firing rules are as follows:

   (a) For $\forall t \in T_D$, the transition firing rules are the same as in Petri net;
   (b) For $\forall t \in T_I$, if $f_I(t)|M = {}_\bullet T_\bullet$, then a logic input transition can be fired, denoted as $M[t > M'$, and for $\forall p \in {}^\bullet t$, $M'(p) = 0$; and for $\forall p \notin {}^\bullet t \cup t^\bullet$, $M'(p) = M(p)$; and for $\forall p \in t^\bullet$, $M'(p) = 1$; and
   (c) For $\forall t \in T_O$, if $\forall p \in {}^\bullet t$, $M(p) = 1$, then a logic output transition can be fired, and for $\forall p \in {}^\bullet t$, $M'(p) = 0$; for $\forall p \in t^\bullet$ must satisfies $f_O(t)|M = {}_\bullet T_\bullet$, and for $\forall p \notin {}^\bullet t \cup t^\bullet : M'(p) = M(p)$.

For example, a logic Petri net is shown in Figure 1. t_1 is an input transition. $I(t_1) = p_1 \vee p_2$ is the logic input function of t_1. From firing t_1, there are three situations:

1. p_1 contains a token, or
2. p_2 contains a token, or
3. both p_1 and p_2 contain a token.

t_3 is an output transition. $O(t_3) = (p_6 \otimes p_7) \wedge p_8$ is the logic output function of t_3. There are two situations when t_3 is fired: each of p_6 and p_8 contain a token, or each of p_7 and p_8 contain a token.

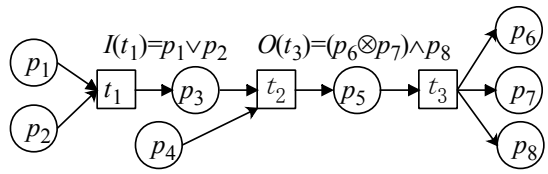


Figure 1. A logic Petri net model LPN_1

Definition 9 (Process tree [18])**.** Let A be a set of actives. $\oplus$ is a given operator set, and τ is an invisible transition, where

1. $a \in A \cup \{\tau\}$ is a process tree;
2. If $PT_1, \ldots, PT_n$ $(n > 0)$ are process trees, then $\oplus(PT_1, \ldots, PT_n)$ is also a process tree; and
3. There are 4 operators:
   - $\times$ stands for a selection relation, and only one sub-tree can occur among $PT_1, \ldots, PT_n$;
   - $\rightarrow$ represents a sequence relation, and the corresponding sub-trees will occur in sequence;
   - $\wedge$ denotes a parallel relation, and the corresponding sub-trees will occur simultaneously; and
   - $\circlearrowleft$ represents a loop structure, and PT_1 denotes a circulatory body, and PT_2, $\ldots$, PT_n $(n \geq 2)$ denotes a loop path.

3 MODEL REPAIRING OF SELECTION STRUCTURES

When business processes or actual working environment changes, event logs generated by actual processes cannot be completely replayed by its original model. Therefore, deviations between an actual event log and its original model should be

identified, and the original process model can be repaired accordingly. In this section, a repairing method is proposed based on a token replay method for models with selection structures. It can find deviations between the original model and the generated logs. Therefore, an original model can be repaired.

3.1 Model Repairing with the Equal Number of Transitions and Log Activities

When a model is repaired, it is necessary to determine the location of deviations between a model and logs. By replaying event log L in a model, the missing and remaining-tokens can be calculated [19]. The position of deviations can be determined. In token replay, tokens will dynamically change from the initial place to the final place when a trace is completely consistent with a model. If there are deviations between a trace and a model, tokens cannot reach the end place according to the missing-tokens, and the fitness of the rest trace cannot be analyzed. Thus an enhanced replaying algorithm is given next.

Algorithm 1 Enhanced Replaying Algorithm

Input: A Workflow net $WFN = (P, T; F, i, o)$ and an event log $L \in \beta(\sigma^*)$;
Output: M.

1: **for** each $p \in P$ **do**
2: **if** $p = i$ **then**
3: $M(i) = 1$;
4: **else**
5: $M(p) = 0$;
6: **end if**
7: **end for**
8: **for** $(j = 1; t_j \in \&(\sigma); j++)$ **do**
9: **if** $t_j \in T$ and $p \in {}^\bullet t_j - t_j^\bullet$ **then**
10: $M(p) \leftarrow M(p) - 1$;
11: **end if**
12: **if** $t_j \in T$ and $p \in t_j^\bullet - {}^\bullet t_j$ **then**
13: $M(p) \leftarrow M(p) + 1$;
14: **end if**
15: **end for**
16: $M(o) \leftarrow M(o) - 1$;
17: **return** M.

In Algorithm 1, all places are initialized in Steps 1–7. There is a token in the initial place, and there is no token in other places. The transitions in traces are replayed in Steps 8–15. If $t_j \in T$ and $p \in {}^\bullet t_j - t_j^\bullet$, then $M(p) = M(p) - 1$. If $t_j \in T$ and $p \in t_j^\bullet - {}^\bullet t_j$, then $M(p) = M(p) + 1$. In Step 16, a token is consumed from the output place, and a final marking is obtained in Step 17. The computational complexity of Algorithm 1 is $O(n)$.

Example 1. A workflow net WFN_1 is shown in Figure 2 where $\sigma_1 = \langle t_1, t_2, t_4, t_5, t_6, t_7 \rangle$ is replayed. Table 1 shows the change of tokens based on Algorithm 1.

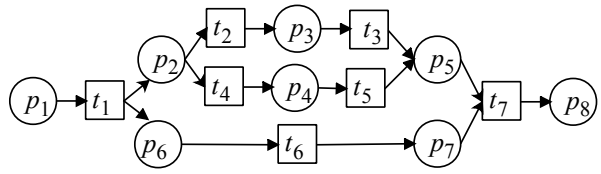


Figure 2. Workflow net model WFN_1

$M(p_8) = 0$ at the end of a replay. After executing Algorithm 1, $M(p_2) = -1$ represents that a token is missing, and $M(p_3) = 1$ represents that a token is remaining. For missing-token places, there should be an arc connecting a place such that tokens can be generated. For remaining-token places, there should be an arc connecting a transition to consume a token. The locations of missing and remaining-token places represent an end position and a start position of an adding arc, respectively. When a model is repaired, another side of the connecting arc needs to be calculated. For example, ${}^\bullet p_3$ should connect t_4 in Figure 2. To repair WFN_1, the arc from p_3 to t_4 should be added based on $\sigma_1 = \langle t_1, t_2, t_4, t_5, t_6, t_7 \rangle$. Another side of an added arc can be determined based on the start position p_2 or the end position p_5 in selection structures. From a process tree and selection relation pairs, start and end pairs are defined with selection structures.

