Human Factor Aspects of Situation Awareness in Autonomous Cars – An Overview of Psychological Approaches

Gábor Kovács

Széchenyi István University, Deparment of Criminal Sciences, Egyetem tér 1,
9026 Győr, Hungary
gkovacs@sze.hu

Ágnes Hőgye-Nagy, Győző Kurucz

University of Debrecen, Institute of Psychology, Department of Social and Work Psychology, Egyetem tér 1, 4032 Debrecen, Hungary,
hogye-nagy.agnes@arts.unideb.hu; kurucz.gyozo@arts.unideb.hu

Abstract: The aim of the article is to give an overview of human factor research in psychology applicable to autonomous driving. The study is centered around situation awareness, a widely used concept in human factor research regarding the operation of automated and semi-automated systems (and communication between autonomous vehicles and humans). A proposal is put forward for structuring situation awareness requirements for autonomous driving, which could be a starting point for defining such requirements, and may foster a discussion on the issues associated with the human factor in relation to autonomous driving. Two models of human error (the SHELL model and the Swiss cheese model) are also introduced, one of which represents an integrated approach of error in situations that involve humans working with complex machinery or instruments, while the other represents a more superficial viewpoint on the multicausal nature of errors. The present overview can provide an appropriate basis for a discussion about the role of the driver in autonomous vehicles, and the place of human factor research in the emerging field of self-driving technology.

Keywords: autonomous car, self-driving car, situation awareness, human factor, human error

1 Introduction

Autonomous driving systems are currently one of the main research and development fields. The number of publications in this field have shown a significant increase in the last few years.

A fundamental issue, from a human factor perspective, is how to design automation so that drivers fully understand the capabilities and limitations of the vehicle, and maintain situation awareness of what the vehicle is doing and when manual intervention is needed – especially for first-generation vehicles that require drivers to resume manual control of automated functions when the vehicle is incapable of controlling itself.

However, the role of the driver, the human factor, is still underrepresented in these studies. The purpose of this paper is to document some of the human factors and challenges associated with the transition from manually driven to self-driving vehicles and to outline possibilities.

A key issue with highly automated driving (HAD) at this stage of its development is that it is not yet fully reliable and safe [1]. Therefore, in situations in which HAD fails or is limited (e.g., sensory degradation in poor weather conditions; the inability of on-board computer algorithms to make a safe decision), the driver will be expected to take control of the vehicle and resume manual driving. For this transition of control to occur safely, it is imperative that the driver fully understands the capabilities and limitations of HAD and maintains full awareness of what the vehicle is doing and when intervention might be needed [2]. In this paper, we document some of the human factor challenges associated with the transition from manually driven to self-driving vehicles.

Psychology, especially traffic psychology is a field that aims to investigate road user behavior, and the psychological aspects, factors, and processes that underlie these behaviors [3], and it should be an essential contributor to the discussion of the issues of autonomous vehicles. Still, psychological approaches are restricted mainly on questions of attitudes to and acceptance of autonomous vehicles. The nature of human cognitive processes, motivations, traits, emotions, moods, and habits all have critical effects on the driver's behavior, perception, and processing of information, which is widely investigated by psychological research. Thus, ignorance of these human characteristics might lead to potential problems and errors in the design and engineering of autonomous vehicles.

The aim of this paper is twofold. On the one hand, we would like to give an overview of the most relevant questions and results on situation awareness in the field of human factor research (see Table 1). We decided to use this expression, which is one among other often used terms in this field (e.g. human-machine interaction), since it emphasizes the place, source of problems we would like to discuss. Situation awareness has been chosen as the central focus since its role is widely accepted in the human factors literature and it is commonly applied in the research on human factors and on autonomous/intelligent systems. Secondly, by highlighting relevant issues and findings, this study would like to show how traffic psychology can contribute to other fields of science like engineering, information technology, and ergonomics. Hopefully, this overview will raise even more questions that need to be answered.

Table 1

Overview of problems discussed in this paper

source of problem		**input**	**process**	**output**
situational awareness	attention	• *locus of attention* • *anticipation and expectations* • *intentional focus of attention* • *divided attention* • *limitation of capacity* • *capacity of working memory*		
	vigilance	• *mental load* • *tiredness* • *boredom*		*reaction time*
	engagement	• *disengagement* • *distracted driving*		
	communication (between autonomous vehicle and human)	*misinterpretation*		• *miscommunication* • *effectiveness of communication*
	information processing		• *comprehension problems* • *relevance of information*	
human error		*typical characteristics of human "operation" and errors (concerning human-machine interaction and communication)*		

2 The Human Factor

The role of the human driver is extensively discussed in "traditional" driving systems. In the case of autonomous cars, the more developed the vehicle control system, the more critical the role of the human factor, – or more precisely, the change in the role of the human factor [4]. It is without question, that in the not-too-distant future autonomous vehicles will reach, or at least will approach the level of total/full automation (level 5, defined by SAE International [5]). Misuse of equipment can cause problems even in partially automated systems [6], hence it is crucial to know if drivers are able to properly use these modern, highly automated systems. Automated systems may not be perfect, they can make wrong decisions and err in some situations. As humans are the only flexible, and adaptive part of the system, only they can prevent these errors by modifying the processes of the system if necessary. Carrying this out may not be easy: the operator has to know the abilities, shortcomings, and limitations of their own and those of the system. Moreover, a high level of vigilance is needed to recognize the errors or malfunctions [7]. Thus, this implies the critical issue of cognitive and other skills in the context of driving. It must be clearly seen that humans and automated systems both can make errors, which – from a psychological perspective – emphasize the need for knowledge about the characteristics of human errors too.

The main subjects of human factor research are operators in nuclear power stations, oil refineries, chemical industrial plants, aviation, and autonomous vehicles. The main issues are trust, acceptance, proper use, and attention. Situation awareness seems to be an effective theoretical framework to discuss these because all of these issues can be built around this concept based on their effect on it.

3 Situation Awareness

Situation awareness is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" ([8], p. 36). It is the mental representation of the current state of the environment, more precisely that of its goal-relevant aspects. Okray and Lubnau [9] define it as a skill to become aware of the relevant, crucial characteristics of the actual situation, which is happening. It is proposed that situation awareness is the basis for decision making that involves complex and highly dynamic systems, and thus it is indispensable in fields like aviation, power plant operation, military tactics, or in more everyday activities like reading or driving in traffic.

Although there are several models of situation awareness (e.g. [10], [11]), the most widely cited and perhaps the most elaborate model in the human factors literature is that of Endsley's framework model [8]. The model contains the proposed structure of situation awareness as the representation of the current state, it's role in behavior involving a complex system, as well as endogenous and exogenous factors affecting it (Figure 1). As of the structure of situation awareness it contains three levels, and is hierarchical in nature. Level 1 is the perception of the elements, their status, and attributes in the environment. Level 2 goes beyond being simply aware of these elements, it involves the comprehension of their role in a situation, recognition of patterns, and identification of significant events. Level 3 is the projection of the current state in the near future based on the knowledge of the attributes of the elements and the understanding of the situation. As we can see the higher levels are based on the lower levels, which also implies that building up situation awareness is inherently time-consuming, and it must be maintained continuously to accurately represent the current situation at every level. For example, in a road traffic scenario when we approach an intersection with a certain speed, we may notice another vehicle from the right, also moving at a considerable speed having the index turned on, and that there aren't any stop signs or give way signs (level 1). From these, we conclude that the vehicle is about to take a left turn and that we have to give way according to the traffic rules (level 2). We also realize that with the current speed and course the two vehicles will crash (level 3). If this representation correctly describes the situation, it can be an apt base for a decision to avoid the collision by changing the speed or course of our vehicle. Also, an inaccurate representation on either level (level 1: overlooking the other vehicle, level 2: neglecting the traffic rules that apply to the situation, level 3: wrongly predicting the route of the vehicles) could result in an accident.

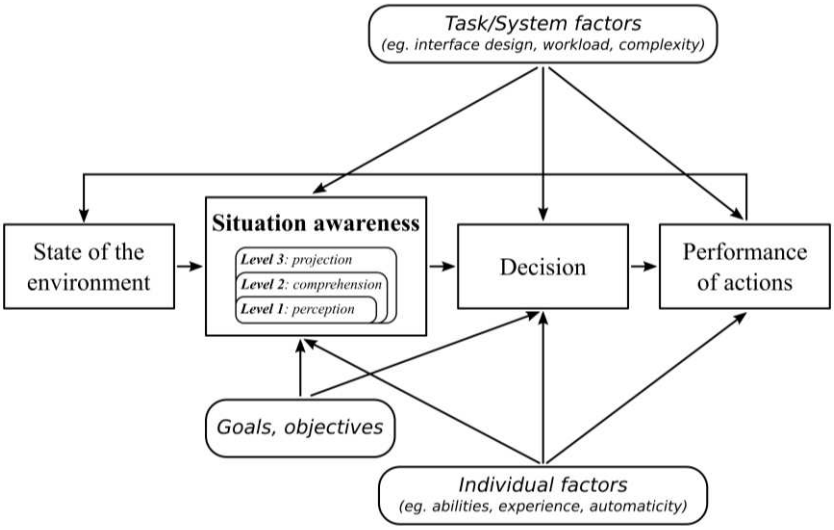


Figure 1
Endsley's framework model of situation awareness (based on [8])

It is important to note, however, that performance in a particular situation depends on situation awareness, but not identical to it. This is also emphasized by Endsley [8] or Adams et al. [12] among others. Errors can stem from an impaired assessment of the situation, from bad decision-making, or a poorly performed action. Thus, by separating the act of decision making and performance from situation awareness also makes it possible to analyze the causes of errors at a deeper level.

3.1 Factors Influencing Situation Awareness

Several studies have investigated the factors which may influence situation awareness. In this section an overview is given of the most frequently cited factors based on Okray and Lubnau [9], integrating it with the most important and relevant findings in traffic psychology.

3.1.1 Attention

Attention is known as an effective predictor of good achievement [13] and as one of the most important factors of safe driving. Some findings suggest that over 25-50% of accidents happen because of drivers' inattention. Attention deficit might be more typical in the case of novice drivers [14], who usually tend to locate their attention ahead in a narrower scope than experienced drivers. With the advancement of driving skills, intentional control of attention becomes more effective. Driving experience is strongly connected to age: younger drivers have less experience. This must be highlighted since attitudes toward autonomous cars are more positive among the youth [15]. Attentional anticipation also improves with the familiarity of the route [16], which may result in a more goal-oriented driving, a smoother driving performance, a decrease in unnecessary declarations, and an increase in travel speed. This improvement is more significant in the case of experienced drivers.

A central characteristic of attention is a limited capacity for information processing, which can be controlled intentionally [17]. Shinoda, Hayhie, and Shrivastava [18] also emphasize the relative importance of expectations in attentional processes. While expectations in some situations are beneficial for drivers, in other situations they are proven to impair attention and perception. A driver of an autonomous car can benefit from his expectations regarding the behavior of the car – they can prepare themselves to intervene. On the other hand, in the absence of expectations they may be surprised and take desperate and improper actions if the car is perceived to act inappropriately.

Since it is impossible to attend to every aspect of the actual situation, effective direction and allocation of attention are crucial for drivers. The direction of attention is determined by both exogenous and endogenous factors [19]. For instance, a change in the environment, like a sudden movement or appearance of a new stimulus usually draws attention to itself. Although this is an unintended process, attention can be directed intentionally too. During driving there is a complex interaction between exogenous and endogenous factors, and it is crucial for the driver to detect all the relevant cues. Unnecessary distractions could impair driving performance; thus, this should be an important aspect in information systems in autonomous vehicles. Insufficient information impairs driver's situation awareness, preventing them from forming an apt picture of the actual situation and therefore inhibits adequate decision making. On the other hand, too much information overloads the cognitive system, making it harder for the driver to attend to the relevant aspects of the situation, also impairing situation awareness and decision making.

Divided attention enables the individual to monitor several events, several aspects of the situation at the same time. According to the early selection theory of attention [20], filtering takes place at an early stage of perception, which means that attention sharing is not possible. We can talk only about quasi-shared attention: the perceiver focuses on only a single aspect of an event at a time, but the focus is (re)directed quickly and often to another aspect, so steps of information processing happen in sequence, shortly after each other. Other theories (e.g., Kahnemann [21]) suppose that there are situations when real, parallel information processing takes place, however, it is a function of attentional capacity and attentional processing if this can happen. Enough attentional capacity enables parallel processes, otherwise, sequential processing occurs.

When talking about attention, working memory must also be mentioned as an important factor in situation awareness. Working memory enables us to remember relevant aspects of a situation and make it available for the human cognitive system. Without this attention is unfocused, and it is not targeted [22]. In an (autonomous) driving task McCarty et al. [23] found that individuals with lower working memory capacity reacted slower. This approves he suggestion that working memory also plays a critical role in safety of autonomous cars.

3.1.2 Communication and Information Processing

Recently communication within the framework of situation awareness has mainly focused on human-human communication in relation to teamwork, not human machines. The findings showed that communication failures may end up in serious errors and even accidents [24]. Miscommunication is more frequent in high-workload situations, e.g., in air traffic control when the pilot is not a native English speaker. This type of communication failure is less important in our current focus. We would rather highlight how human drivers may misinterpret the information perceived in direct communication between autonomous vehicles and humans.

Drivers' performance varies during information processing. Shinar et al. [25] provide a cross-national study on traffic sign symbol comprehension. They concluded that a good sign design should follow stereotypes (for example red is connected to danger), differing shapes of signs help us to distinguish among different types of messages (prohibition, warning or guiding). Relevancy is also important: relevant signs require easier recognition. There are some characteristics of traffic signs that help the recognition [26] like familiarity, concreteness, complexity, meaningfulness, and semantic difference. These findings could be important in intelligent vehicle design too since these can help human drivers' understanding of the system processes.

In case of autonomous systems, it is crucial that the system communicates in a way that helps the drivers' information processing. Potential candidates to deal with these kinds of problems would be cognitive info-communication (CogInfoCom) [27], distributed cognition [28], or the field of human-computer interaction [29]. These multidisciplinary fields could provide recommendations regarding e.g., the type and modality of information the system provides for the user. In the case of autonomous cars, such an approach will be especially important, as the car sometimes has to effectively communicate complex representations of the system state to the driver.

3.1.3 Boredom and Engagement: Maintaining Vigilance

Autonomous driving systems aim to lower human mental load in order to optimize human performance. However, this approach is contradictory since the decreased mental load can cause passive tiredness and lower vigilance, which impairs attention and in a critical situation it may lead to wrong decisions. Also, drivers tend to prevent boredom, so in low-activity situations, they engage in non-driving activities behind the wheel (e.g. [30]), which is a problem because distracted driving causes higher reaction times in task-shift situations.

Reaction time is closely related to vigilance. Scientists usually treat it as a constant skill, though it is affected by several factors, like age, gender, and the actual mental state of the individual, among others. Reaction time tends to decrease with age, [13], [31]. Drivers under 30 have the lowest reaction times, and drivers above 60 the highest. These differences are enhanced by mental load [31]. Men react faster than women at any age [32]. In simulation research it was found that reaction time is also connected to human circadian rhythm and level of arousal: one tends to respond slower at 6 am, 2 pm. and 2 am, than at 10 am, 6 pm or 10 pm [33].

From the results of research in the fields of aviation and air traffic control, it is known that high level of automation can contribute to boredom [34]. Boredom can lead to distraction, higher reaction times, and even more mistakes [35] however this effect isn't specifically tied to autonomous vehicles. For example, Dahlen et al. [36] found that boredom proneness can predict unsafe driving. Cummings et al. [34] concluded that subjects during a simulation of an unmanned vehicle's supervisory control spent a lot of time being distracted, and they noted that the problem can be mitigated by efficiently switching attention. Driver boredom also shows connections with personality; a higher level of enthusiasm is associated with a lower level of boredom [35], while neuroticism seems to be positively correlated with boredom. Age also seems to have an effect on boredom: younger drivers experience boredom more often than older drivers [37]. It can be an aim of design to reduce boredom; for example, Steinberger [38] has found that gamification might be a key to solve this problem.

3.2 Situation Awareness Requirements for Driving

There are areas such as air-to-air combat fighters [39], infantry platoon leaders [40] or air traffic control personnel [41] where we can find information requirements for situation awareness. In the case of driving this research is at best in its infancy (see [42] for an example). However, the basis for this research in the form of descriptive driving models is very promising (see [43], [44]).