Transisions	$M(p_1)$	$M(p_2)$	$M(p_3)$	$M(p_4)$	$M(p_5)$	$M(p_6)$	$M(p_7)$	$M(p_8)$
Start	1	0	0	0	0	0	0	0
t_1	0	1	0	0	0	1	0	0
t_2	0	0	1	0	0	1	0	0
t_4	0	−1	1	1	0	1	0	0
t_5	0	−1	1	0	1	1	0	0
t_6	0	−1	1	0	1	0	1	0
t_7	0	−1	1	0	0	0	0	1
End	0	−1	1	0	0	0	0	0

Table 1. Change of token in WFN_1

Definition 10 (Selection relation)**.** Let PT be a process tree of $WFN = (P, T; F, M, i, o)$. $n =$ "$\times$" is a node of PT, and it represents a selection structure. $C_{RP} = (t_1, t_2)$ is called a selection relation where $t_1 = n_l$ and $t_2 = n_r$ where n_l and n_r represent the leftmost and rightmost subtree of a selection structure, respectively.

S_{CRP} represents a set of selection relations, i.e., $S_{CRP} = \{(t_1, t_2) \mid t_1 = n_l, t_2 = n_r, \forall n = \times\}$. For example, $S_{CRP} = \{(t_2, t_5)\}$ in Figure 3.

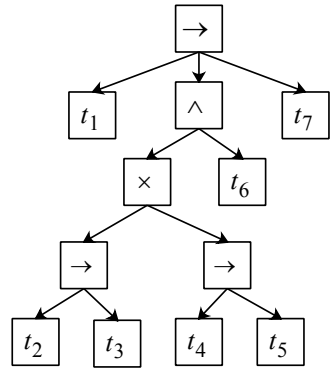


Figure 3. The process tree PT_1 of WFN_1

Definition 11 (Selection structure)**.** PT is a process tree of $WFN = (P, T; F, M, i, o)$. $n =$ "×" is a node of PT, and it represents a selection structure. $C_{RP} \in S_{CRP}$ represents a selection relation pair. $C_{SEP} = (p_1, p_2)$ represents the start and end pair of a selection structure, and $p_1 = {}^\bullet(\pi_1(C_{RP}))$, $p_2 = (\pi_2(C_{RP}))^\bullet$. S_{CSEP} represents a start and end pair set of selection structures. $S_{CSEP} = \{(p_1, p_2) \mid p_1 = {}^\bullet(\pi_1(C_{RP})), p_2 = (\pi_2(C_{RP}))^\bullet, \forall C_{RP} \in S_{CRP}\}$.

For example, $C_{RP} = (t_2, t_5)$, and ${}^\bullet t_2 = \{p_2\}$, $t_5^\bullet = \{p_5\}$, in Figure 3. Therefore, $C_{SEP} = (p_2, p_5)$, and $S_{CSEP} = \{(p_2, p_5)\}$. Thus, an algorithm of calculating S_{CSEP} is given as follows.

Algorithm 2 gives a calculating method of start and end pair sets with selection structures. S_{CRP} and S_{CSEP} are initialized in Step 1. Steps 2–15 find out the leftmost and rightmost sub-tree pairs of all selection structures in WFN, and they are saved in S_{CRP}. Steps 16–18 calculate the start and end pairs of selection structures, according to S_{CRP}, and store them in S_{CSEP}. Step 19 returns S_{CSEP}.

Definition 12 (Mapping function)**.** Given log activity $a \in \&(\sigma)$ and a transition t corresponding to it, $Map(a, t)$ is the mapping function from a log activity to a model transition for trace σ in log L.

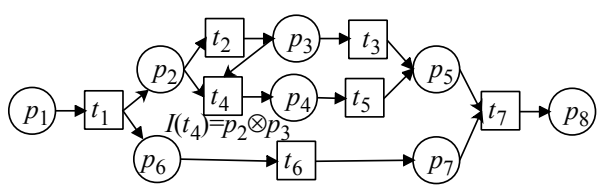


Figure 4. A repaired model of WFN_1 with L_1

Algorithm 3 gives a model repairing method. Step 1 calls Algorithm 1 to replay the logs. Step 2 calls Algorithm 2 to calculate the start and end pair set of

Algorithm 2 S_{CSEP} Calculation

Input: A Workflow net WFN, the non-leaf node of process tree PT denoted by n;
Output: A start and end pair set of selection structures, S_{CSEP}.

1: $S_{CRP} \leftarrow \phi$, $S_{CSEP} \leftarrow \phi$;
2: **for** each $n \in PT$ **do**
3: **if** $n! = \phi$ and $n \in \oplus$ **then**
4: **if** $n =$ "$\times$" **then**
5: $S_{CRP} \leftarrow S_{CRP} \cup \{(n_l, n_r)\}$;
6: **else**
7: **for** all the sub nodes $SN \in n$ **do**
8: $n \leftarrow SN$;
9: Skip to Setp 3;
10: **end for**
11: **end if**
12: **else**
13: break;
14: **end if**
15: **end for**
16: **for** $C_{RP} \in S_{CRP}$ **do**
17: $S_{CSEP} \leftarrow S_{CSEP} \cup \{({}^{\bullet}(\pi_1(C_{RP})), (\pi_2(C_{RP}))^{\bullet})\}$;
18: **end for**
19: **return** S_{CSEP}.

S_{CSEP}. Steps 3–13 calculate a pre-set for remaining-places, remaining-places, the number of remaining-places, a post-set for missing-places, missing-places, and the number of missing-places. Their results are stored in $T_s[\ ]$, $P_s[\ ]$, m, $T_q[\ ]$, $P_q[\ ]$, and n, respectively. All remaining-places and missing-places are judged whether they belong to a start or end place of selection structures in Steps 14–29. If the answer is no, an arc should be added from a remaining-place to $T_q[i]$, and a logic input expression $I(T_q[i]) \leftarrow P_s[j] \otimes P_q[i]$ should be added, for the remaining-places. Moreover, an arc should be added from $T_s[j]$ to $P_q[i]$, and a logic output expression $O(T_s[j]) \leftarrow P_s[j] \otimes P_q[i]$ should be added too, for the missing-places. Steps 30 and 31 mean P' and T' are the same as P and T, respectively. Finally, Algorithm 3 returns a repaired model LPN in Step 32. The computational complexity of Algorithm 3 is $O(n^3)$.