This study tries to add to this endeavor with a proposal regarding the structure of such requirements, which is also applicable to autonomous cars. At the core of this proposition is a sharp distinction between traditional driving tasks (e.g., choice of speed, steering, navigation, and monitoring of the traffic) and system-monitoring tasks (e.g., supervision of the decision of vehicles, or checking of operational conditions) of the driver. The relevance of these tasks is a function of the level of automation. Viewed simply the relevance of traditional driving tasks is higher at the lower levels of automation, while the relevance of system-monitoring tasks may be (but is not necessarily) higher at the higher levels of automation. To demonstrate this point recommendations for the levels of automation for on-road motor vehicles by the Society of Automotive Engineers [5] are used to take a look at the role of traditional driving tasks and of system monitoring tasks during driving at different levels of automation. The attentional demands and risk of boredom are also taken into consideration. These are overviewed in Figure 2.

The SAE recommendation proposes six levels of automation, ranging from level 0 (no automation) to level 5 (full automation). The difference of the levels can be easily grasped by stating who is responsible for performing the dynamic driving task (DDT) at different levels. The DDT contains essentially what Michon [43] defined as tactical and operational levels of driving tasks, and incorporates object and event detection and response (OEDR, monitoring of the driving environment), which is highly dependent on situation awareness.

At level 0 there is no automation, the driver is responsible for the whole DDT, however some vehicle systems may provide warnings or support, such as momentary emergency intervention. The role of traditional driving tasks are essential, while there is no need for monitoring of the system’s state thus there isn’t an increased need for the sharing of attention between these two. However, the driver needs to be receptive to the signals of the aforementioned vehicle systems. The driver needs to actively sustain their attention, and actively manage the DDT while they continuously receive feedback of their actions, so the risk of boredom is low.

Level 1 automation incorporates several driving assistant systems which manage parts of the DDT, such as lateral control of the vehicle (e.g., lane centering assist systems), or the longitudinal control of the vehicle (e.g., adaptive cruise control systems). The role of traditional driving tasks may be similarly crucial as at level 0, as these driving assistant systems are not able to manage the full operative control of the vehicle. The driver has to monitor the functioning of the driving assistant system, and intervene, if necessary, hence the role of system monitoring tasks is more important, and there is a need to share attentional capacity between the two tasks. The driver needs to stay ready to take over the full DDT. The risk of boredom is low because of the highly active role in the DDT, and the need for continuous monitoring of the traffic situation.

A car with level 2 automation manages part of the DDT, both the lateral and longitudinal control of the vehicle (i.e., the operative driving tasks). The driver manages the rest of the DDT, that is the tactical driving tasks, like selecting the correct maneuver for the current situation. System monitoring tasks are more important than on the previous level, the driver has to supervise many aspects of the system. Because of the relatively low-level of activity, the risk of boredom is higher than before.

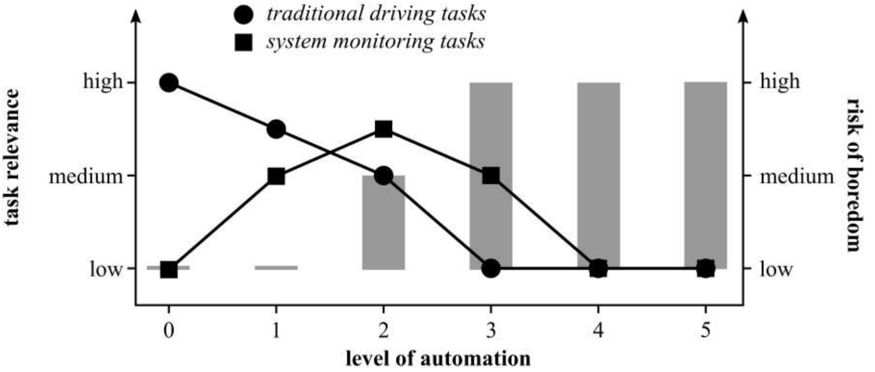


Figure 2
Supposed relevance of traditional driving tasks and system monitoring tasks and risk of boredom at different levels of automation

On level 3 the system manages the entire DDT, including the operative and tactical driving tasks. Besides these, it also monitors itself, is able to detect system malfunctions, and leaving of the operational domain. The driver does not practice any of the traditional driving tasks, and system monitoring tasks, however, all these tasks will be crucial if the driver has to take over the DDT (based on their own decision or because of system failure). Sharing of attention is not needed under normal circumstances, but in the case of takeover, there is an increased need for attention sharing. The risk of boredom is high in such scenarios, which is exceptionally dangerous if the driver needs to take over the full control of the vehicle.

Level 4 and level 5 are covered together. These are similar in that the vehicle at these levels deals with the full DDT, and is capable of performing fallback functions, or achieving minimal risk conditions in case of emergency, or if it leaves its operational domain. The main difference between these two levels is the range of situations it can handle, that is the broadness of their operational domains. As the driver does not have to do traditional driving tasks or system monitoring tasks, shared attention is not needed, but the risk of boredom is high. However, as the system is capable of intervention, and bringing the vehicle to safety, boredom should not be a cause of problem even in emergency situations.

Naturally, requirements regarding specific traditional driving tasks and system monitoring tasks can be distinct in different traffic scenarios and can vary depending on the system implementation at hand. These specific information must surely be taken into consideration when specifying the exact SA requirements.

4 Understanding Human Error: SHELL and Swiss Cheese Model

Technical development has solved numerous problems and ruled out errors in the field of traffic and transport, though the role of human error is still significant [45]. According to the findings of the International Civil Aviation Organization (ICAO), the last few decades can be described by a decreased number of machine errors and an increased number of human errors. The Civil Aviation Authority refers to Admiral Donald Engen that it is high time to focus on the human since the hardware field of traffic is quite reliable [46]. Two known and accepted frameworks of human error are to be considered: the SHELL model and the Swiss cheese model.

The conceptual framework of the SHELL model describes how humans as operators function in a complex system [46], [47], [48], [49], [50]. The model is widely used in aviation and traffic psychology and provides a framework for understanding and modeling human errors in a complex system. The original concept is introduced by Edwards [51], and modified by Hawkins [52]. The model places the human factor in the center, as it is the most crucial, yet the most flexible element of the system. The human factor is characterized by inconstant performance and limitations. This premise holds for autonomous vehicles as well since it is profoundly the driver, who monitors the functioning of the system and intervenes if necessary.

Figure 3 shows the building blocks and their connection points in the model. It is useful to notice that the edges of the blocks are irregular. It indicates that the components must be fitted carefully to the central component, aka the human component [46]. Each and every component has crucial characteristics that have to be taken into consideration when designing a system. The most important features of the liveware/human component are physical size, shape, physiological demands (food, oxygen, water, etc.), input characteristics (features of perception), characteristics of information processing, output characteristics (movement, communication), and environmental tolerances (temperature, noise, pressure, light, etc.).

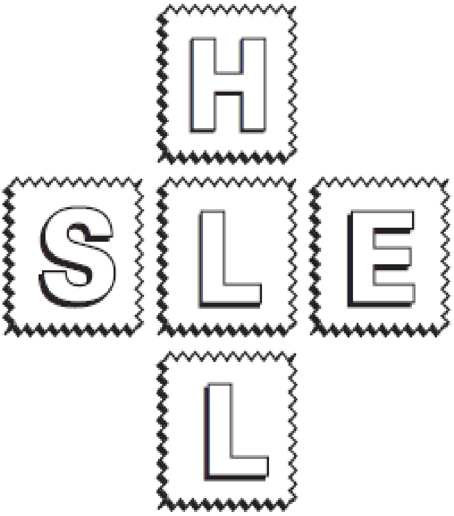


Figure 3
The building blocks of the SHELL model ([46], p. 12; S: software, H: hardware, L: liveware /human/, E: environment)

The fitting of liveware and hardware includes the aspects of human-machine interaction. Possible discrepancies are masked by human features, so problems usually emerge later in time. The liveware and environment fit was a focus of early research (e.g., protective clothing against harmful environmental effects). Lately, there are attempts to reshape the environment, so it fits the human needs (e.g., virtual functions). The liveware-liveware interaction is highly important in traffic psychology since human communication in teams and leadership are crucial in aviation. However, when dealing with autonomous vehicles it remains a question, how much importance it holds.

The liveware-software interaction means all the non-physical aspects of the system. Masking of errors might be highly critical here. The software element has an important role in the case of autonomous vehicles. Moreover, the functioning of the car and the IT system has to be extensively recognized. This means that it also has a training aspect, the driver has to study how the IT system of the vehicle works, how it decides, reacts in certain situations, and the limitations of the system have to be clear too. The system has to provide enough information to the driver in order to enable them to effectively monitor its functioning and to have a clear picture of the actual state of the system. At first glance, this seems to be easy, but it raises several questions. The information must be given "economically", it must be sufficient to gain an overview, but not too much, as it may unnecessarily increase the mental load. The timing of information is also critical, too early and too late are both problematic. For example, on motorways warnings are given earlier that is traffic signs are further away from the related elements than in a small village since at higher speeds more time may be needed to respond appropriately. Another question is the case of multiple warnings at the same time. Will, it mentally overload the driver, and if so, how much will it increase reaction times?

A new aspect emerges in the case of autonomous cars, namely software-software interaction that is when the different systems communicate with each other. However, the question of this kind of communication is outside the reach of human factor research.

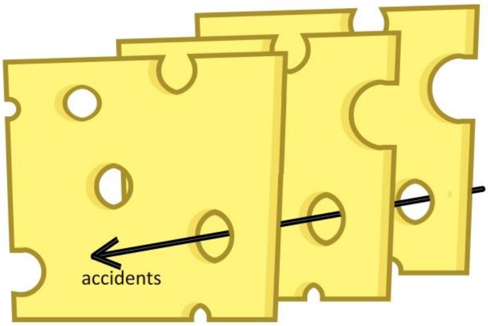


Figure 4
Swiss cheese model of errors

The Swiss cheese model introduced by Reason [53] sets the different levels/layers of prevention in focus (Figure 4). In an organization (e.g. aviation), and even in engineering, there are many types of defense in order to avoid accidents. These are represented by the layers (layered security). The holes on the layers represent both active failures and latent conditions (e.g. engineering, mechanical, control, maintenance, or application problems). The model proposes that accidents happen when these active failures and/or latent conditions coincide. It also illustrates how a potential accident can get through a safety system and how multicausal it can be.

Conclusions

In this paper, several psychological findings have been reviewed together with the role of the human factor in autonomous vehicles. Situation awareness was used as a central concept to guide the review. It is argued that during the design and engineering of autonomous vehicles the models and empirical findings of such research could be beneficial, and using this knowledge would help in constructing autonomous cars that are more user friendly and safer. A structure has also been proposed for situation awareness requirements that are applicable to traditional driving as well as to driving of autonomous vehicles. Having observed the abilities and limitations of human drivers regarding the operation of a self-driving vehicle, the legal aspects of autonomous driving should be considered. With this review, we hope to stimulate discussions on human factor issues amongst more technical approaches to autonomous vehicles since it is obvious that humans will play an integral part in this complex system in the foreseeable future.

Acknowledgement

This work has been funded by the European Union through the project titled "Felsőoktatási és Ipari Együttműködési Központ a Széchenyi István Egyetemen" (GINOP-2.3.4-15-2016-00003).

References

[1] M. Martens and A. van den Beukel, *The road to automated driving: Dual mode and human factors considerations*. 2013

[2] M. L. Cummings and J. Ryan, "Who is in charge?: The promises and pitfalls of driverless cars," *TR News*, pp. 25-30, 2014

[3] T. Rothengatter, "Psychological Aspects of Road User Behaviour," *Applied Psychology*, Vol. 46, No. 3, pp. 223-234, 1997, doi: https://doi.org/10.1111/j.1464-0597.1997.tb01227.x

[4] M. Kyriakidis *et al.*, "A human factors perspective on automated driving," *Theoretical Issues in Ergonomics Science*, Vol. 20, No. 3, pp. 223-249, May 2019, doi: 10.1080/1463922X.2017.1293187

[5] SAE International, "J3016B: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 2018 [Online] Available: https://www.sae.org/standards/content/j3016_201806/

[6] A.-M. Feyer, A. M. Williamson, and D. R. Cairns, "The involvement of human behaviour in occupational accidents: Errors in context," *Safety Science*, Vol. 25, No. 1, pp. 55-65, 1997, doi: 10.1016/S0925-7535(97)00008-8

[7] M. Cunningham and M. A. Regan, "Autonomous Vehicles: Human Factors Issues and Future Research," p. 12, 2015

[8] M. R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems," *Hum Factors*, Vol. 37, No. 1, pp. 32-64, Mar. 1995, doi: 10.1518/001872095779049543

[9] R. Okray and T. Lubnau, *Crew Resource Management for the Fire Service*, 50125th edition, Tulsa, Okla: Fire Engineering Books & Videos, 2003

[10] F. T. Durso and A. Sethumadhavan, "Situation awareness: understanding dynamic environments," *Hum Factors*, Vol. 50, No. 3, pp. 442-448, Jun. 2008, doi: 10.1518/001872008X288448

[11] K. Smith and P. A. Hancock, "Situation Awareness Is Adaptive, Externally Directed Consciousness:," *Human Factors*, 1995, doi: 10.1518/001872095779049444

[12] M. J. Adams, Y. J. Tenney, R. W. Pew, Y. J. Tenney, and R. W. Pew, "Situation Awareness and the Cognitive Management of Complex Systems," *Situational Awareness*, Jul. 2017, doi: 10.4324/9781315087924-4

[13] J. A. Groeger, *Understanding driving: Applying cognitive psychology to a complex everyday task*. New York, NY, US: Psychology Press, 2000

[14] G. Underwood, "Visual attention and the transition from novice to advanced driver," *Ergonomics*, Vol. 50, No. 8, pp. 1235-1249, 2007, doi: 10.1080/00140130701318707

[15] L. M. Hulse, H. Xie, and E. R. Galea, "Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age," *Safety Science*, Vol. 102, pp. 1-13, 2018, doi: 10.1016/j.ssci.2017.10.001

[16] K. J, N. V-M, K. K, and J. T, "Driving Characteristics and Development of Anticipation of Experienced and Inexperienced Drivers When Learning a Route in a Driving Simulator," Paris, France, 2012, p. 5

[17] E. A. Styles, *The psychology of attention, 2nd ed*. New York, NY, US: Psychology Press, 2006

[18] H. Shinoda, M. M. Hayhoe, and A. Shrivastava, "What controls attention in natural environments?," *Vision Research*, Vol. 41, No. 25-26, pp. 3535-3545, 2001, doi: 10.1016/S0042-6989(01)00199-7

[19] G. J. Hole, *The Psychology of Driving*, 1st ed. Mahwah, NJ: Routledge, 2006

[20] D. E. Broadbent, *Perception and communication*. Elmsford, NY, US: Pergamon Press, 1958

[21] D. Kahneman, *Attention and effort.* Englewood Cliffs N.J.: Prentice-Hall, 1973

[22] J. Wiley and A. F. Jarosz, "Working memory capacity, attentional focus, and problem solving," *Current Directions in Psychological Science*, Vol. 21, No. 4, pp. 258-262, 2012, doi: 10.1177/0963721412447622

[23] M. McCarty, K. Funkhouser, J. Zadra, and F. Drews, "Effects of Auditory Working Memory Tasks while Switching between Autonomous and Manual Driving," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60, No. 1, pp. 1741-1745, Sep. 2016, doi: 10.1177/1541931213601399

[24] B. R. C. Molesworth and D. Estival, "Miscommunication in general aviation: The influence of external factors on communication errors," *Safety Science*, Vol. 73, pp. 73-79, Mar. 2015, doi: 10.1016/j.ssci.2014.11.004

[25] D. Shinar, R. Dewar, H. Summala, and L. Zakowska, "Traffic sign symbol comprehension: a cross-cultural study," *Ergonomics*, Vol. 46, No. 15, pp. 1549-1565, 2003, doi: 10.1080/0014013032000121615

[26] S. J. P. McDougall, M. B. Curry, and O. de Bruijn, "Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness,

familiarity, and semantic distance for 239 symbols," *Behavior Research Methods, Instruments & Computers*, Vol. 31, No. 3, pp. 487-519, 1999, doi: 10.3758/BF03200730

[27] P. Baranyi, A. Csapo, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)* Springer International Publishing, 2015

[28] J. Hollan, E. Hutchins, and D. Kirsh, "Distributed Cognition: Toward a New Foundation for Human-computer Interaction Research," *ACM Trans. Comput.-Hum. Interact.*, Vol. 7, No. 2, pp. 174-196, Jun. 2000, doi: 10.1145/353485.353487