Example 2. Algorithm 3 is used to repair WFN_1 in Figure 2 where $L_1 = \{\langle t_1, t_2, t_4, t_5, t_6, t_7\rangle\}$. The repaired result is shown in Figure 4.

Example 3. Algorithm 3 is used to repair WFN_1 in Figure 2 where $L_2 = \{\langle t_1, t_2, t_3, t_5, t_6, t_7\rangle\}$. The repaired result is shown in Figure 5.

Algorithm 3 Model Repairing Algorithm

Input: A workflow net $WFN = (P, T; F, M, i, o)$ and an event log $L \in \beta(\sigma^*)$;
Output: A repaired logic Petri net, denoted by $LPN = (P, T; F, I, O, M)$.

Call Algorithm 1 to replay event logs;
Call Algorithm 2 to calculate the set of a start and end pair S_{CSEP};
for $(i = 1;\ i < |T|;\ i++)$ **do**
    $m = n = 0$;
    if $M(p_i) > 0$ **then**
        $T_s[m] \leftarrow {}^\bullet p_i$;
        $P_s[m] \leftarrow p_i$;
        $m++$;
    end if
    if $M(p_i) < 0$ **then**
        $T_q[n] \leftarrow p_i^\bullet$;
        $P_q[n] \leftarrow p_i$;
        $n++$;
    end if
end for
for $(j = 1;\ j <= m;\ j++)$ **do**
    for $(i = 1;\ i <= n;\ i++)$ **do**
        for each $\sigma \in L(k = 1;\ k < |\sigma|;\ k++)$ **do**
            if $T_s[j] \in \sigma_k$ and $T_q[i] \in \sigma_{k+1}$ **then**
                if $P_s[j] \notin \pi_1(S_{CSEP})$ and $P_s[j] \notin \pi_2(S_{CSEP})$ **then**
                    $F' \leftarrow F' \cup P_s[j] \rightarrow T_q[i]$;
                    $I(T_q[i]) \leftarrow P_s[j] \otimes P_q[i]$;
                end if
                if $P_q[i] \notin \pi_1(S_{CSEP})$ and $P_q[i] \notin \pi_2(S_{CSEP})$ **then**
                    $F' \leftarrow F' \cup T_s[j] \rightarrow P_q[i]$;
                    $O(T_s[j]) \leftarrow P_s[j] \otimes P_q[i]$;
                end if
            end if
        end for
    end for
end for
$P' \leftarrow P$;
$T' \leftarrow T$;
return $LPN = (P', T'; F', I, O, M)$.

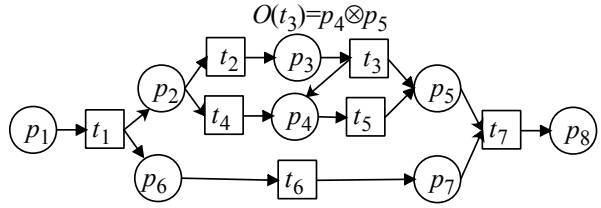


Figure 5. A repaired model of WFN_1 with L_2

3.2 Model Repairing with Different Number of Transitions and Log Activities

It is supposed that model activities and log activities are the same in Algorithm 3. However, log activities generated by an actual process are generally more than model activities. When this situation happened, Algorithm 3 cannot complete a model repairing. To take with this problem, newly added log activities need to be calculated first. A sub-model can be mined according to the newly added log activities and the inductive algorithms. Finally, a sub-model can be inserted into the original model. The algorithm for calculating the newly added log activities is given as follows.

Algorithm 4 New Log Activities

Input: A workflow net $WFN = (P, T; F, M, i, o)$ and an event log $L \in \beta(\sigma^*)$;
Output: The set of newly added log activities, denoted by $NewAct$.

1: $NewAct \leftarrow \phi$, $T_M \leftarrow \phi$;
2: **for** $(i = 1; i <= |T|; i++)$ **do**
3: $\quad T_M \leftarrow T_M \cup \{t_i\}$;
4: **end for**
5: **for** each $\sigma \in L$ **do**
6: $\quad$ **for** $(j = 1, a_j \in \sigma; j <= |\sigma|; j++)$ **do**
7: $\quad\quad Map(a_j, t_k)$;
8: $\quad\quad$ **if** $t_k \notin T_M$ **then**
9: $\quad\quad\quad NewAct \leftarrow NewAct \cup \{a_j\}$;
10: $\quad\quad$ **end if**
11: $\quad$ **end for**
12: **end for**
13: **return** $NewAct$.

New log activities are calculated in Algorithm 4. Step 1 initializes $NewAct$ and T_M as an empty set. From Steps 2–4, all model activities are added in T_M. Log activities are mapped to model activities in Steps 5–13, and it is judged whether the model activities belong to T_M. If the answer is no, then the log activities are added to $NewAct$. Step 14 returns $NewAct$ of new activities. The computational complexity of Algorithm 4 is $O(n^2)$.

WFN_2 is shown in Figure 6, and $L_3 = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5\} = \{\langle t_1, t_2, t_4, t_8\rangle, \langle t_1, t_9, t_5, t_6, t_{12}, t_8\rangle, \langle t_1, t_3, t_5, t_6, t_7, t_8\rangle, \langle t_1, t_9, t_{10}, t_{12}, t_8\rangle, \langle t_1, t_9, t_{11}, t_{12}, t_8\rangle\}$. Newly added log activities t_9, t_{10}, t_{11} and t_{12} can be calculated based on Algorithm 4.

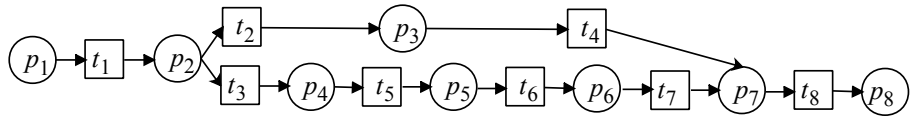


Figure 6. Workflow net model WFN_2

Algorithm 5 Order Relation from an Event Log

Input: An event log $L \in \beta(\sigma^*)$;
Output: An event log relation set R.

1: $R \leftarrow \phi$, $R' \leftarrow \phi$;
2: **for** each $\sigma \in L$ **do**
3: $R' \leftarrow R' \cup \{a_i >_L a_{i+1}\}$;
4: **if** $a \in \sigma$ and $b \notin \sigma$ or $b \in \sigma$ and $a \notin \sigma$ **then**
5: $R \leftarrow R \cup \{a \#_L b\}$;
6: **end if**
7: **if** $a >_L b$ and $b \not>_L a$ **then**
8: $R \leftarrow R \cup \{a \rightarrow_L b\}$;
9: **end if**
10: **if** $a >_L b$ and $b >_L a$ **then**
11: $R \leftarrow R \cup \{a||_L b\}$;
12: **end if**
13: **end for**
14: **return** R.

Algorithm 5, the order relationship of logs can be found. $a \rightarrow_L b$, and $a||_L b$, $a\#_L b$ represent a causality, parallel, choice relation between a and b, respectively. The computational complexity of this algorithm is $O(n)$.

A repairing method of including sub-models with selection structures is shown in Algorithm 6. Algorithm 4 is called to calculate a new log activities set *NewAct* in Step 1. Step 2 calculates the projection of *NewAct* in L to find sub log *SL*. *InductiveMiner*(*SL*) algorithm is called to mine a sub-model $WFN' = (P', T'; F', M', i', o')$ in Step 3. Step 4 calls Algorithm 5 to calculate the order relation of event logs. A sub-model is inserted into the original model in Steps 5–13. Step 14 calls Algorithm 3 to repair the model.

Example 4. Workflow net WFN_2 can be repaired based on Algorithm 6, and $L_3 = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5\} = \{\langle t_1, t_2, t_4, t_8\rangle, \langle t_1, t_9, t_5, t_6, t_{12}, t_8\rangle, \langle t_1, t_3, t_5, t_6, t_7, t_8\rangle, \langle t_1, t_9, t_{10}, t_{12}, t_8\rangle, \langle t_1, t_9, t_{11}, t_{12}, t_8\rangle\}$.

Algorithm 6 Model Repairing Method for Sub-Models with Choice Structures

Input: A workflow net $WFN = (P, T; F, M, i, o)$ and an event log $L \in \beta(\sigma^*)$;
Output: The repaired logic Petri net $LPN = (P', T'; F', I, O, M)$.

1: Call Algorithm 4 to calculate a new log activities set $NewAct$;
2: Calculating the projection of $NewAct$ in L to find its sub log SL;
3: Call $InductiveMiner(SL)$ algorithm to mine a sub-model $WFN' = (P', T'; F', i', o', M')$;
4: Call Algorithm 5 to calculate the order relationship of event logs;
5: **for** each $(t_i \rightarrow_L i'^{\bullet}) \in R$ **do**
6: $\quad F' = F' \cup (t_i^{\bullet} \rightarrow i'^{\bullet})$;
7: **end for**
8: **for** each $({}^{\bullet}o' \rightarrow_L t_i) \in R$ **do**
9: $\quad F' = F' \cup ({}^{\bullet}o' \rightarrow {}^{\bullet}t_i)$;
10: **end for**
11: $F' = F \cup F' - (i' \rightarrow i'^{\bullet}) - ({}^{\bullet}o' \rightarrow o')$;
12: $T = T \cup T'$;
13: $P = P \cup P' - i' - o'$;
14: Call Algorithm 3 to repair the model.

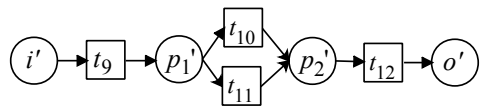


Figure 7. Sub Workflow net model WFN'

$NewAct = \{t_9, t_{10}, t_{11}, t_{12}\}$ is calculated first according to the Algorithm 4. Then the sub log $SL = \{\langle t_9, t_{12}\rangle, \langle t_9, t_{10}, t_{12}\rangle, \langle t_9, t_{11}, t_{12}\rangle\}$ is calculated based on $NewAct$ and L_2. A sub-model WFN' can be mined according to the inductive algorithm, and the result shows in Figure 7. The relationship between the new log activities t_9–t_{12} and the original model transition is calculated. Besides, the relationships are $t_1 \rightarrow t_9$, $t_2 \# t_9$, $t_3 \# t_9$, and $t_{12} \rightarrow t_8$. Finally, WFN' is inserted into the original

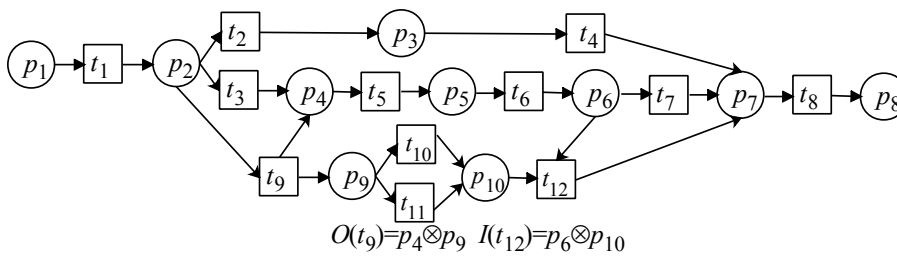


Figure 8. Repaired model of WFN_2

model and the model is repaired according to Algorithm 3. The repaired model is shown in Figure 8.

4 SIMULATION EXPERIMENTS

Simulation experiments are conducted in this section. The data is from an inspection department from a hospital in Qingdao, China, and event logs can be accessible at: `https://pan.baidu.com/s/1AG4TvrxF62uAP2QOowOSEQ`. The experimental results of the proposed method in this paper are compared and analyzed with those of Fahland's method [5] and Goldratt's method [11], where the former is implemented by the corresponding plug-ins in ProM 6.10 available at `http://www.promtools.org/`; and the latter is implemented in the DOS window and edited in ProM 6.10.

4.1 Model Repairing

The Petri net model in Figure 9 can be mined by α algorithm [7] according to event logs, and it shows the whole process of patients from outpatient appointments to treatment and departure. First, a patient makes an appointment at the triage station or by phone call. Then he (or she) needs to book and then gets a reservation number. Some patients may not make an appointment, but they need to register at first. After that, the patients need to wait for calling their number and are inquired by the doctor. Then the patients may need to do some examinations, i.e., common CT, PET-CT, chest enhanced scan, ESR, biochemical full set, blood gas analysis, and blood routine. Then, the doctor makes a diagnosis according to the results of examinations and decides whether the patients leave the hospital or be hospitalized.