[29] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, *Human-Computer Interaction*. GBR: Addison-Wesley Longman Ltd., 1994

[30] H. Clark and J. Feng, "Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation," *Accident Analysis & Prevention*, Vol. 106, pp. 468-479, Sep. 2017, doi: 10.1016/j.aap.2016.08.027

[31] H. Makishita and K. Matsunaga, "Differences of drivers' reaction times according to age and mental workload," *Accid Anal Prev*, Vol. 40, No. 2, pp. 567-575, Mar. 2008, doi: 10.1016/j.aap.2007.08.012

[32] L. Warshawsky-Livne and D. Shinar, "Effects of uncertainty, transmission type, driver age and gender on brake reaction and movement time," *Journal of Safety Research*, Vol. 33, No. 1, pp. 117-128, 2002, doi: 10.1016/S0022-4375(02)00006-3

[33] M. G. Lenné, T. J. Triggs, and J. R. Redman, "Time of day variations in driving performance," *Accid Anal Prev*, Vol. 29, No. 4, pp. 431-437, Jul. 1997, doi: 10.1016/s0001-4575(97)00022-5

[34] M. L. Cummings, C. Mastracchio, K. M. Thornburg, and A. Mkrtchyan, "Boredom and Distraction in Multiple Unmanned Vehicle Supervisory Control," 2013, Accessed: Jul. 22, 2020 [Online] Available: https://dspace.mit.edu/handle/1721.1/86942

[35] S. Heslop, "Driver boredom: Its individual difference predictors and behavioural effects," *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 22, pp. 159-169, 2014, doi: 10.1016/j.trf.2013.12.004

[36] E. R. Dahlen, R. C. Martin, K. Ragan, and M. M. Kuhlman, "Boredom proneness in anger and aggression: Effects of impulsiveness and sensation seeking," *Personality and Individual Differences*, Vol. 37, No. 8, pp. 1615-1627, 2004, doi: 10.1016/j.paid.2004.02.016

[37] J. Harvey, S. Heslop, and N. Thorpe, "The categorisation of drivers in relation to boredom," *Transportation Planning and Technology - TRANSPORT PLANNING TECHNOL*, Vol. 34, pp. 51-69, 2011, doi: 10.1080/03081060.2011.530829

[38] F. Steinberger, R. Schroeter, and C. Watling, "From Road Distraction to Safe Driving: Evaluating the Effects of Boredom and Gamification on Driving Behaviour, Physiological Arousal, and Subjective Experience," *Computers in Human Behavior*, Vol. 75, Jun. 2017, doi: 10.1016/j.chb.2017.06.019

[39] M. R. Endsley, "A survey of situation awareness requirements in air-to-air combat fighters," *The International Journal of Aviation Psychology*, Vol. 3, No. 2, pp. 157-168, 1993, doi: 10.1207/s15327108ijap0302_5

[40] M. Matthews, L. Strater, and M. Endsley, "Situation Awareness Requirements for Infantry Platoon Leaders.," *Military Psychology*, Vol. 16, pp. 149-161, Jul. 2004, doi: 10.1207/s15327876mp1603_1

[41] M. R. Endsley and M. D. Rodgers, "Situation Awareness Information Requirements Analysis for En Route Air Traffic Control:," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1994, doi: 10.1177/154193129403800113

[42] M. Matthews, D. Bryant, R. Webb, and J. Harbluk, "Model for Situation Awareness and Driving: Application to Analysis and Research for Intelligent Transportation Systems," *Transportation Research Record*, Vol. 1779, pp. 26-32, 2001, doi: 10.3141/1779-04

[43] J. A. Michon, "A Critical View of Driver Behavior Models: What Do We Know, What Should We Do?," in *Human Behavior and Traffic Safety*, L. Evans and R. C. Schwing, Eds. Boston, MA: Springer US, 1985, pp. 485-524

[44] A. J. McKnight and B. B. Adams, "Driver Education Task Analysis. Volume I: Task Descriptions. Final Report (August 1969-July 1970)," Nov. 1970, Accessed: Jul. 23, 2020 [Online]

[45] A. Hobbs, "Human Factors: The Last Frontier of Aviation Safety?," *International Journal of Aviation Psychology - INT J AVIAT PSYCHOL*, Vol. 14, pp. 331-345, Oct. 2004, doi: 10.1207/s15327108ijap1404_1

[46] CAA, Civil Aviation Authority, Safety Regulation Group, and International Civil Aviation Organization, *Fundamental human factors concepts: (previously ICAO digest No. 1)* West Sussex, UK: Civil Aviation Authority, 2002

[47] M. Martinussen, D. R. Hunter, and D. R. Hunter, *Aviation Psychology and Human Factors*. CRC Press, 2017

[48] G. Molloy and C. A. O'Boyle, "The SHEL Model: A Useful Tool for Analyzing and Teaching the Contribution of Human Factors to Medical Error," *Acad.Med.*, Vol. 80, No. 2, pp. 152-155, 2005, doi: 10.1097/00001888-200502000-00009

[49] D. A. Wiegmann and S. A. Shappell, “Human error perspectives in aviation,” *The International Journal of Aviation Psychology*, Vol. 11, No. 4, pp. 341-357, 2001, doi: 10.1207/S15327108IJAP1104_2

[50] D. A. Wiegmann, S. A. Shappell, and S. A. Shappell, *A Human Error Approach to Aviation Accident Analysis : The Human Factors Analysis and Classification System*. Routledge, 2003

[51] E. Edwards, “Man and Machine: Systems for Safety,” in *Proceedings of British Airline Pilots Association Technical Symposium*, London, 1972, pp. 21-36

[52] F. H. Hawkins, *Human Factors in Flight*. Routledge, 2017

[53] J. Reason, *Human Error*. Cambridge: Cambridge University Press, 1990

Pedestrian Crosswalk Detection Using a Column and Row Structure Analysis in Assistance Systems for the Visually Impaired

Krešimir Romić, Irena Galić, Hrvoje Leventić, Marija Habijan

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Kneza Trpimira 2b, HR-31000 Osijek, Croatia
kresimir.romic@ferit.hr, irena.galic@ferit.hr, hrvoje.leventic@ferit.hr, marija.habijan@ferit.hr

Abstract: Computer vision-based approaches have become more common in assistance systems for the blind and visually impaired where portable devices can be used to assist users in their free movement. The method for pedestrian crosswalk detection with the main goal to facilitate the crossing of the road is proposed in this paper. The proposed crosswalk detection method is based on analyzing the image column and row structure. The Performance of this kind of approach relies on the input image resolution and quality. Therefore, guidance for selecting the appropriate input image resolution for this kind of approach is given. This approach is tested on the realistic input data captured with a monocular camera and using portable devices for image processing.

Keywords: assistance systems for the visually impaired; crosswalk detection; column structure; morphology

1 Introduction

The World Health Organization claims that there are 285 million people with some kind of visual impairment and 39 million of them are totally blind [1]. Most of them still do not use some advanced technological assistance systems, therefore, a white cane is the most widely spread assistance tool [2]. Advances in technology facilitate the development of advanced automated systems for navigation and orientation of blind and visually impaired people. This applies to the movement in both known and unknown environments where people are faced with different problems and obstacles. There are many types of problems and obstacles in the free movement of the blind and visually impaired, but this paper will deal with the problem of independent road crossing on designated zebra-style pedestrian crosswalks. This approach has the main goal to recognize a pedestrian crosswalk in front of a person in order to facilitate the road crossing on zebra crosswalks.

The proposed approach uses image processing techniques to get useful information about the environment from input frames from video sequences taken with a camera. Camera-based assistance systems are a very common solution for this problem [3], [4], [5]. In this case, we propose the method for pedestrian crosswalk recognition from video frames. The image processing techniques used for this purpose include white balancing, edge detection, and morphological operations extended with a specially tailored vertical and horizontal analysis of binary frames.

An algorithm is performed using a simple monocular camera and a portable computer as a processing device. The final output is information about the presence of a crosswalk in front of a person. Additional information about the position of the crosswalk in the image is also provided. The way of informing a person about the crosswalk is out of the scope of this paper, but some preliminary tests were conducted by using stereo sound signals [6]. Similarly, Mascetti et al. in [7] proposed the sonification of guidance data during road crossing.

This paper is organized as follows: the second chapter explains the important preprocessing actions; the third chapter explains the proposed method in detail; the fourth chapter provides the experimental results with the obtained performance and accuracy and the last chapter draws up a conclusion and provides guidelines for further work. Figure 1 shows a brief overview of the main steps of the proposed method.

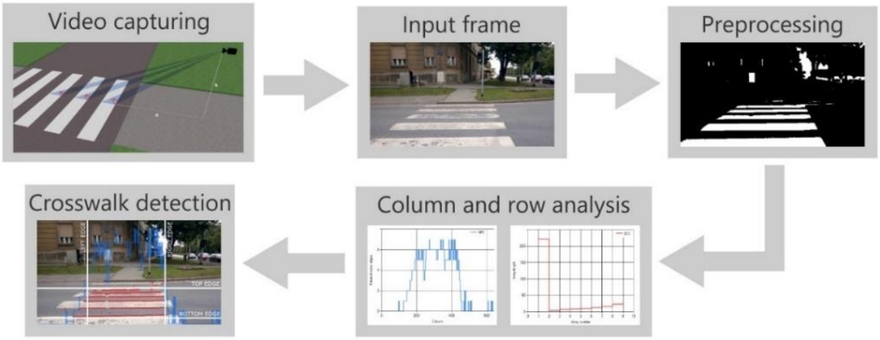


Figure 1
The method overview

1.1 Related Work

According to the survey [8], there are three groups of problems in individual movement of the blind: far, intermediate, and near distance tasks. Intermediate and near distance tasks like obstacle detection [9], space perception [10], and reading [11] are often solved using camera and image processing techniques.

Independent road crossing belongs to the group of intermediate distance tasks. Our previous work and similar approaches that deal with this problem based on image processing will be presented below.

When talking about zebra-style crosswalks, detection algorithms are often based on techniques used for detecting similar objects, e.g. staircases like in [4] and [12]. Our previous research on staircase detection proved that column-level analysis can be useful to localize free space and route the visually impaired people on the staircases. This research tries to implement and adopt similar principles to the problem of crosswalk detection. Furthermore, our previous research on crosswalk detection was more focused on potential crosswalk ROI extraction in preprocessing step in order to avoid whole image processing in higher resolutions, while this research presents the detailed column and row structure analysis method for crosswalk detection which yielded better results.

Similar to staircases, pedestrian crosswalks are characterized by rectangular shapes that appear periodically in an image making it suitable for using Hough transformation for detecting parallel lines as presented in [4]. The authors in [4] used RGBD camera which provides useful depth information; however, it is more expensive and impractical to use in motion. Detection of parallel vanishing lines is presented in [13] where a crosswalk slope is also approximated, but unlike the approach [4], the authors used a standard camera.

The use of a classic monocular camera is more frequent in real-time crosswalk detection systems. The mean shift segmentation and morphological processing are used to separate the crosswalk from the environment in [14] forming the robust and nonparametric system. The main disadvantages are lower detection rates due to larger shadows and very bright areas. Another approach [15] proposes a crosswalk detection algorithm based on bipolarity and projective invariant. This approach emphasizes the performance in various lighting conditions caused by sun, clouds or rain. In [16] and [17], figure-ground segmentation is used to detect crosswalks and the direction of the crosswalk is provided by finding a vanishing point. Figure-ground segmentation proved to be more robust than Hough transformation in cases of local deformations of straight lines which are often a part of crosswalk stripe edges. Cheng et al. [18] compared the conventional bipolarity method with a more novel adaptive extraction and consistency analysis (AECA) method on their own dataset in order to cope with inevitable aggravating capturing conditions. Adaptations of methods from driver assistance systems are also common solutions for crosswalk detection [19]. However, it can be concluded that the field of assistive technologies for the visually impaired is here neglected compared to vehicle and driver assistance systems, so this work certainly increases the awareness of this kind of problem.

Some authors recommend the development of algorithms tailored for execution on smartphones [3], [5] to take advantage of small devices with integrated cameras and additional help of GPS [20]. The mentioned smartphone-based approaches also investigate the usage of panoramic images in combination with satellite imagery. Authors in [21] present a software module called *ZebraRecognizer* with the main aim to remove projection distortion on acquired images to improve the crosswalk recognition accuracy.

In some recent scientific papers, machine learning techniques were used to solve crosswalk detection problems. Deep learning was successfully employed to detect crosswalks in [22] but the system is tailored for driver help and autonomous vehicles, so the camera perspective is different. The authors in [23] have used deep learning and they focus their research to detect crosswalks in different orientations. Authors also claimed the necessity for dataset extension to achieve better results with this approach. In [24], authors have dealt with crosswalk start and end-point detection with an accent on pedestrian traffic light recognition using convolutional neural networks. In mentioned work, testing was conducted only on images with crosswalks and lights. Even though the deep learning methods are becoming more common in this field, we have noticed that there is still a difference in approaches and methodology in such papers which, in combination with the lack of public datasets, makes them hardly comparable.

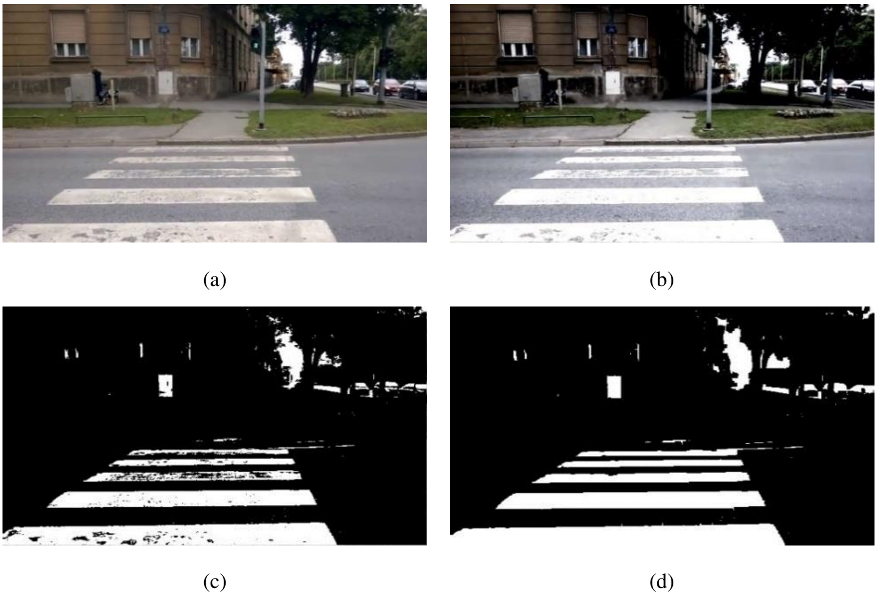
The shortcomings of the already proposed solutions can be summarized as follows: problems with shadows, obstacles and lighting conditions, problems with distortions caused by cameras, and the lack of unified test data especially for machine learning-based techniques.

Therefore, the goal of our method was in achieving good results without previous corrections regarding the projection distortions. Our approach actually uses the capturing perspective in order to detect characteristics of the crosswalk in particular camera views. This paper aims at a robust crosswalk detection method which will work in aggravating capturing conditions. Integrated crosswalk localization will be the additional benefit of the proposed method on the way to the final navigation system. The focus will be on the usage of medium and low-performance devices and cameras as the final goal is to develop a cheap and widely available system for the navigation of visually impaired persons.

2 Preprocessing

It is important to develop a method that will be able to work with frames captured from video sequences. Those frames can be considered as input data and a number of processed frames per second will show the performance of the method. The example of an input image is shown in Figure 2 a). The method is also geared to work on a video captured with a standard cheap monocular camera. Regarding capturing conditions and camera quality, we encountered several problems with input data. Those problems include different lighting conditions, unfocused and blurred frames due to camera movement. In order to solve the aforementioned problems, the preprocessing step enhances the visibility of the potential crosswalk and thereby prepares the image for further processing.

The initial preprocessing step starts with the white balancing of an input frame. Even though the human color perception of zebra crosswalk stripes indicates a white color, those stripes are often represented with slight gray and yellow shades in digital images depending on lighting conditions and camera specifications. To obtain clear white stripes of a potential crosswalk, white balancing is performed. For this purpose, we used a specialized Gray-world algorithm [25] to emphasize the white color as much as possible. This algorithm is chosen because of its simple implementation and low processing time. The effect of this process is presented in Figure 2 b).