The event logs L_1–L_3 are shown in Table 2 including the number and length of traces, and the number of events and activities.

Logs	Traces	Events	Activities	Length
L_1	105	980	24	$7 \sim 11$
L_2	210	1 960	24	$7 \sim 11$
L_3	314	2 938	24	$7 \sim 11$

Table 2. Event logs

There are some deviations between event logs and the process model in Figure 9 since there are some new occurring activities. For example, a patient can make an appointment by internet or WeChat. Besides, the doctor may choose other decisions after diagnosis, i.e., referral or conservative treatment.

The model in Figure 9 can be repaired by three methods. The repaired model of Fahland's method is shown in Figure 10. The repaired model of Goldratt's method is shown in Figure 11. The repaired model of our approach is shown in Figure 12. From Figures 10, 11 and 12 the repaired models of Fahland's and Goldratt's methods

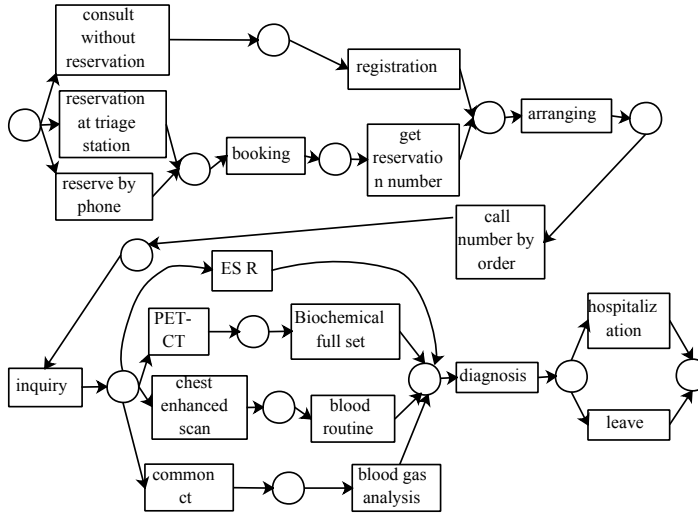


Figure 9. A Petri net of thoracic surgery

add some self-loops and invisible transitions to improve the fitness of models. Self-loops and invisible transitions can decrease the model precision, and increase the model complexity.

The following situation will occur according to actual occurrence logs in this paper. After the common CT, it is possible to check blood gas analysis, ESR, blood routine, biochemical full set, or entering the diagnosis directly. In this paper, suppose that the logic output function of t_{12} is $O(t_{12}) = p_9 \otimes p_{10} \otimes p_{11} \otimes p_{12}$. After checking chest enhanced scan, blood routine, biochemical full set, or directly entering the diagnosis can be done. Thus the logic output function of t_{13} is set as $O(t_{13}) = p_{10} \otimes p_{11} \otimes p_{12}$. After checking PET-CT, biochemical full set or directly entering the diagnosis can be done. The logic output function of t_{14} is set as $O(t_{14}) = p_{11} \otimes p_{12}$. Common CT can be done before checking ESR. Therefore, the logic input function of t_{16} is set as $I(t_{16}) = p_8 \otimes p_9$. After inquiry blood routine, biochemical full set, blood gas analysis or directly entering the diagnosis can be done. Therefore, the logic input function of t_{15}, t_{17}, t_{18}, t_{19} are set as $I(t_{15}) = p_8 \otimes p_9$, $I(t_{17}) = p_8 \otimes p_{10}$, $I(t_{18}) = p_8 \otimes p_{11}$, $I(t_{19}) = p_8 \otimes p_{12}$, respectively.

In the process model of Figure 12, the mapping relationship between transitions and activities is shown in Table 3.

4.2 Model Evaluation

The simplicity of the repaired process model can be analyzed by three repairing methods according to the principle of Occam's Razor [20]. The increased number of places, transitions, invisible transitions and arcs are compared in Table 4, after

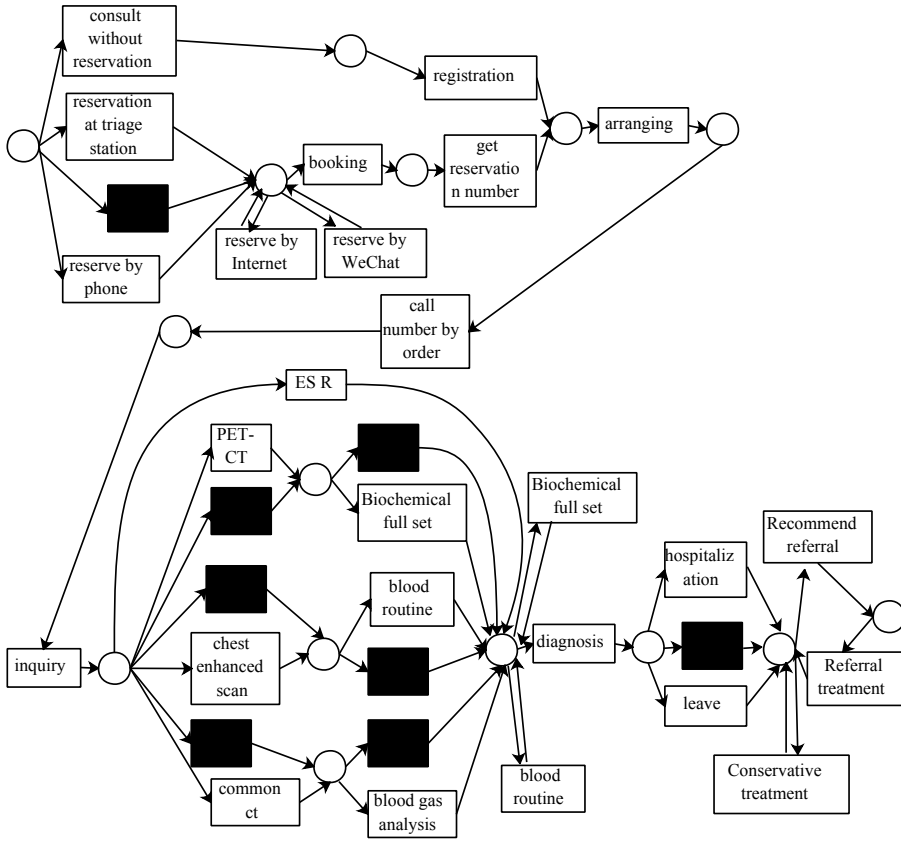


Figure 10. Fahland's method

Transitions	Activities	Transitions	Activities
t_1	reserve by phone	t_{13}	chest enhanced scan
t_2	reservation at triage station	t_{14}	PET-CT
t_3	reserve by Internet	t_{15}	blood gas analysis
t_4	reserve by WeChat	t_{16}	ESR
t_5	consult without reservation	t_{17}	blood routine
t_6	booking	t_{18}	biochemical full set
t_7	get reservation number	t_{19}	diagnosis
t_8	registration	t_{20}	leave
t_9	arranging	t_{21}	hospitalization
t_{10}	call number by order	t_{22}	conservative treatment
t_{11}	inquiry	t_{23}	recommend referral
t_{12}	common CT	t_{24}	referral treatment

Table 3. The mapping relationship of transitions and activities in Figure 9

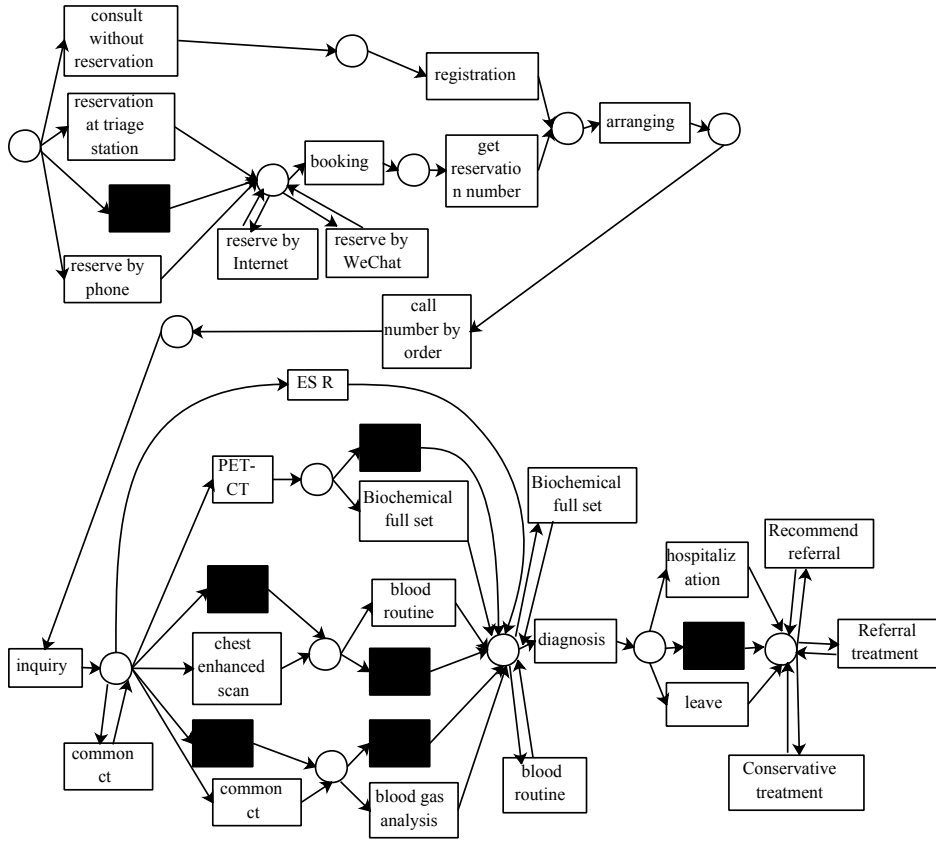


Figure 11. Goldratt's method

the model is repaired by three methods. From the analysis, the invisible transitions are not added in our method, while 8 and 7 invisible transitions are increased by Fahland's and Goldratt's methods, respectively. From the increasing number of arcs, 21 arcs are added by the repairing method of this paper, while 30 arcs are increased by Fahland's and Goldratt's methods, respectively. Thus, the simplicity of our approach is lower than the two repairing methods.

Models	Places	Transitions	Invisible transitions	Arcs
Our approach	1	5	0	21
Fahland's method	1	5	8	30
Goldratt's method	0	5	7	30

Table 4. Comparison of simplicity in three repairing models

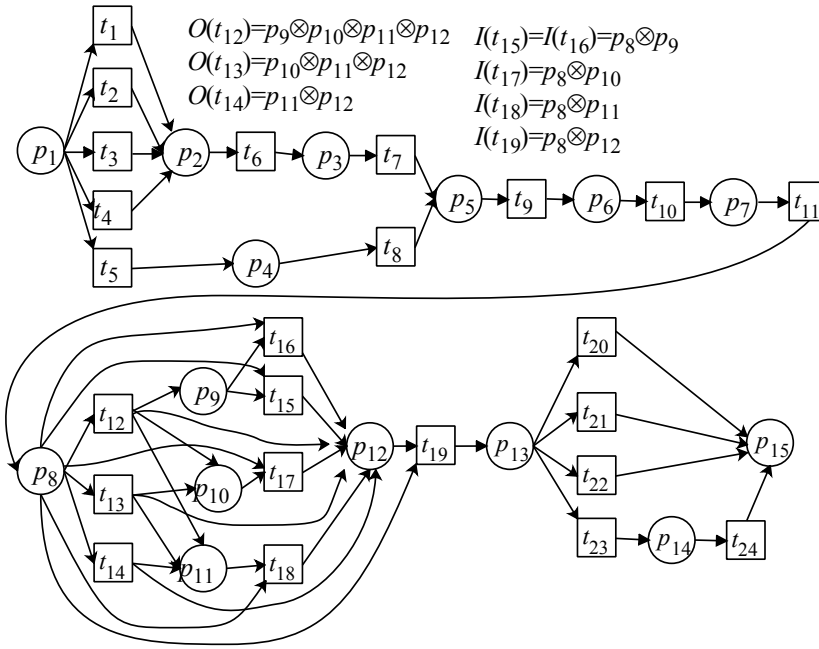


Figure 12. The repaired model by our approach

The fitness [21] is another important metric of conformance checking of process models. The fitness of different models is analyzed based on the number of traces in Figure 13. The fitness of Fahland's and Goldratt's methods are calculated according to the tool of ProM 6.10 plug-in "Replay a Log on Petri Net for Performance Analysis". The fitness of the logical Petri net model proposed in this paper is calculated manually according to reference [21]. From Figure 13, the fitness of every method is more than 0.9. besides, Fahland's method is slightly lower than our approach and Goldratt's method. The fitness of our approach is 1.