(a) (b)

(c) (d)

Figure 2
Preprocessing: (a) Input image, (b) White-balancing, (c) Thresholding, (d) Closing

The next step is thresholding where near-white regions become white and all other colored regions become black. This step is necessary to emphasize the white regions of interest and thereby eliminate irrelevant image regions. Thresholding is performed on every color channel (red, green, and blue). For every pixel, all three channel intensities must be larger than the empirically obtained threshold value (220) and additionally, the difference between the two-channel intensities must be less than 20. Pixels that match the given criteria become white and all channels are set to value 255, otherwise to 0. The previous white color balancing allows us to use the fixed threshold values rather than adaptive methods thus reducing processing time. The final result of this step is a binary image as shown in Figure 2 c). The following equation shows thresholding conditions:

$$f^*(x,y) = \begin{cases} 255, & \begin{array}{c} if\ f_r(x,y)>220 \\ AND\ f_g(x,y)>220 \\ AND\ f_b(x,y)>220 \\ AND\ |f_r(x,y)-f_g(x,y)|<20 \\ AND\ |f_r(x,y)-f_b(x,y)|<20 \\ AND\ |f_g(x,y)-f_b(x,y)|<20 \end{array} \\ 0, & otherwise \end{cases} \tag{1}$$

where $f_r(x,y)$ represents the value of the red color channel of pixel with coordinates x,y. Similar to this, $f_g(x,y)$ represents the green channel and $f_b(x,y)$ the blue channel. Values that form the new binary image are represented with $f^*(x,y)$.

The final part of preprocessing is associated with possible imperfections on crosswalk stripes due to age and faded white paint on the road. First, the morphological closing operation is performed to fill narrow and small black regions with the white color [26]. The appropriate size of the structuring element for closing operation is chosen based on the resolution of the image which is explained in more detail in the experimental phase of this research (Chapter 4). An additional algorithm is tailored for finding and filling gray gaps on the road between the white color regions. Horizontal arrays of black pixels are compared in length with the surrounding horizontal white pixel arrays on the right and left. Every black array of pixels shorter than the surrounding white arrays is substituted with white pixels. This step successfully fills the most of black gaps between white horizontal regions which are characteristic of crosswalk stripes. The final binary image after the preprocessing step is shown in Figure 2 d).

3 The Method for Crosswalk Detection

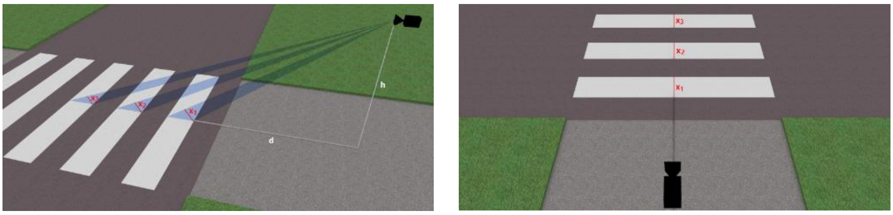
The main part of the proposed method for pedestrian crosswalk detection is explained in this chapter. The method is primarily based on the structure of image columns in the crosswalk region. The success of the method relies on the assumption that finding the specific characteristics of the column structure will allow us to detect which columns of the image belongs to the crosswalk region. The chapter is divided into three sections - column analysis, row analysis, and crosswalk localization.

3.1 Column Analysis

As opposed to the existing approaches which are often based on searching for nearly parallel horizontal lines [4], [5], [17], the proposed method is essentially based on a vertical analysis, i.e. a column analysis. The idea is to analyze every column of an input image and find features characteristic for a crosswalk region. If a camera is located on height h (approx. 1.5 m) and the crosswalk is in front of a camera on distance d (approx. 1-3 m), it is necessary to analyze how the white stripes are getting thinner with the increase of distance from a camera as shown in Figure 3 a). Further, in Figure 3 b), the camera perspective is shown with designated vertical red lines that represent white stripe width (x_1, x_2, x_3). It is obvious that those lines will gradually increase in length observing them from the top to bottom on captured video frames. It is experimentally obtained that the vertical white and black lines increase in length by 20% to 100% as shown in the following equation:

$$1.2 < k = \frac{x_n}{x_{n+1}} < 2.0 \tag{2}$$

where x_n is the length of n-th vertical array and k is the length increasing coefficient. It can be said that the vertical white and black lines increase in length in every particular column likewise members of geometric series, but with a certain aberration of k parameter. The estimated aberrations of k parameter shown in (2) are valid when the frame is captured from the distance 1-3 meters with a camera of 1.5 m height.



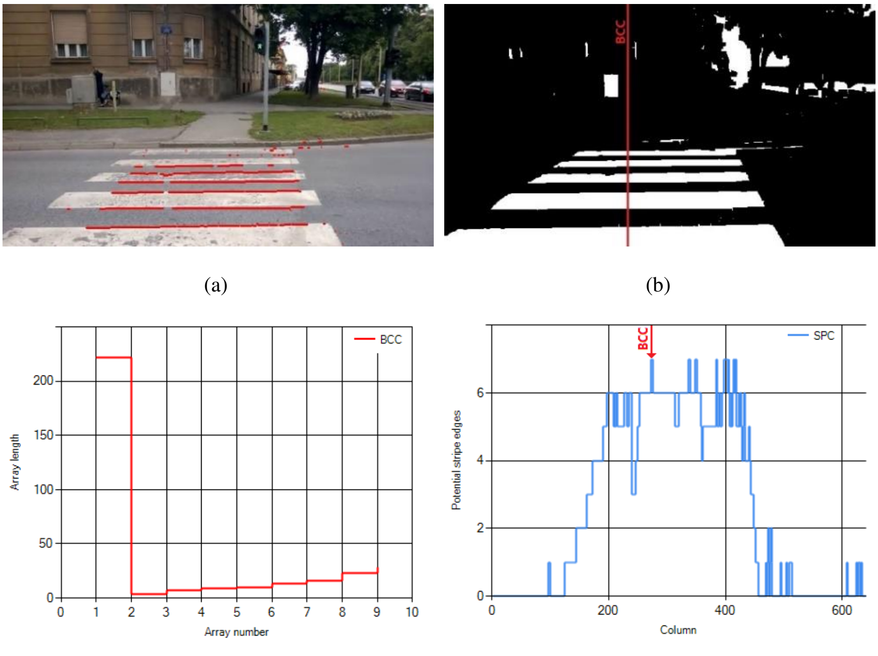
(a) (b)

Figure 3

Scene with camera and crosswalk: (a) Side perspective, (b) Camera perspective

An algorithm is developed to analyze every column of the binary image (Figure 4 b) obtained after preprocessing. The analysis is performed by measuring the length of the vertical white and black pixel arrays. The length of consecutive vertical pixel arrays of the same color (black or white) are compared and every increase that matches the required k parameter is counted as a potential stripe edge. Those potential stripe edges are marked with red dots on the input image illustrated in Figure 4 a). The vertical black pixel arrays are also tested because the space between white stripes acts similarly to stripes on the captured images.

When the column designated with a red line in Figure 4 b) is analyzed, the lengths of black and white vertical arrays have values as shown in Figure 4 c). To illustrate, this column is chosen because it is the column with the highest number of potential stripe edges and it is called the best column case (BCC). The graph in Figure 4 c) shows that arrays 2 to 9 gradually increase in length as requested in the mentioned condition (2) and 7 potential stripe edges are detected in that particular column. This process is repeated for every column and the graph of potential stripe edges per column (SPC) is generated (Figure 4 d). It is evident that the columns in the crosswalk region have much higher values. Supposing that a crosswalk consists of at least 2 stripes, columns with 4 or more potential stripe edges are counted. A higher number of such columns implies a higher possibility for finding the crosswalk in the image. Images with at least 10 columns with 4 or more potential stripe edges are forwarded to further analysis. Otherwise, images with less than 10 such columns are discarded and a decision about not finding a crosswalk is made. Once we have detected columns where crosswalk stripes are present we can use that information to find discontinuation in an array of crosswalk columns. This can help us to detect potential obstacles present on the crosswalk which is another benefit of this approach compared to related work.



(a)

(b)

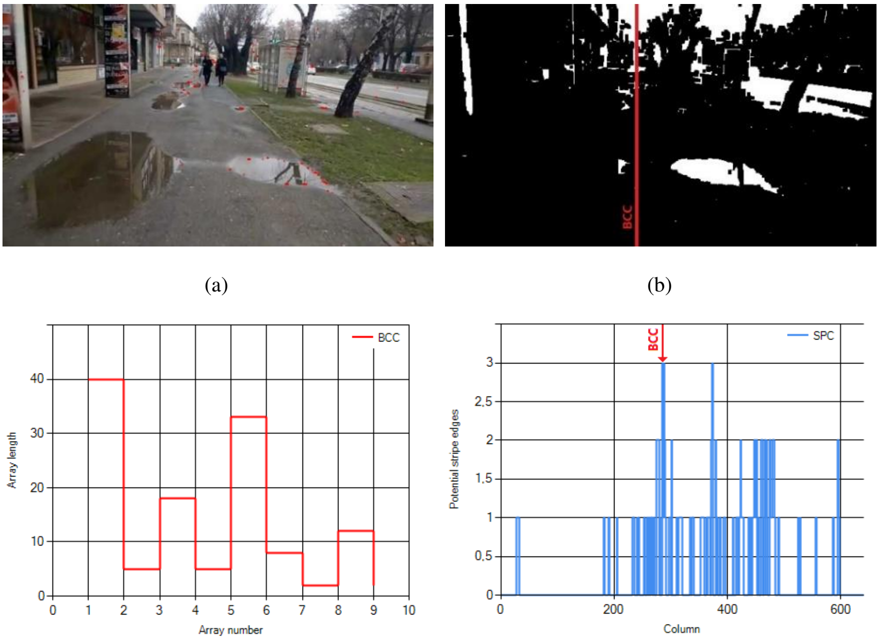
(c)

(d)

Figure 4
Column analysis (with crosswalk): (a) Detected potential crosswalk edges, (b) Analyzed binary image, (c) Array lengths in the best column case, (d) Potential stripe edges per column

In comparison to the previous example, Figure 5 shows a detection process for an image example where a crosswalk is not present. Figure 5 a) shows the input image without the crosswalk where some points (red dots) were declared as potential stripe edges and those points are scattered on the entire image. The potential stripe edges are obtained by analyzing the binary image (Figure 5 b). If we observe a particular column with the highest number of potential stripe edges (red line in Figure 5 b), it is clear that the lengths of black and white vertical arrays do not have a constant increase in size characteristic for a crosswalk region. Those array lengths for the best column case are shown in Figure 5 c). The graph in Figure 5 d) shows that none of columns have more than 3 potential stripe edges. The overall number of potential stripe edges per column is much lower than in the previous case (Figure 4 d). An image like the one in Figure 5 a) is discarded at this moment considering that it does not have at least 10 columns with at least 4 potential stripe edges.

Putting the column analysis in the first place makes this method resistant to smaller obstacles that obstruct the view on a crosswalk. By having a larger range of parameter k values, the method is more resistant to different capturing angles, which are often problematic in similar approaches.



(a) (b)

(c) (d)

Figure 5

Column analysis (without crosswalk): (a) Detected potential crosswalk edges, (b) Analyzed binary image, (c) Array lengths in the best column case, (d) Potential stripe edges per column

3.2 Row Analysis

The last step of the method is the row analysis, i.e. the horizontal analysis which can be considered as an additional check. The row analysis is performed on a new image generated only from columns that passed the aforementioned column analysis with more than three potential crosswalk stripe edges. Generated images are narrower than the input image and are trimmed at the nearest and farthest potential stripe edge. In the trimming process, it is important to exclude isolated potential stripe edges that appear in distant positions from other potential edges because they unnecessarily increase the generated image size. The examples of generated images for the row analysis are shown in Figure 6.

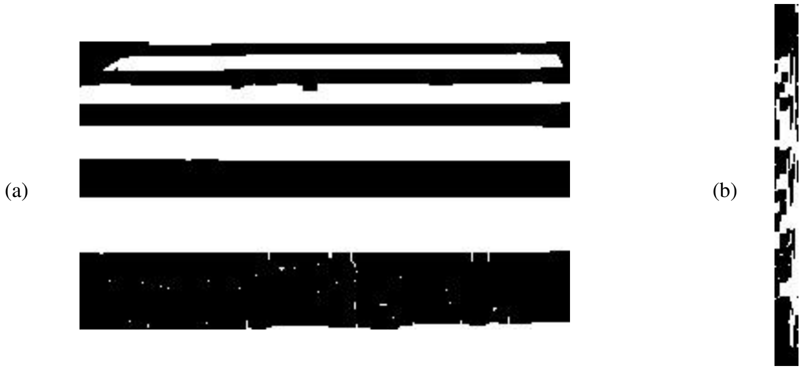


Figure 6

Generated images for row analysis from input frame: (a) with crosswalk, (b) without crosswalk

The idea is to check whether the isolated columns are actually part of a crosswalk region or they have accidentally passed the previous step. The images generated from the input frames with a crosswalk (Figure 6 a) are characterized by wide rectangular shapes with lower rates of horizontal transitions from white to a black color and vice versa. In comparison, the images generated from the input frames without a crosswalk (Figure 6 b) are often narrower and have higher rates of horizontal transitions. Those images do not have distinctive horizontal stripes because they consist of columns from different parts of the input frame that accidentally passed the column analysis. In order to distinguish whether the generated images contain a crosswalk region or not, a special parameter called horizontal energy (HE) is calculated. The parameter is tailored to analyze row characteristics. It is calculated by the following equation:

$$HE = \frac{\sum_{i=0}^{h} l_{max}(i)}{\sum_{i=0}^{h} tr(i)} \tag{3}$$

where h is the image height, $l_{max}(i)$ is the length of the longest array of consecutive white or black pixels in row i, while $tr(i)$ is the number of transitions in row i. In this way, it can be decided whether the column analysis successfully detected a crosswalk region or it was a wrong assumption because HE will have higher values for images generated from the input frames with a crosswalk. On the other hand, images generated from the input frames without a crosswalk have lower HE. The threshold value for the final decision is obtained experimentally and is set to 2. Images with HE higher than 2 are marked as positive (i.e. contain a crosswalk) and other ones are discarded and marked as negative (i.e. do not contain a crosswalk).

3.3 Crosswalk Localization

Regarding assistance systems for the blind and visually impaired, it is important to provide simple information about a crosswalk position relative to a person [20]. If we observe data from a camera installed on a person, a vertical position of the crosswalk in the image is considered as irrelevant because it depends on a camera positioning angle and it is assumed that the crosswalk is on the same level as a person. On the other hand, it is important to emphasize the horizontal position of the crosswalk relative to a person. It can be very useful for a person to have real-time information whether the crosswalk is on the left, right, or in front.

In order to provide the crosswalk position on images, the previously shown layout and concentration of potential stripe edges (Figure 4 a) and Figure 4 d) are used. The combination of the required data is shown in Figure 7. When observing the columns from the left to the right, the first column with more than 3 potential stripe edges can be considered as the left edge of the crosswalk region. Similarly, the last column with more than 3 potential stripe edges is marked as the right edge of the crosswalk region. Those edges are marked with vertical white lines in Figure 7.

In images with the detected crosswalk, points that represent a potential stripe edge are concentrated in the crosswalk region, while a smaller number of points is isolated in other parts of the image. Those isolated points should be excluded and then the top and the bottom of the crosswalk can be approximately detected as the highest and the lowest potential stripe edge. The positions of those edges that determine the vertical position are marked with horizontal white lines in Figure 7. Having the left, right, top, and bottom edges, it is possible to outline the rectangle around the approximate position of the crosswalk. Other examples of localized crosswalks will be presented later in the chapter with the experimental results.

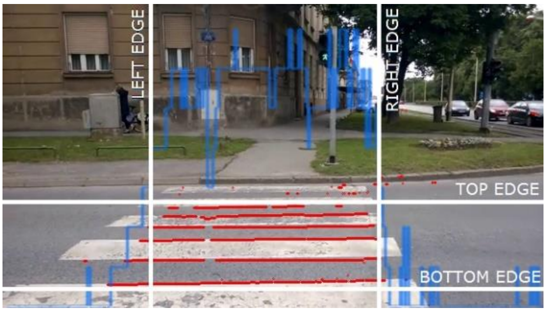


Figure 7
Crosswalk localization

4 Experimental Results

This chapter presents the experimental results obtained in realistic conditions. After initial experiment settings, the accuracy and performance of the proposed method are presented and compared to similar methods.