The precision of the three repairing methods is shown in Figure 14. The precision of Fahland's and Goldratt's methods are calculated by plug-in "Check Precision based on Align-ETCformance" in ProM 6.10. The precision of the logical Petri net model proposed in this paper is calculated manually according to reference [21]. When the given amount of available repair resources is small, the repaired model by Goldratt's method cannot display all the new log activities. If the value of R is set to 16 in Goldratt's method, all log activities can appear in the repaired model. Therefore, the precision of Goldratt's method is relatively low, around 0.67. The precision of Fahland's method is about 0.73. The precision of our approach is about 0.85. Obviously, our approach has higher precision than others.

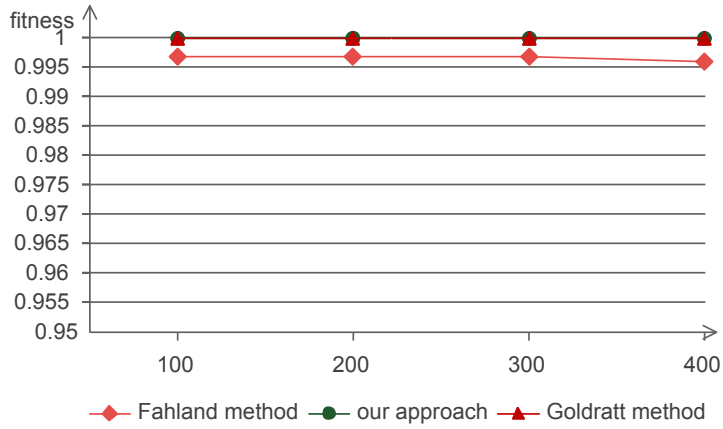


Figure 13. The fitness between different models

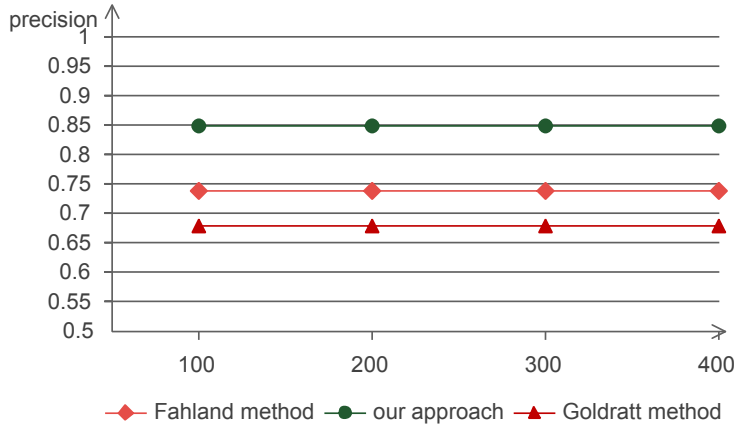


Figure 14. The precision between different models

5 CONCLUSIONS

Aiming at the low precision of Fahland's and other repairing methods, a repairing method of token replay is proposed based on logical Petri nets in this paper. The deviation locations are firstly determined by token replay. A sub-model of newly added log activities is mined by the inductive algorithm. Then an insertion location of the sub-model is determined based on the relationship between the input and output places of the sub-model and event logs. Finally, an original model can be repaired by given algorithms via logical Petri nets. The repairing methods of process models with selection structures are discussed in this paper. Therefore, the

repairing methods of process models with parallel structures or loop structures will be analyzed in our future research.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61903229 and Grant No. 61973180, and in part by the China Electronics Technology Group Corporation (CETC).

REFERENCES

[1] van der Aalst, W. M. P.—Stahl, C.: Modeling Business Processes: A Petri Net Oriented Approach. The MIT Press, Cambridge, USA, 2011, doi: 10.7551/mitpress/8811.001.0001.

[2] Conforti, R.—Dumas, M.—García-Bañuelos, L.—La Rosa, M.: BPMN Miner: Automated Discovery of BPMN Process Models with Hierarchical Structure. Information Systems, Vol. 56, 2016, pp. 284–303, doi: 10.1016/j.is.2015.07.004.

[3] Wang, L.—Du, Y. Y.—Liu, W.: Aligning Observed and Modelled Behaviour Based on Workflow Decomposition. Enterprise Information Systems, Vol. 11, 2017, No. 8, pp. 1207–1227, doi: 10.1080/17517575.2016.1193633.

[4] Wang, Y. Y.—Du, Y. Y.: Comformance Checking Based on Extended Footprint Matrix. Journal of Shandong University of Science and Technology (Natural Science), Vol. 37, 2018, No. 2, pp. 9–15.

[5] Fahland, D.—Van der Aalst, W. M. P.: Model Repair – Aligning Process Models to Reality. Information Systems, Vol. 47, 2015, pp. 220–243, doi: 10.1016/j.is.2013.12.007.

[6] van der Aalst, W. M. P.—Weijters, T.—Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering, Vol. 16, 2004, No. 9, pp. 1128–1142, doi: 10.1109/tkde.2004.47.

[7] Wen, L. J.—van der Aalst, W. M. P.—Wang, J. M.—Sun, J. G.: Mining Process Models with Non-Free-Choice Constructs. Data Mining and Knowledge Discovery, Vol. 15, 2007, No. 2, pp. 145–180, doi: 10.1007/s10618-007-0065-y.

[8] Wen, L. J.—Wang, J. M.—Sun, J. G.: Mining Invisible Tasks from Event Logs. In: Dong, G., Lin, X., Wang, W., Yang, Y., Yu, J. X. (Eds.): Advances in Data and Web Management (APWeb 2007, WAIM 2007). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 4505, 2007, pp. 358–365, doi: 10.1007/978-3-540-72524-4_38.

[9] van der Aalst, W. M. P—de Medeiros, A. K. A.—Weijters, A. J. M. M.: Genetic Process Mining. In: Ciardo, G., Darondeau, P. (Eds.): Application and Theory of Petri Nets 2005 (ICATPN 2005). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 3536, 2005, pp. 48–69, doi: 10.1007/11494744_5.

[10] de Medeiros, A. K. A.—Weijters, A. J. M. M.—van der Aalst, W. M. P.: Genetic Process Mining: An Experimental Evaluation. Data Mining and Knowledge Discovery, Vol. 14, 2007, No. 2, pp. 245–304, doi: 10.1007/s10618-006-0061-7.