4.1 Experiment Settings

The proposed method is implemented in the C# programming language and the test environment is made for the sake of easier parameter monitoring and analyzing the method steps. The input video for processing can be captured directly from a camera or a video file. The resulting images of a particular method step can be analyzed for every frame of the input video along with a decision about detecting crosswalks and the crosswalk position.

The experiment was conducted by capturing video sequences in realistic conditions. The videos were captured in motion with a camera installed on a person at approximately the height of 1.5 meters (chest level). For this purpose, we used a standard monocular camera with 16:9 aspect ratio. Captured video sequences are prepared for testing in three different resolutions: 320×180, 640×360, and 1280×720. In contrast, to the RGB-D cameras used in a similar work [4], monocular cameras are cheaper and smaller, which makes them easier to use when they are installed on a person in motion. It is necessary to test the method on the frames from video sequences where people are approaching pedestrian crosswalks, but also from sequences without crosswalks. On video sequences that contain a crosswalk, it is important to capture a moment of approaching where a person is located on a distance of 1-3 meters from the crosswalk. A distance of 1-3 meters is a prerequisite for the proposed method to work properly.

Since this method is supposed to be used in assistance systems to help the blind in movement, two portable devices are chosen for the experiment. The first processing device is a laptop which brings an advantage in terms of higher computational power, but it is less portable due to its dimensions and weight. The second processing device is a mini PC with benefits in small dimensions, weight, and low power consumption, but with significantly less computational power.

4.2 Accuracy

When processing video sequences from files or directly from a camera, frames are processed one after another and real-time information about crosswalk presence is provided along with crosswalk approximate position. In order to obtain accuracy results, the representative frames were extracted from the video sequences for our database. There are 150 frames with a scene containing a crosswalk and 150 frames with a random scene not containing a crosswalk. It is crucial to test both groups of situations to avoid as much as possible false positive detections in scenes without a crosswalk. When collecting the input dataset, the videos captured in various aggravating conditions were chosen. This considers crosswalks with faded or soiled stripes on videos captured in different lighting conditions and from different angles. The input dataset does not include night scenes and the proposed method is primarily tailored to work with day scenes with natural light present. For night scenes, it would be necessary to use an additional lighting source, e.g. infrared lamp.

Testing is performed using images in three different resolutions and with three different structuring element sizes for closing operation as shown in Table 1. This experimental setup allows us to choose appropriate input resolution for this method and to choose the best closing parameter for a particular resolution. The results have shown that the resolution of 640×360 and closing structuring element size of 5×5 yield best accuracy rates. In that case, there are 98.7% correctly detected crosswalks and false-positive detections are present in only 1.3% of cases. These results are slightly better in comparison to our previous multiresolution approach [27].

Similar crosswalk detection accuracy is obtained on higher resolutions but with more false-positive detections due to unnecessary details visible in high-resolution images. Therefore, it can be concluded that it is unnecessary to use higher resolutions for this problem. On the other hand, it has been shown that lower resolution requires a smaller structuring element for closing and 76.67% of correct crosswalk detections can be obtained.

Table 1
Accuracy results based on image resolution and closing parameters

Resolution	320×180			640×360			1280×720		
Closing structuring element size	3×3	5×5	7×7	3×3	5×5	7×7	3×3	5×5	7×7
Detections (%) [with crosswalk]	76.7	52.7	22.7	94.7	98.7	93.3	94.7	97.3	97.3
Detections (%) [without crosswalk]	0	0	0	3.3	1.3	0.7	10	8.7	10

To our knowledge, there is no widely-used and available image collection for testing these kinds of methods, thus the direct comparison with the methods presented in the related work is not feasible. Authors mostly use their own databases for testing and there is a possibility that the results may be biased to some extent because the dataset is collected by the authors. Therefore, the proposed method is compared with our previous method on the same private dataset of 300 images. Furthermore, the proposed method is compared with the AECA (adaptive extraction and consistency analysis) method proposed in [18] on the dataset that authors provided online [28]. Comparison on our private database helps us to follow the progress of our own research in this field, while the comparison on the dataset provided by other authors helps us to test the robustness of our method on new images taken in different environments with different cameras. We found this comparison useful because mentioned work has the accent on crosswalk detection in aggravating conditions as is the case in this paper. Those conditions include pedestrian occlusion, low-contrast crosswalks, various illuminances, etc. Both datasets include images with and without crosswalks. We have measured the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for both comparisons and precision and recall metrics were calculated using the following equations (4).

$$Precision\ (\%) = \frac{TP}{TP+FP} * 100 \qquad Recall\ (\%) = \frac{TP}{TP+FN} * 100 \tag{4}$$

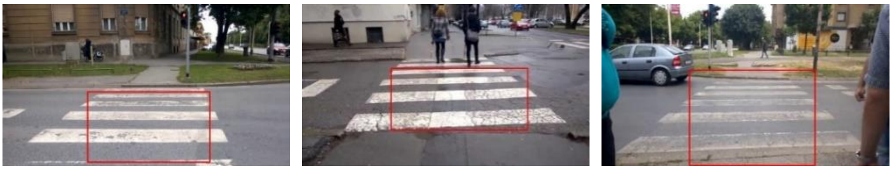
In the first comparison of our private dataset we have achieved slightly better results with precision and recall at 98.7% which is more than the previous 96.6% (precision) and 94.7% (recall). This method is computationally more demanding than the previous one but gives better accuracy rates which suggest that a certain trade-off between accuracy and processing speed is still unavoidable. The second comparison of the aforementioned public dataset implies that the proposed method is competitive with the AECA method. The proposed method gained a higher precision rate (94.8%) compared to the AECA method (84.6%). Similarly, the recall rate is also higher, 73.1% compared to 60.1%. It must be said that only overall results are compared and none of the subsets of data were tested individually. All accuracy rates of the proposed method compared to similar approaches are shown in Table 2.

Table 2

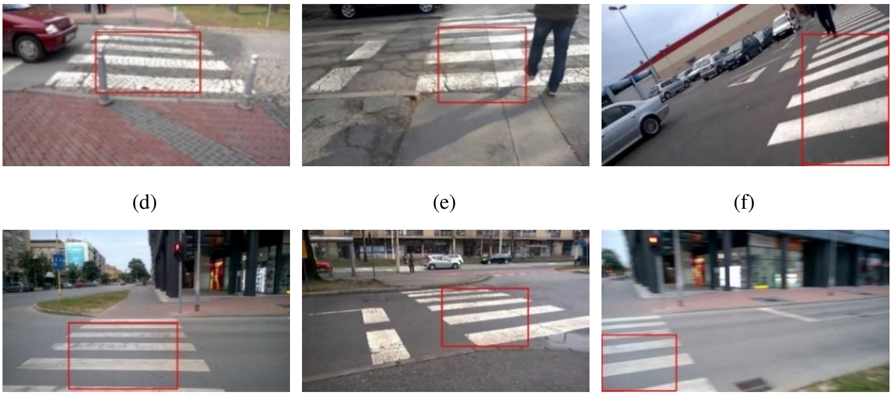
Accuracy results compared to similar methods

Dataset	Method	TP	TN	FP	FN	Recall (%)	Precision (%)
Dataset 1 (300 images)	Proposed method	148	148	2	2	98.7	98.7
	Previous multiresolution method	142	145	5	8	94.7	96.6
Dataset 2 (452 images)	Proposed method	239	112	13	88	73.1	94.8
	AECA algorithm	187	104	39	122	60.1	84.6

The examples of positive detections with localized crosswalks are shown in Figure 8. The test frames show that the method works very well in various situations. Figure 8 a) shows detection in usual daylight conditions, while Figure 8 b) shows detection on wet roads due to rain. Detection is also successful on faded crosswalk stripes (Figure 8 c) or on barely focused frames (Figure 8 d). Detection is possible even with the pedestrian presence on the crosswalk as shown in Figure 8 e). Figure 8 f) shows a situation when a camera is not in a completely horizontal position. Lower light conditions are present in the example in Figure 8 g). A situation when a person approaches the crosswalk at a certain angle is shown in Figure 8 h). The method proved to be robust to blurred frames as visible in Figure 8 i) due to thresholding and closing operation in the preprocessing step. For this reason, the deblurring procedure is not included in preprocessing in order to reduce processing time. Figure 8 i) also shows detection when the camera is not strictly directed to the center of the crosswalk and the pedestrian should be directed to the left side. Some test images show that missed detections are caused by shadows on a crosswalk surface, which makes binary images unusable for detection. Using the proposed method, false-positive detections are very rare and can occur only when shapes very similar to crosswalk stripes appear in an image.



(a) (b) (c)



(d) (e) (f)

(g) (h) (i)

Figure 8
Examples of positive detections (a)-(i)

4.3 Performance

According to [29], real-time implies that the response time must be enough to avoid a failure of the system. When talking about real-time video processing for this purpose, it is important to determine a minimal processing frame rate sufficient to beforehand inform a user about a crosswalk (i.e. before standing on the crosswalk). If we assume that the average walking speed is about 1.4 m/s [30] and the method detects a crosswalk from at least 1 meter distance, it is sufficient enough to process 2 frames per second because in that case, a user will only pass 0.7 meters while processing one frame. However, it is preferred not to take this information for granted because of a possible fault in particular frames due to capturing conditions in motion (e.g. unfocused and blurred frames). Furthermore, it would be good to make a final detection decision based on several consecutive frames, so it is preferable to process several frames per second.

The average processing speed obtained using the proposed method on input images in three different resolutions is given in Tab. 3. It is clear that the processing speed will also depend on the hardware specifications of processing devices, so two portable devices are chosen for the purpose of testing. The first device is a mid-range laptop and the second one is a mini PC with lower performance and power consumption. The detailed specifications of the used devices are also given in Tab. 3. Since we are aiming at developing a system that is dedicated only for assisting the visually impaired persons in their movement, we did not consider using smartphones as target devices. Even though it will be easy to deliver the system to potential users as a smartphone application, smartphones are also used for other everyday activities and the availability of processing resources could be occupied by other working applications on the smartphone or users can be distracted by other smartphone features. Another disadvantage is the battery life which can be very low due to demanding processing tasks. Therefore, we decided to test our method on a laptop, but also on a mini PC. This device uses 64-bit Intel Atom processor so it was convenient to use the same developed algorithm and libraries on both test devices. For further development of the fully usable assistance system, it will be necessary to adapt algorithms for embedded devices. However, for this stage of research we have found laptops and mini PC-s as good prototype platforms for our system.

Table 3

Processing speed on portable devices

		Laptop	Mini-PC
Specifications	Processor	Intel Core i5-3210M @ 2.50 GHz (2 cores)	Intel Atom Z3735F @ 1.33 GHz (4 cores)
	RAM	8 GB	2 GB
	Dimensions (W×D×H)	37.6×25.7×2.8 cm	10×3.8×1.5 cm
	Weight	2.5 kg	0.3 kg
Processing speed (per frame)	320×180 px	≈41 ms	≈158 ms
	640×360 px	≈150 ms	≈607 ms
	1280×720 px	≈634 ms	≈2579 ms

Using the 640×360 resolution frames, which yield the best accuracy results, the average processing speed is about 150 milliseconds on the laptop and about 607 milliseconds on the mini PC. When using the laptop, the information about the crosswalk presence is provided 6.67 times in a second, which can be considered as sufficiently fast to use such a system in reality. On the other hand, when using the mini PC with lower computational power, the information is provided 1.65 times in a second, which can be insufficient for some real situations. However, it is important to note that the size and weight of the mini PC is more suitable for use in motion. Since this paper puts the emphasis on accuracy, there is still enough space for algorithm optimization or parallel implementation, which will increase the processing speed. One of the solutions for improving the processing speed is to adapt the method to increase the accuracy on low-resolution images (e.g. 320×180) where a processing speed of nearly 24 fps is obtained.

Conclusion

The main contribution of this paper is a column-based approach for detecting crosswalks in video frames. This work brings some benefits to this research field by putting the accent on detecting the characteristic constitution of the columns in the crosswalk region. This concept allows robustness even when the part of the crosswalk is occluded. Additional contribution is proposed, horizontal energy parameter which is necessary for final decision making. This work has also brought the guidelines to choose the appropriate input resolution and preprocessing parameters for this kind of approach.

The proposed algorithm for crosswalk detection brings benefits to assistance systems for easier movement of the blind and visually impaired. The developed method for crosswalk detection is robust and yields high rates of correct detection. The method proved to be successful in aggravated conditions caused by lighting variations, faded crosswalk stripes, capturing angle, and camera shaking. Problems like variations in white color representation caused by shadows make space for the improvement of the method. Further research activities in this field will be focused on improving the method and adding additional features like detecting other pedestrians on the crosswalk.

The developed algorithm is tailored to perform on extensively used devices like laptops or mini PCs with a simple monocular camera. It is important to note that the proposed method is not dependent on camera type and the algorithm is tailored to work with images from different sources (e.g. web cameras, sports cameras, and camera glasses). Those facts make a potential assistance system cheaper and more discreet when installed on a blind person. The latter is very important not to additionally designate people with special needs in public.

The processing speed of the proposed algorithm is already sufficient for performing on mid-range laptops, but there is still space for optimization and improvements for performing on small devices like mini PCs. Although small devices will have increased computational power in the future, for now, it is necessary to make a compromise between the computational power and the size of a device. The final idea is to eventually develop a system that will integrate several types of aid for the movement of the blind. Therefore, the development and constant improvement of systems like the one presented in this paper is a necessity.

To sum up the contributions of this research it can be said that it brings better accuracy results proved on a private and public dataset. The presented column level approach brings easier crosswalk localization and potentially detection of obstacles on the crosswalk. Another benefit is accuracy and performance analysis on two different devices with several input image resolutions which is not common in related work and could be useful for further research.

References

[1] World Health Organization, “Global data on visual impairment.” http://www.who.int/blindness/publications/globaldata/en/ (accessed Jul. 06, 2020)

[2] S. Shoval, I. Ulrich, and J. Borenstein, “NavBelt and the Guide-Cane [obstacle-avoidance systems for the blind and visually impaired],” *IEEE Robot. Autom. Mag.*, Vol. 10, No. 1, pp. 9-20, Mar. 2003, doi: 10.1109/MRA.2003.1191706

[3] V. N. Murali and J. M. Coughlan, “Smartphone-based crosswalk detection and localization for visually impaired pedestrians,” in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2013, pp. 1-7

[4] S. Wang, H. Pan, C. Zhang, and Y. Tian, “RGB-D image-based detection of stairs, pedestrian crosswalks and traffic signs,” *J. Vis. Commun. Image Represent.*, Vol. 25, No. 2, pp. 263-272, Feb. 2014, doi: 10.1016/j.jvcir.2013.11.005

[5] V. Ivanchenko, J. Coughlan, and H. Shen, “Detecting and locating crosswalks using a camera phone,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008, CVPRW’08.*, 2008, pp. 1-8

[6] K. Romić, I. Galić, and T. Galba, “Technology assisting the blind - Routing on the staircases using wide-angle camera,” in *2017 International Symposium ELMAR*, Sep. 2017, pp. 43-46, doi: 10.23919/ELMAR.2017.8124431

[7] S. Mascetti, L. Picinali, A. Gerino, D. Ahmetovic, and C. Bernareggi, “Sonification of guidance data during road crossing for people with visual impairments or blindness,” *Int. J. Hum.-Comput. Stud.*, Vol. 85, pp. 16-26, Jan. 2016, doi: 10.1016/j.ijhcs.2015.08.003

[8] L. Hakobyan, J. Lumsden, D. O’Sullivan, and H. Bartlett, “Mobile assistive technologies for the visually impaired,” *Surv. Ophthalmol.*, Vol. 58, No. 6, pp. 513-528, Nov. 2013, doi: 10.1016/j.survophthal.2012.10.004

[9] J. Zhang, C. W. Lip, S. K. Ong, and A. Nee, “A multiple sensor-based shoe-mounted user interface designed for navigation systems for the visually impaired,” in *The 5th Annual ICST Wireless Internet Conference (WICON)*, Apr. 2010, pp. 1-8

[10] P. Strumillo, “Electronic Systems Aiding Spatial Orientation and Mobility of the Visually Impaired,” in *Human – Computer Systems Interaction: Backgrounds and Applications 2*, Z. S. Hippe, J. L. Kulikowski, and T. Mroczek, Eds. Springer Berlin Heidelberg, 2012, pp. 373-386

[11] C. Yi, Y. Tian, and A. Arditi, “Portable Camera-Based Assistive Text and Product Label Reading From Hand-Held Objects for Blind Persons,” *IEEEASME Trans. Mechatron.*, Vol. 19, No. 3, pp. 808-817, Jun. 2014, doi: 10.1109/TMECH.2013.2261083