[11] Polyvyanyy, A.—van der Aalst, W. M. P.—Ter Hofstede, A. H. M.—Wynn, M. T.: Impact-Driven Process Model Repair. ACM Transactions on Software Engineering and Methodology, Vol. 25, 2017, No. 4, Art. No. 28, 60 pp., doi: 10.1145/2980764.

[12] Du, Y. Y.—Qi, L.—Zhou, M. C.: A Vector Matching Method for Analyzing Logic Petri Nets. Enterprise Information Systems, Vol. 5, 2011, No. 4, pp. 449–468, doi: 10.1080/17517575.2010.541943.

[13] Rozinat, A.—van der Aalst, W. M. P.: Conformance Checking of Processes Based on Monitoring Real Behavior. Information Systems, Vol. 33, 2008, No. 1, pp. 64–95, doi: 10.1016/j.is.2007.07.001.

[14] Teng, Y. X.—Qi, L.—Du, Y. Y.: A Logic Petri Net-Based Repair Method of Process Models with Incomplete Choice and Concurrent Structures. Computing and Informatics, Vol. 39, 2020, No. 1-2, pp. 264–297, doi: 10.31577/cai_2020_1-2_264.

[15] Wang, Z.—Du, Y. Y.—Qi, L.: Extended Colored Logic Petri Net and Its Reachability Analysis. Journal of Shandong University of Science and Technology (Natural Science), Vol. 39, 2020, No. 1, pp. 84–98.

[16] Qi, H. D.—Du, Y. Y.—Liu, W.: Process Model Repairing Method Based on Reachable Markings. Journal of Shandong University of Science and Technology (Natural Science), 2017, pp. 118–124.

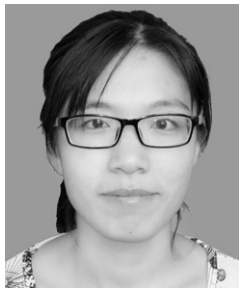
[17] Wen, L.—Wang, J.—van der Aalst, W. M. P.—Huang, B.—Sun, J.: Mining Process Models with Prime Invisible Tasks. Data and Knowledge Engineering, Vol. 69, 2010, No. 10, pp. 999–1021, doi: 10.1016/j.datak.2010.06.001.

[18] Buijs, J. C. A. M.—van Dongen, B. F.—van der Aalst, W. M. P.: A Genetic Algorithm for Discovering Process Trees. Proceedings of the 2012 IEEE Congress on Evolutionary Computation (CEC), 2012, pp. 1–8, doi: 10.1109/cec.2012.6256458.

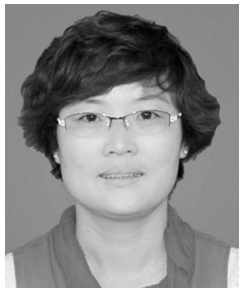
[19] Adriansyah, A.—van Dongen, B. F.—van der Aalst, W. M. P.: Conformance Checking Using Cost-Based Fitness Analysis. Proceedings of the 2011 IEEE 15th International Enterprise Distributed Object Computing Conference (EDOC), 2011, pp. 55–64, doi: 10.1109/edoc.2011.12.

[20] Witten, I.—Frank, E.: Data Minning: Practical Machine Learning Tools and Techniques. Second Edition. Morgan Kaufmann, 2005.

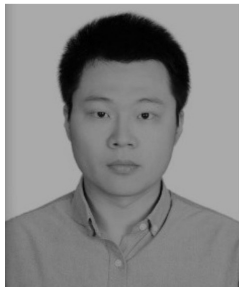
[21] Adriansyah, A.: Aligning Observed and Modeled Behavior. Ph.D. Thesis, Technische Universiteit Eindhoven, 2014, pp. 139–149, doi: 10.6100/IR770080.



Erjing Bai received her B.Sc. degree from the Hebei Normal University, Hebei, China, in 2000, her M.Sc. degree from the Qingdao University of Science and Technology, Qingdao, China, in 2014. She is currently Associate Professor at the College of Qingdao Huanghai University, Qingdao, China. Her current research interests are process mining, Petri nets and workflow.



Na Su received her B.Sc. degree from the Liaocheng Normal University, Shangdong, China, in 2000, her M.Sc. degree from the Qingdao University of Science and Technology, Qingdao, China, in 2015. She is currently Associate Professor at the College of Qingdao Huanghai University, Qingdao, China. Her current research interests are process mining, Petri nets and workflow.



Yu Liang received his Bachelor degree in computer science and technology from the Shandong Agricultural University, China in 2015. He received his Master degree in software engineering from the Shandong University of Science and Technology, China in 2019. He is currently pursuing Doctorate in the Department of Computer Science and Technology, Tongji University, Shanghai. His research interests include lightweight deep neural networks (DNNs), interpretation methods for DNNs and graph DNNs.

Liang Qi received his B.Sc. degree in information and computing science and his M.Sc. degree in computer software and theory from the Shandong University of Science and Technology, Qingdao, China, in 2009 and 2012, respectively, and his Ph.D. degree in computer software and theory from the Tongji University, Shanghai, China in 2017. From 2015 to 2017, he was Visiting Student with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA. He is currently Associate Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China. He has over 80 papers in journals and conference proceedings, including the IEEE Transactions on Intelligent Transportation Systems, the IEEE/CAA Journal of Automatica Sinica, the IEEE Transactions on System, Man and Cybernetics: Systems, the IEEE Transactions on Computational Social Systems, the IEEE Transactions on Automation Science and Engineering, the IEEE Transactions on Cybernetics, the IEEE Transactions on Network Science and Engineering, IEEE Transactions on Image Processing, and IEEE Signal Processing Letters. He received the Best Student Paper Award-Finalist in the 15$^{\text{th}}$ IEEE International Conference on Networking, Sensing and Control (ICNSC 2018). His current research interests include Petri nets, optimization algorithms, machine learning, and intelligent transportation systems.

Yuyue Du received the B.Sc. degree from the Shandong University, Jinan, China, in 1982, his M.Sc. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1991, and his Ph.D. degree in computer application from the Tongji University, Shanghai, China, in 2003. He is currently Professor at the College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, China. He has taken in over 10 projects supported by the National Nature Science Foundation, the National Key Basic Research Developing Program, and other important and key projects at provincial levels. He has published over 200 papers in domestic and international academic publications. His research interests are in formal engineering, Petri nets, real-time systems, process mining, and workflows.