[12] K. Romic, I. Galic, and T. Galba, “Technology assisting the blind - Video processing based staircase detection,” in *ELMAR 2015 57th International Symposium*, Zadar, Sep. 2015, pp. 221-224, doi: 10.1109/ELMAR.2015.7334533

[13] S. Se, “Zebra-crossing detection for the partially sighted,” in *IEEE Conference on Computer Vision and Pattern Recognition, 2000, Proceedings*, 2000, Vol. 2, pp. 211-217, doi: 10.1109/CVPR.2000.854787

[14] M. Radvanyi, B. Varga, and K. Karacs, “Advanced crosswalk detection for the Bionic Eyeglass,” in *12th International Workshop on Cellular Nanoscale Networks and Their Applications (CNNA)*, 2010, pp. 1-5

[15] M. S. Uddin and T. Shioyama, “Robust zebra-crossing detection using bipolarity and projective invariant.,” in *ISSPA*, 2005, pp. 571-574

[16] H. Shen, K.-Y. Chan, J. Coughlan, and J. Brabyn, “A mobile phone system to find crosswalks for visually impaired pedestrians,” *Technol. Disabil.*, Vol. 20, No. 3, p. 217, 2008

[17] J. M. Coughlan and H. Shen, “A fast algorithm for finding crosswalks using figure-ground segmentation,” presented at the 2nd Workshop on Applications of Computer Vision, in conjunction with ECCV, 2006

[18] R. Cheng *et al.*, “Crosswalk navigation for people with visual impairments on a wearable device,” *J. Electron. Imaging*, Vol. 26, No. 5, Oct. 2017, doi: 10.1117/1.JEI.26.5.053025

[19] J. Jakob and T. József, “Camera-based On-Road Detections for the Visually Impaired,” *Acta Polytech. Hung.*, Vol. 17, pp. 125-146, Jan. 2020, doi: 10.12700/APH.17.3.2020.3.7

[20] J. M. Coughlan and H. Shen, “Crosswatch: a System for Providing Guidance to Visually Impaired Travelers at Traffic Intersections,” *J. Assist. Technol.*, Vol. 7, No. 2, Apr. 2013, doi: 10.1108/17549451311328808

[21] S. Mascetti, D. Ahmetovic, A. Gerino, and C. Bernareggi, “ZebraRecognizer: Pedestrian crossing recognition for people with visual impairment or blindness,” *Pattern Recognit.*, Vol. 60, pp. 405-419, Dec. 2016, doi: 10.1016/j.patcog.2016.05.002

[22] V. Tümen and B. Ergen, “Intersections and crosswalk detection using deep learning and image processing techniques,” *Phys. Stat. Mech. Its Appl.*, Vol. 543, Apr. 2020, doi: 10.1016/j.physa.2019.123510

[23] M. Poggi, L. Nanni, and S. Mattoccia, “Crosswalk Recognition Through Point-Cloud Processing and Deep-Learning Suited to a Wearable Mobility Aid for the Visually Impaired,” in *New Trends in Image Analysis and Processing -- ICIAP 2015 Workshops*, Sep. 2015, pp. 282-289, doi: 10.1007/978-3-319-23222-5_35

[24] S. Yu, H. Lee, and J. Kim, “LYTNet: A Convolutional Neural Network for Real-Time Pedestrian Traffic Lights and Zebra Crossing Recognition for the Visually Impaired,” in *Computer Analysis of Images and Patterns*, Cham, 2019, pp. 259-270, doi: 10.1007/978-3-030-29888-3_21

[25] A. Gijsenij and T. Gevers, “Color Constancy Using Natural Image Statistics and Scene Semantics,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 33, No. 4, pp. 687-698, Apr. 2011, doi: 10.1109/TPAMI.2010.93

[26] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Pearson Education, 2011

[27] K. Romić, I. Galić, H. Leventic, and K. Nenadić, “Real-time Multiresolution Crosswalk Detection with Walk Light Recognition for the Blind,” *Adv. Electr. Comput. Eng.*, Vol. 18, No. 1, pp. 11-20, Feb. 2018, doi: 10.4316/AECE.2018.01002

[28] Ruiqi Cheng, Zhejiang University, Hangzhou, China, “Pedestrian Crosswalks Recognition (PCR) Public Database.” Zhejiang University, Hangzhou, China, Accessed: Mar. 20, 2020 [Online] Available: http://www.wangkaiwei.org/project.html

[29] Information Resources Management Association, *Image Processing: Concepts, Methodologies, Tools, and Applications*, 1 edition. Hershey, PA: IGI Global, 2013

[30] R. C. Browning, E. A. Baker, J. A. Herron, and R. Kram, “Effects of obesity and sex on the energetic cost and preferred speed of walking,” *J. Appl. Physiol.*, Vol. 100, No. 2, pp. 390-398, Feb. 2006, doi: 10.1152/japplphysiol.00767.2005

Separation of Several Illnesses Using Correlation Structures with Convolutional Neural Networks

Attila Zoltán Jenei, Gábor Kiss, Miklós Gábriel Tulics, Dávid Sztahó

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar tudósok krt. 2, 1117 Budapest, Hungary,
e-mails: jenei@tmit.bme.hu, kiss.gabor@vik.bme.hu, tulics.miklos@vik.bme.hu, sztaho.david@vik.bme.hu

Abstract: There is already a lot of research in the literature on the binary separation of healthy people and people with some illnesses that affects speech. However, there are only a few examinations where more illnesses are recognized together. The examination of the latter is justified by the fact that a person may suffer from several illnesses at the same time to a certain extent. In the present study, multiclass classification of depression, Parkinson's disease, and general voice disorders (organic and functional dysphonia) was performed using speech samples. Foremost, several acoustic features were examined as input (such as Mel-Frequency Cepstral Coefficients (MFCCs), mel-band energy values, formants and their bandwidths). Using the inputs, auto- and cross-correlation structures were formed as image representations and fed to a convolutional neural network (CNN). Parameter optimization of the correlation structures and the CNN model was applied to achieve the highest accuracy. Moreover, the result of the tuned process was compared to the result of a baseline process. Finally, multiclass (5 and 4 classes) classification was performed with the best parameters. The prominent feature set was the MFCCs (55.9% accuracy, 52.2% macro F-score) for 5 class classification. 64.3% accuracy and 60.0% macro F1-score was obtained for 5 classes after parameter optimization. For classifying 4 classes (merging dysphonic ones together), 74.9% accuracy and 71.7% macro F1-score was achieved.

Keywords: depression; voice disorders; Parkinson's disease; speech; Convolutional Neural Network

1 Introduction

Several illnesses have effect on speech production making speech a very important biomarker. There is much research in the literature on recognizing a disease while comparing samples with healthy control. However, some illnesses can occur at the same time, for example, Parkinson's disease is often accompanied by depression. In this paper, we demonstrate multiclass classification process separating illnesses such as depression, Parkinson's disease, and dysphonia, supplemented with healthy class.

Depression is one of the most common psychiatric illnesses, affecting more than 300 million people worldwide. Nearly 800,000 people commit suicide each year according to World Health Organization (WHO) due to depression [1]. Possible triggers of depression can be stressful or negative life events, physiological disorders, social problems [2]. Early detection of the disease is not always clear, as its symptoms vary widely from individual to individual [3]. The diagnostic process of depression is further complicated by the fact that the person can be completely isolated from society [4].

Parkinson's disease is a neurological degenerative disease that mainly occurs in the elderly. The source of this illness is the death of dopamine-producing cells in the brain. Typical symptoms are resting tremor, muscle rigidity, instability, bradykinesia. The disease also affects the vocal cords and the muscles of the face, thus appearing during speech production [5]. The importance of its early diagnosis is given by the fact that it is currently an incurable disease, the progression, and symptoms of which can only be alleviated [6].

Dysphonia (the auditory-perceptual symptoms of voice disorders) is a disease that occurs regardless of age and gender causing changes in speech quality. It is observed with an increased frequency in people who use their voice heavily, such as singers and teachers [7]. It directly affects the patient's quality of life, which can also bring about isolation from society, triggers depression, anxiety. It can also present as an accompanying symptom of tumours, which can be fatal if not properly diagnosed and treated [8]. Dysphonia is classified as either an organic or a functional dysphonia, where organic dysphonia results from some sort of physiological change in one of the subsystems of speech, while the latter refers to a voice problem in the absence of a physical condition.

In the conference article [9], the three disease classes – mentioned above – were included in addition to the healthy control class. Approximately 270 features were calculated per recording, including voice quality measures (e.g., jitter, shimmer), pitch and intensity related measures, spectral indicators (e.g., formant frequencies, MFCCs), prosodic features (e.g. Pairwise Variability Indices [PVI]), energy metrics (e.g. Soft Phonation Index [SPI]). Parkinson's disease, depression, and general voice disorders were classified with 10 fold cross-validation among the healthy class. As a result, the accuracy ranged from 71.7% to 86.6%. The former result was achieved by the k-nearest neighbours (k-NN) classifier, the latter was accomplished with support vector machine (SVM) with radial basis function. In addition, when feature selection was used, the accuracy of the SVM with radial basis function improved from 86.6% to 88%.

In a previous research, we have already examined the recognition of these three disease classes (depression, Parkinson's disease, and dysphonia) using auto and cross-correlation structures from a limited set of acoustic-phonetic features [10]. The correlation structure was created following the work of Williamson et al., who have already successfully applied this solution in several researches [11-12]. The eigenvalues of the structures were used as input in the classification process created in RapidMiner Studio. k-NN and SVM algorithms were executed with 10 fold cross-validation. 78% accuracy was achieved using formants frequencies, MFCCs, mel-energy values, and fundamental frequency together.

Correlation structures have been already used for feature selection (Parkinson's disease, dysphonia) and recognition (dysphonia) by examining the sum of the upper triangular of the correlation matrix structure [13-14]. However, the correlation matrices as images for CNN have not been studied yet.

Numerous publications have been already published in the literature reporting binary classifications for the disease classes presented here. In these, high classification accuracy has been achieved (above 85%).

In a previous publication, the automatic separation of depression and healthy control was performed with 83%-86% accuracy with SVM. In this research audio recordings from 48 depressed subjects were used [15]. In a Chinese study, depression was detected with 82% accuracy using male speech samples with a regression procedure (for females the accuracy was 75%) [16].

In the case of Parkinson's disease, higher accuracy values (around 90%) can be found with both the sustained vowels and continuous speech in the literature [17-20]. However, small Parkinson's databases were usually used. According to the mPower research, 5.826 participants were tested by their sustained "a" sound, 86% accuracy was achieved [21].

Features like jitter, shimmer, MFCCs, and formant frequencies were the most commonly used acoustic features in recognizing dysphonia [22-23]. Both sustained vowels and continuous speech were examined. High accuracy (above 90%) also can be achieved to recognize dysphonia [24].

In this work, correlation structures were also created from certain features, but were used as an image representation and were fed into a CNN for classification. The application of correlation structures as an image on convolutional networks is novel, such a process has not been studied in these disease classes yet. Respectively, there is less research in the literature for examining these three illness groups simultaneously. However, such an investigation is justified by the fact that these three groups of illnesses may even be present simultaneously in a person's speech [25-27]. Furthermore, these illnesses are rarely suspected in the early stages. Such a device may help point out any of them by using a speech sample at the general practitioner.

In this study, a baseline CNN model was created first and a 5-class classification (depression, Parkinson’s disease, organic-, functional dysphonia, and healthy) was performed on it using several features. Secondly, parameter selection was done in the correlation structure and the CNN model using a specific group of features. Finally, 4-class classification was executed with the tuned correlation structure and model.

The content of the article follows the next structure: In Section 2, the speech databases are presented. The process and methods are described in Section 3. In Section 4, the result of multiclass classification and parameter tuning is summarized. In Section 5, a conclusion is drawn from the research and the results.

2 Speech Databases

The database contained speech samples of the three illnesses (Parkinson's disease, depression, voice disorders) and healthy recordings as a control class. Prior to each recording, the patients (and the control subjects) signed an informed consent in which they agreed to use their voice recordings for research purposes.

Each subject read out loud “The North Wind and the Sun” in Hungarian language, a text that is often used in speech technology research. This resulted in about a one-minute-long recording for each subject. The database contained recordings in which the subjects did not have any other illnesses (other than Parkinson's disease, depression, voice disorders) that could have affected his or her speech. The presence of one disease (exclusion of other diseases) is certified by the doctor treating the patient. Audio materials were recorded at a sampling frequency of 44.1 kHz with a clip-on microphone in a quiet room. The recordings were stored in 16 bits in PCM format.

2.1 Depressed Speech Database (DE)

Several versions of BDI (Beck Depression Inventory) questionnaire were created. The latest version of which was published in 1996, named BDI-II was used in this research [28]. This version consists of 21 questions (0 to 3 score for each question). Speech recordings from people suffering from depression were approximately evenly distributed among the depression severity categories defined by the BDI-II (Beck Depression Inventory-II) as mild depression (score: 14-19), moderate depression (score: 20-28), and major depression (score: 29-63). Below the score of 14, the patient is considered healthy.

Speech samples from individuals suffering from depression were collected from the Psychiatric and Psychotherapy Clinic of Semmelweis University, Budapest.

A total of 91 speech samples were used from the Depressed Speech Database: 58 female subjects (mean BDI score: 27.6 (±9.3); mean age: 37.5 (±16.7) and 20 male subjects (mean BDI score: 26.6 (±8.6); mean age: 40.6 (±15.9)).

2.2 Voice Disorder Speech Database (UD)

Speech samples were recorded from patients diagnosed with different voice disorder by the Outpatients' Clinic of the Head and Neck Surgery Department of the National Institute of Oncology, Budapest.

The voice disorder database included patients' voice suffering from disorders such as functional dysphonia, recurrent paresis, tumours at the vocal tract, cysts, tract stenosis, vocal node, laryngitis, laryngeal paralysis, spasmodic dysphonia. Overall, these were divided into two major groups: organic dysphonia (OD) and functional dysphonia (FD). Together, OD and FD form the UD database.

The RBH (Roughness, Breathiness, Hoarseness) scale describes the severity of voice disorders that is widely used in Hungary [29]. The scale scores the roughness, breathiness, and hoarseness of the voice with integers between 0 and 3. The integer 3 is the most severe category. The severity of dysphonia was determined by the clinician who made the diagnosis during the consultations.

167 recordings (74 men and 93 women) were used from OD, while 68 (20 men, 48 women) were used from FD. Their mean ages were 51.6 (OD) and 55.8 (FD) years, respectively, and their standard deviations were 14.4 (OD) and 16.1 years (FD).

The hoarseness (H) value was used to describe the severity of the voice disorder. The mean hoarseness of functional dysphonia (FD) was 1.5 and the standard deviation was 0.7 for male subjects. For women, the mean was 1.3 and the standard deviation was 0.6. For organic dysphonia (OD), the mean of H was 2.1 and the standard deviation was 0.9 for male patients. For women, the values were 1.8 and 0.8, respectively.

2.3 Parkinson's Disease Speech database (PD)

Audio recordings of patients diagnosed with Parkinson's disease (PD) were collected from two locations in Budapest: Semmelweis University (25 recordings) and Virányos Clinic (55 recordings).

H&Y (Hoehn and Yahr) scale was used to describe the severity of the disease, which ranges from 1 to 5 [30]. The 5 indicates the most severe condition, while a 1 indicates mild symptoms. Furthermore, the scale is non-linear, from which it follows that H&Y 2 does not present twice as severe symptoms as H&Y 1.

80 speech samples were collected from patients with PD: 43 males (mean H&Y score: 2.7 (±1.2); mean age: 62.6 (±13.5)) and 37 females (mean H&Y score: 2.6 (±1.2); mean age: 65.2 (±9.2)).

2.4 Healthy Control Database (HC)

In addition, a database of healthy people's recordings was also created as a control group (HC). According to their own statement, healthy individuals did not have any illnesses (and have not been diagnosed with any known illnesses) that would affect their speech at the time of recording.

140 healthy speech samples were recorded: 85 female speakers (mean age: 49.6 (±15.2)) and 55 male speakers (mean age: 51.4 (±21.6)).

3 Methods

The examination process is illustrated in Figure 1. Firstly, acoustic features were obtained from the speech recordings. From these, auto and cross-correlation structures were generated. Finally, classification with the help of CNN was executed.

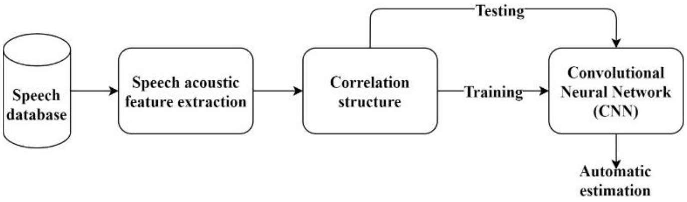


Figure 1
Outline of the applied process. (Speech database, feature extraction, correlation structure, training / testing on convolutional network)

The two-dimensional correlation matrices were input into a 2D convolutional neural network. After training the model, testing was done for automatic estimation.

Five feature sets were examined with a baseline process. From here, one set of features is selected for further studies. Furthermore, parameter optimization was performed on the correlation structure as well as on the machine learning model. In the tuned process, multiclass classification (with 5, 4 classes) was finally executed.

3.1 Acoustic Features Extraction

Before calculating the acoustic features, the speech samples were normalized to the peak. Then, acoustic features were calculated in a 50 ms Hamming window with the time step 10 ms using Praat software [31]. With this technique, a time series (later on referred as a vector) can be assigned to each feature per recording.

Then, the following speech acoustic features were obtained [32-33]:

Mel-Band Energy Values: The frequency range of the speech can be converted to a mel scale, from which mel-bands can be derived. The energy spectrum of speech can be passed through on these mel-bands, which resulted in cumulative energy values. The first 27 mel-band energy value was calculated from 100 Hz. Further on it is referred to as Melfilters.

Mel Frequency Cepstral Coefficient: This is determined from the power spectrum by summing the energy values within a defined mel-bands. Then, the discrete cosine transform of its logarithm value is calculated. The values of the first 14 coefficients were determined. These are hereinafter referred to as MFCCs.

Formant frequencies: The maximum amplitude's locations of the spectral envelope curves of the overtone beams amplified by human resonator cavities are called formant frequencies. The first three formant frequencies were calculated, which are hereinafter referred to as Formants.

Bandwidth of formant frequency: Bandwidth means the frequency range measured at a decrease of 3 dB from the amplitude peak of the formant frequency. The bandwidths of the 1st, 2nd and 3rd formant frequencies were calculated, which are hereinafter referred to as Bandwidths.

Finally, formant frequencies and their bandwidths were also used in a combination as a fifth set of features, referred to as Form-Band. So that set included Formants and Bandwidths vectors.

Melfilters and MFCCs were calculated from the total speech sample, while formant frequencies and their bandwidths were calculated from the voiced sections. Thus, the vectors of MFCCs had the same length as Melfilters'. The length of the formant frequency vectors and the bandwidth vectors were also the same.

Where the feature extraction program could not determine a value that data itself was removed from the vector and also deleted from the other feature vectors (in the same set) on the same index even if it was a numeric value. Thus, the feature vectors did not shift relative to each other in time.

As a summary, Table 1 contains the extracted features, the name of the feature set, and the number of vectors in a set.

Table 1
Extracted acoustic features with the Praat software

Feature	Name of set	Number of vectors
Mel-Band Energy Values	Melfilters	27
Mel Frequency Cepstral Coefficient	MFCCs	14
Formant frequencies	Formants	3
Bandwidth of formant frequency	Bandwidths	3
Combination of formant frequencies and their bandwidths	Form-Band	6

3.2. The Structure of Correlation Matrices

Instead of using one single vector, several new vectors were created by shifts along time. At each shift, the elements were displaced by a certain extent (hereinafter referred as displacement rate) so that the last elements were placed at the beginning of the vector. At each shift, a new vector is produced. A general approach is shown in Eq. (1), where X_0 is the original feature vector, $x_1, x_2, \dots, x_m$ are its vector components (features from time to time). X_1 is a new vector with one element (displacement rate is 1) shift. X_i is the i^{th} new vector after i element (displacement rate is 1) shift.

$$\begin{aligned} X_0 &= [x_1, x_2, \dots, x_m] \\ X_1 &= [x_m, x_1, \dots, x_{m-1}] \\ X_i &= [x_{m-(i-1)}, x_{m-(i-2)}, \dots, x_m, x_1, \dots, x_{m-i}] \end{aligned} \tag{1}$$

Pearson's correlation coefficient was used to describe the linear relationship of two feature vectors [34]. Calculating this correlation coefficient between the two original and their shifted feature vectors, a matrix can be filled.

Denoting (k - 1) as the number of shifts, a submatrix of size $k \times k$ can be created using two feature vectors and their shifted variants from one set. This is shown on the right side of Figure 2. The rows stand for the first while the columns stand for the second feature vector. The first row and column indicate the original vectors whilst the other rows and columns represent the shifted vectors. The cells of the matrix contain the correlation coefficients of the two specific feature vectors. For example, the cell of the 3rd row and 2nd column (marked by a blue rectangle in Figure 2) includes the correlation coefficient of the two times-shifted Vector 2 and 1 time shifted Vector 1, respectively. For instance, Vector 1 can be the first and Vector 2 can be the second formant frequency vector.

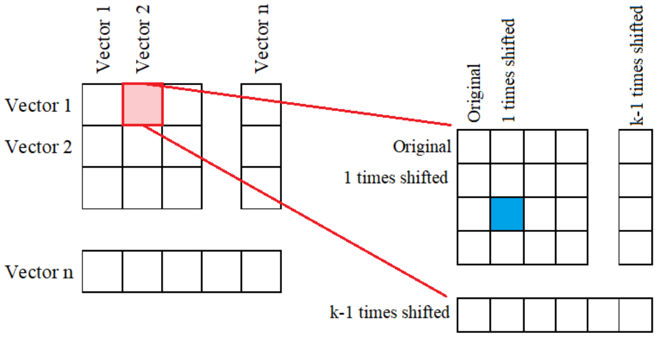


Figure 2

Structure of the correlation matrix: on the right side the submatrix of two feature vectors (size: $k \times k$). On the left there is the complete structure using a feature set with n vectors (size:$(k \times n)(k \times n)$).

One set of features was used up at once to create a correlation structure. Denoting the number of vectors in a set with n, the total size of the correlation structure is $(k \times n)(k \times n)$.

The total correlation structure is shown on the left side of Figure 2. The constructed structure is symmetrical, with autocorrelation coefficients in the main diagonals and cross-correlation coefficients in the sub-diagonals.

9 times shift (k = 10) with displacement rate 1 were set to create a baseline process. These baseline parameters (for the correlation structures and CNN) were successfully applied in preliminary research [35]. Later on, 4 times (k = 5) and 14 times shift (k = 15) were also examined with the displacement rate 1, 4, and 8 as parameter tuning.

One correlation structure was constructed for each person from each feature set (Overall, 5 correlation structures were available for each person). These as image representations were the input to the classification algorithm.

3.3 Construction of CNN Model

A simple CNN was created in Python (version 3.7.0) using Tensorflow (version 1.12.0). The baseline parameters are based on preliminary research [35].

A sequential CNN model was created with two convolutional layers, followed by a maxpooling, a flatten and a dense layer. The arrangement of the CNN layers is shown in Figure 3.

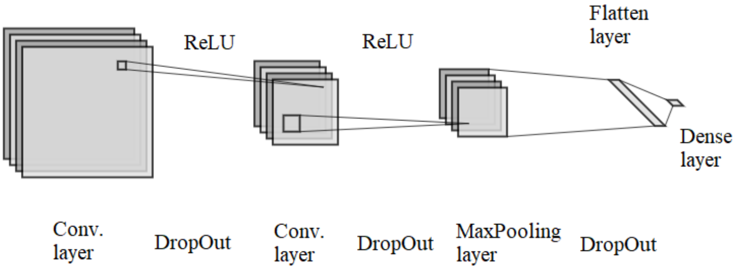


Figure 3

The structure of the CNN model: two Convolutional layers, one MaxPooling, Flatten and Dense layer

Correlation structures with the size of $(k \times n)(k \times n)$ described above were used as input to the CNN. 32 kernels were used in the first Convolutional layer. Kernel size $k \times k$ and stride k have been set according to the size of the submatrices of the input images.

The kernel size and stride of the second Convolution layer were chosen so that 2 × 2 size matrices were at the output of the layer. The size of the stride was equal to any dimension of the square kernel. 32 kernels were also used here.

The pool size of the MaxPooling layer was 2 × 2 as default.

ReLU (Rectified Linear Unit) activation functions were set up after the first two Convolutional layers and 25% DropOut regulations after each of the first 3 layers.

Finally, the output values of the Dense layer were converted into probability values with the SoftMax function. A vector with x components resulted for each test subject, where x denotes the number of classification categories. The predicted class is the one having the highest probability output.

ADAM optimization was applied during training sessions. Herewith, the automatic adjustment of the learning rate is realized taking into account the cost function. The cost function used here was the categorical cross-entropy [36].

For pre-processing, the imported data was shuffled wherein one correlation structure belonged to one subject. Normalization between 0 and 1 was also done on the elements of the matrices.

3.4 Tests and Evaluation Methods

With the created process (feature extraction from speech recordings, the built-up of the correlation structures, creating the CNN model), the following examinations were performed. The first two tests were executed with 5 classes: DE, PD, FD, OD, HC. Then the last test was performed with the database where FD and OD were merged to UD.

Leave-one-out cross-validation (LOOCV) was used for model evaluation for all tests. During this process, one subject is selected as the test element while the remaining samples are used as the training set. This is repeated until every sample was a test element. This means the training and testing process is repeated as many times as many samples are in the database. Moreover, the testing and training set were always disjoint.

For evaluation, confusion tables were created from the output of the CNN models. Metrics such as recall, precision, accuracy, and F1-score were derived.

a) Examination of feature sets: the 5 feature sets were tested separately in the baseline model. This gave sequential results on which features most appropriate for separation using this certain process.

b) Parameter optimization: In this test, the baseline process was tuned. Specifically, the number of shifts and the displacement rate were changed in the correlation structure. By changing the displacement rate, the parameters of the neural network model were not changed. However, by changing the number of shifts, the kernel size and strides of the first convolution layer were adjusted as shown in Table 2. Finally, 4 time (k = 5), 9 time (k = 10) and 14 time (k = 15) shifts were examined. The displacement rate was set as 1, 4, 8. The number of kernels remained 32 for both Convolutional layers in this case.

Table 2
The kernel size and stride of the first Convolutional layer based on the number of shifts

Number of shift	***k***	Kernel size	Kernel stride
4	5	5 × 5	5
9	10	10 × 10	10
14	15	15 × 15	15

After choosing the right parameters for the correlation structures, the CNN parameter settings followed. The number of iterations during training and the number of kernels were changed in the new CNN model. The number of iterations was set to 25, 50, 75, 100, 125, 150, and the number of kernels was set to 16, 32, 64, 128. The kernel numbers were chosen so that the first convolution layer had half the kernel number as had the second convolution layer. Thus, kernel numbers 16/32, 32/64, and 64/128 were used, where the first number is the kernel number of the first convolution layer and the second is the kernel number of the second convolution layer.

The parameter optimization was done before the training cycle and then it was tested by all the subjects separately. With this in mind, the separation of a third independent set was not necessary as the test set was already independent for the models.

c) 4 classes classification: By combining organic (OD) and functional (FD) dysphonia, the general voice disorders (UD) group was created to investigate the 4-class classification in the optimized process.

4 Results

4.1 Examination of the 5 Feature Sets

The 5 feature sets were examined in the baseline process (9 times shift, 1 displacement rate, 32 kernels, 100 iterations) based on accuracy and F1-score. All subject was used from the 5 classes. Results are shown in Figure 4.

Accuracy values ranged from 43% to 56% for all feature sets, while macro F1-score values ranged from 36% to 52%. The highest accuracy and macro F1-value were achieved with the MFCCs feature set (55.9% accuracy, 52.2% macro F1-value). Melfilters achieved the second-best accuracy (51.1%) and macro F1-score (47.4%).

Using the formants and their bandwidths separately, we obtained an accuracy of 43.0% (formants) and 45.5% (bandwidths). While using them together, the results improved (52.0% accuracy, 44.3% macro F1-score).

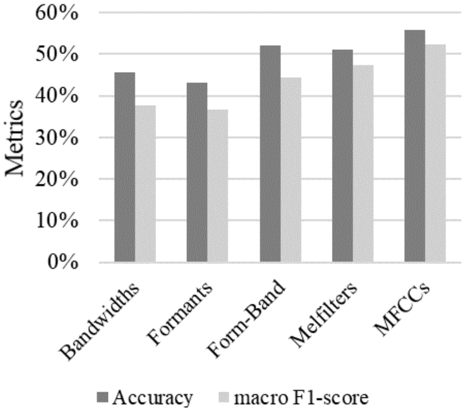


Figure 4
Achieved accuracy and macro F1-score with different feature sets on the baseline process using 5 classes

Table 3 shows the confusion matrix created from the feature set that achieved the highest accuracy (MFCCs). The columns are the original classes and the rows are the classifier's decisions. Precision and recall by classes are also noted.

The recall of the DE class was low (38.5% recall), while the precision was high (64.8% precision) compared to the other classes. The recall of the HC class was 73.6%. On the other hand, many samples from originally positive classes were classified as healthy, resulting in a 53.1% precision.

It can also be seen that subjects with functional dysphonia tended to be classified as organic dysphonia or healthy (10.3% recall, 50% precision).

Table 3
The confusion matrix derived from the MFCCs feature set. The columns represent the original classes the rows represent the decision of the algorithm.

		Original classes					
		HC	DE	FD	OD	PD	Precision
Predicted classes	HC	103	26	26	30	9	53.1%
	DE	8	35	1	3	7	64.8%
	FD	2	0	7	5	0	50.0%
	OD	18	18	30	118	22	57.3%
	PD	9	12	4	11	42	53.8%
	Recall	73.6%	38.5%	10.3%	70.3%	52.5%	

4.2 Parameter Optimization

MFCCs were selected to adjust the correlation structure and parameters of CNN to achieve a better separation of classes.

The results obtained by changing the number of shifts (k) and displacement rate are shown in Figure 5. The accuracy is given on the left diagram, the macro F1-score is given on the right diagram. The displacement rates are on the category axis while the shades of the bars indicate the number of shifts.

According to Figure 5, it is worthwhile to use a correlation structure with a higher displacement rate. However, further changes were not experienced with the displacement rate of 8. Changing the number of shifts is significant at a low displacement rate. While at a higher displacement rate, changing the number of shifts will only cause small changes in the metric values.

The highest, 61.7% accuracy was achieved at the displacement rate 4 with 15 shifts. Using macro F1-score, 55.5% is reached as the peak at the displacement rate 8 with 10 shifts.

The displacement rate 8 and 4 times shift were selected for further examination because at this displacement rate, all three shift numbers gave similar results. Nonetheless, the CNN parameters increased polynomial by linearly increasing the number of shifts.

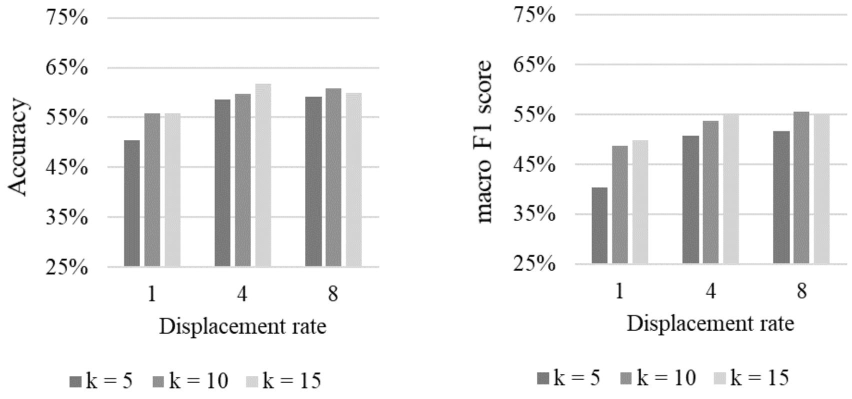


Figure 5

Results obtained by varying the number of shifts and displacement rates in the correlation structure. The left side chart shows the accuracy, while the right side chart shows the macro F1-score.

In the case of the CNN model, several parameters can be set, from which the number of iterations during the training and the number of kernels of two convolutional layers were selected for analysis.

The results obtained using different iteration numbers are shown in Figure 6. The horizontal axis shows the number of iterations, the vertical axis the percentage of accuracy.

The mean (black line) and standard deviation (grey band) of the accuracy was calculated and plotted during the training process. The accuracy of the test set was also plotted (grey curve). In the latter case, the standard deviation could not be calculated.

Based on Figure 6, the accuracy of both the test set and the training set increases. Over 125 epochs, a decrease can be observed at the test set accuracy. Thus, the iteration number in the CNN model was set from 100 to 125 in the following experiments.

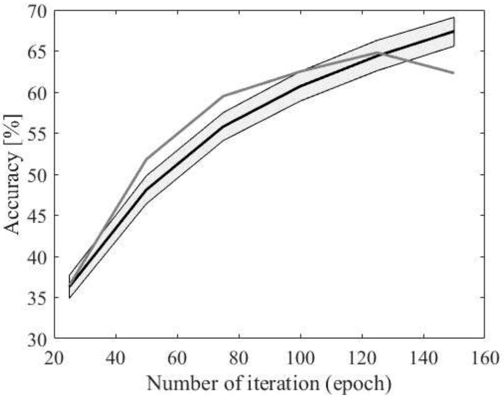


Figure 6

The accuracy of the training (black line with grey band) and test set (grey curve) as a function of the epoch number for 5 classes

The second parameter was the kernel number of the two convolutional layers to set. Kernel numbers were determined as the power of two so that the kernel number of the second convolutional layer was twice that of the first. Based on this, 16/32, 32/64, and 64/128 kernels were applied with 125 iterations (k = 5, displacement rate 8, MFCCs feature set).

The results are shown in Figure 7, where the kernel numbers are shown on the category axis. The vertical axis shows the percentages of the accuracy and F1-score.

The value of accuracy ranged from 61.2% to 64.3% in this examination. The maximum of 64.3% was reached by setting the 32/64 kernels. Similarly, the macro F1-score had a maximum of 60.0% at 32/64 kernels. The lowest value of the macro F1-score was 54.6% at 16/32 kernels. The precision averaged along the classes varied over a narrow range, from 60.0% to 61.8% along the category axis. Recall increased from 55.2% to 60.0% at 32/64 kernels.

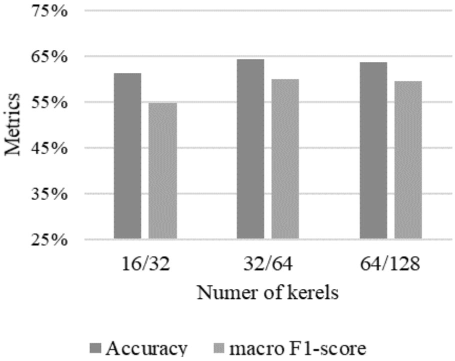


Figure 7

The classification results by choosing different kernel numbers. The first number is the kernel number of the first Convolution layer, the second is the kernel number of the second Convolution layer on the category axis.

Summarizing the results obtained by setting the two parameters of the network: the maximum of 125 iterations are worth using in the present construction. Setting 32/64 kernels brought the highest accuracy and macro F1-score in the present method. It should also be noted that this choice of kernel numbers increased the number of free parameters in the model approximately polynomially. Thus, setting these parameters can also be considered when designing such an experiment.

4.3 4 Classes Classification

Based on the results of the classification with the baseline process, organic and functional dysphonia were difficult to distinguish from each other. Therefore, these two groups were combined and examined as the general voice disorder group. This test has been done by applying the optimized parameters (iteration: 125, 32/64 kernels, $k = 5$, displacement rate 8, MFCCs feature set) on the system.

Recognition of depression was reduced by 7 samples, healthy by 4 samples, and Parkinson's disease by 1 sample in the classification of 4 classes compared to 5 classes. However, the recognition of UD was improved by 70 samples. The overall results can be seen in Table 4. Accuracy 64.3% was achieved for 5 classes and 74.9% for 4 classes on the tuned system. Macro F1-score increased from 60.0% to 71.7% by using 4 classes instead of 5 on the optimized process.

Table 4

Result of 5 and 4 classes classification with the optimized process using MFCCs

	Accuracy	macro F1-score
5 classes	64.3%	60.0%
4 classes	74.9%	71.7%

Discussion

Using the baseline process, the MFCCs feature set performed the best (55.9% accuracy, 52.2% macro F1-score). The MelFilters feature set resulted in the second-best output (47.4% accuracy, 51.1% macro F1 value). Its drop compared to the MFCCs is probably due to the fact that the 27 mel-band energy values contain everything up to 8 kHz, including signals that do not contribute to separation but are interfering.

Furthermore, the Form-Band features indicated that the combination of formant and bandwidth could improve the separation of the classification algorithm compared to applying them separately.

Increasing the number of shifts increased accuracy and macro F1-score. A possible reason for this may be that the first convolution layer may have performed better convolution from multiple samples (from a larger input context) than from a few samples.

Metrics also improved by increasing the displacement rate, but a slight decrease was already experienced at a displacement rate of 8. The decrease may be due to the disappearance of the correlation relationship between the two feature vectors in the sub-diagonals. Thus, strong correlations in the structure are limited to the main diagonal, which adversely affects the classification.

The highest test accuracy was obtained at 125 iterations, where even the results of the training and test set together progressed within the deviation band. The risk of overfitting on the training set increases at higher iterations. Also, the result of the test set has already decreased. At a low number of iterations, the accuracy of the test set changes more dynamically compared to the training set. This can presumably be caused by underfitting.

Changing the kernel numbers brought a bigger change in the macro F1-score compared to the accuracy value. Their highest performance (accuracy: 64.3%, macro F1-score: 60.0%) occurred at 32/64 kernel number. Using a higher kernel number (than 32/64) did not improve the classification, but it did increase significantly the running time of the training.

Combining OD and FD together as UD improved their correct recognition (accuracy: 74.9%, macro F1-score: 71.7%) while maintaining the optimized parameters. This is certainly influenced by the relatively large number of elements in the UD class. In contrast, the average improvement across classes is greater for 4 classes than 5 classes.

Conclusions

In the present work, the recognition of depression, Parkinson's disease, and general voice disorders were examined using a method in a new approach. In this procedure, acoustic features were calculated from speech. Component shifts were performed on the feature vectors, from which a correlation matrix was created.

These matrices were the input of a CNN model to execute the separation. First, a baseline process served the feature selection purpose from several feature sets. Secondly, parameter optimization of the correlation structures and the CNN model was also performed. Finally, 4 and 5 class classification were performed.

The advantage of this method is that it does not require more complex speech processing (such as segmentation). Furthermore, the convolutional network itself extracts the essential information from the image representations.

Also, the classification results of 4 classes can be compared to the results discussed in [10]. Higher accuracy (74.9%) was achieved here against 69.4% in [10] using only MFCCs. Unfortunately, exceeding 86.6% that had been achieved in the [9] was not successful. However, multiple features were applied there while only MFCCs were applied in the present study. Moreover, many features required segmentation in [9] while the presented method here does not need segmentation which can be a huge advantage.

Acknowledgement

Project no. K128568 has been implemented with the support provided by the National Research, Development and Innovation Fund of Hungary, financed under the K_18 funding scheme. The research was partly funded by the CELSA (CELSA/18/027) project titled: “Models of Pathological Speech for Diagnosis and Speech Recognition”.

References

[1] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators: Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017, The Lancet, 392(10159), pp. 1789-1858, 2018, https://doi.org/10.1016/S0140-6736(18)32279-7

[2] J. W. Kanter, A. M. Busch, C. E. Weeks, S. J. Landes: The nature of clinical depression: Symptoms, syndromes, and behavior analysis. The Behavior Analyst, 31(1), pp. 1-21, 2008, https://doi.org/10.1007/bf03392158

[3] G. S. Malhi, J. J. Mann: Depression, The Lancet, 392(10161), pp. 2299-2312, 2018, https://doi.org/10.1016/S0140-6736(18)31948-2

[4] L. Ge, C. W. Yap, R. Ong, B. H. Heng: Social isolation, loneliness and their relationships with depressive symptoms: A population-based study, PLoS ONE, 12(8), e0182145, 2017, https://doi.org/10.1371/journal.pone.0182145

[5] E. A. C. Pereira, T. Z. Aziz: Parkinson’s disease and primate research: Past, present, and future, Postgraduate Medical Journal, 82(967), pp. 293-299, 2006, https://doi.org/10.1136/pgmj.2005.041194

[6] L. V. Kalia, A. E. Lang: Parkinson's disease, The Lancet, 386(9996), pp. 896-912, 2015, https://doi.org/10.1016/S0140-6736(14)61393-3

[7] A. C. Gama, J. N. Santos, E. F. Pedra, A. T. Rabelo, M. C. Magalhães, E. B. Casas: Vocal dose in teachers: correlation with dysphonia, CoDAS, 28(2), pp. 190-192, 2016, https://doi.org/10.1590/2317-1782/20162015156

[8] R. J. Stachler, D. O. Francis, S. R. Schwartz, C. C. Damask and et al.: Clinical Practice Guideline: Hoarseness (Dysphonia) (Update), Otolaryngology-Head and Neck Surgery, 159(1), pp. 1-42, 2018, https://doi.org/10.1177/0194599817751030

[9] D. Sztahó, G. Kiss, M. G. Tulics, B. Hajduska-Dér, K. Vicsi: Automatic discrimination of several types of speech pathologies, in 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, pp. 1-6, 2019, https://doi.org/10.1109/SPED.2019.8906556

[10] D. Sztahó, G. Kiss, M. G. Tulics, K. Vicsi: Automatic Separation of Various Disease Types by Correlation Structure of Time Shifted Speech Features, in 2018 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, Greece, pp. 1-4, 2018, https://doi.org/10.1109/TSP.2018.8441395

[11] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, D. D. Mehta: Vocal and facial biomarkers of depression based on motor incoordination and timing, in Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 65-72, 2014, https://doi.org/10.1145/2661806.2661809

[12] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, D. D. Mehta: Vocal biomarkers of depression based on motor incoordination, in Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, pp. 41-48, 2013, https://doi.org/10.1145/2512530.2512531

[13] T. Bocklet, E. Noth, G. Stemmer, H. Ruzickova, J. Rusz: Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis, in 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, pp. 478-483, 2011, https://doi.org/10.1109/ASRU.2011.6163978

[14] T. Dubuisson, T. Dutoit, B. Gosselin, M. Remacle: On the use of the correlation between acoustic descriptors for the normal/Pathological voices discrimination, EURASIP Journal on Advances Signal Processing, 173967 (2009), 2009, https://doi.org/10.1155/2009/173967

[15] G. Kiss, K. Vicsi: Comparison of read and spontaneous speech in case of automatic detection of depression, in 8th IEEE International Conference on

Cognitive Infocommunications (CogInfoCom), Debrecen, Hungary, pp. 213-218, 2017, https://doi.org/10.1109/CogInfoCom.2017.8268245

[16] H. Jiang, B. Hu, Z. Liu, G. Wang, L. C. Zhang, X. Li, H. Kang: Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features, Computational and Mathematical Methods in Medicine, 2018, https://doi.org/10.1155/2018/6508319

[17] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, L. O. Ramig: Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson’s Disease, IEEE Transactions on Biomedical Engineering, 59(5), pp. 1264-1271, 2012, https://doi.org/10.1109/TBME.2012.2183367

[18] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene: Detecting Parkinson’s disease from sustained phonation and speech signals, PLoS ONE, 12(10), pp. 1-16, 2017, https://doi.org/10.1371/journal.pone.0185613

[19] A. Benba, A. Jilbab, A. Hammouch, S. Sandabad: Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson’s disease, in 2015 International Conference on Electrical and Information Technologies (ICEIT), Marrakeck, Morocco, pp. 300-304, 2015, https://doi.org/10.1109/EITech.2015.7163000

[20] H. Hazan, D. Hilu, L. Manevitz, L. O. Ramig, S. Sapir: Early diagnosis of Parkinson’s disease via machine learning on speech data, in 2012 IEEE 27th Convention Electrical and Electronics Engineers in Israel, Eilat, Israel, pp. 1-4, 2012, https://doi.org/10.1109/EEEI.2012.6377065

[21] B. M. Bot, C. Suver, E. C. Neto, et al: The mPower study, Parkinson disease mobile data collected using ResearchKit”, Scientific Data, 3, 2016, https://doi.org/10.1038/sdata.2016.11

[22] J. P. Teixeira, P. O. Fernandes: Acoustic Analysis of Vocal Dysphonia, Procedia Computer Science, 64, pp. 466-473, 2015, https://doi.org/10.1016/j.procs.2015.08.544

[23] H. T. Lathadevi, S. P. Guggarigoudar: Objective acoustic analysis and comparison of normal and abnormal voices, Journal Clinical and Diagnostic Research, 12(12), pp. 1-4, 2018, https://doi.org/10.7860/JCDR/2018/36782.12310

[24] J. I. Godino-Llorente, P. Gomez-Vilda: Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors, in IEEE Transactions on Biomedical Engineering, 51(2), pp. 380-384, 2004, https://doi.org/10.1109/TBME.2003.820386

[25] L. Marsh: Depression and Parkinson's disease: current knowledge, Current Neurology and Neuroscience Reports, 13, 2013, https://doi.org/10.1007/s11910-013-0409-5

[26] A. Hu, A. Hillel, W. Zhao, T. Meyer: Anxiety and depression in spasmodic dysphonia patients, World Journal of Otorhinolaryngology - Head and Neck Surgery, 4(2), pp. 110-116, 2018, https://doi.org/10.1016/j.wjorl.2018.04.004

[27] L. J. White, E. R. Hapner, A. M. Klein, et al.: Coprevalence of anxiety and depression with spasmodic dysphonia: a case-control study, Journal of Voice, 26(5), pp. 1-6, 2012, https://doi.org/10.1016/j.jvoice.2011.08.011

[28] A. T. Beck, R. A. Steer, R. Ball, W. F. Ranieri: Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients, Journal of Personality Assessment, 67(3), pp. 588-597, 2010, https://doi.org/10.1207/s15327752jpa6703_13

[29] T. Haderlein, C. Schwemmle, M. Döllinger, et al: Automatic Evaluation of Voice Quality Using Text-Based Laryngograph Measurements and Prosodic Analysis, Computational and Mathematical Methods in Medicine, 2015, pp. 1-11, 2015, https://doi.org/10.1155/2015/316325

[30] J. M. Rabey, A. D. Korczyn: The Hoehn and Yahr Rating Scale for Parkinson's Disease, in Instrumental Methods and Scoring in Extrapyramidal Disorders, Heidelberg: Springer, Berlin, Heidelberg, pp. 7-17, 1995, https://doi.org/10.1007/978-3-642-78914-4_2

[31] P. Boersma: Praat, a system for doing phonetics by computer, Glot International, 5(9/10), 341-345, 2002

[32] B. H. Story: Vowel and consonant contributions to vocal tract shape, The Journal of the Acoustical Society of America, 126(2), pp. 825-836, 2009, https://doi.org/10.1121/1.3158816

[33] J. Saini, R. Mehra: Power Spectral Density Analysis of Speech Signal using Window Techniques, International Journal of Computer Applications, 131(14), pp. 33-36, 2015, https://doi.org/10.5120/ijca2015907549

[34] M. M. Mukaka: Statistics corner: A guide to appropriate use of correlation coefficient in medical research, Malawi Medical Journal, 24(3), pp. 69-71, 2012, PMCID: PMC3576830

[35] A. Z. Jenei, G. Kiss: Possibilities of Recognizing Depression with Convolutional Networks Applied in Correlation Structure, In: 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), Milan, Italy, pp. 101-104, 2020, https://doi.org/10.1109/TSP49548.2020.9163547

[36] K. P. Murphy: Probability, in Machine Learning: A Probabilistic Perspective, (ed.) The MIT Press, Cambridge, Massachusetts, 2012

A Comprehensive Comparison between Finite Control Set Model Predictive Control and Classical Proportional-Integral Control for Grid-tied Power Electronics Devices

Abdulrahman J. Babqi[*], Basem Alamri

Department of Electrical Engineering, College of Engineering, Taif University
KSA, P. O. Box 11099, Taif 21944, Saudi Arabia, b.alamri@tu.edu.sa
*Corresponding author: ajbabqi@tu.edu.sa

Abstract: Recently finite control set model predictive control (FCS-MPC) becomes a promising solution for controlling power electronics devices (PEDs). Although FCS-MPC produces variable switching frequency and steady-state error, it has many advantages such as the ease of implementation, controlling multiple parameters at the same time, and generating the switching signals internally. This paper aims to assess the performance of the FCS-MPC by constructing a comparative study of the FCS-MPC current control with the proportional-integral (PI) current control. For a fair comparison, the FCS-MPC average switching frequency was made to be equal or lower than the PI switching frequency. The study was performed on three different grid-connected PEDs, which are the three-phase two-level, single-phase full-bridge, and H5 inverters. Both control strategies were compared considering the switching frequency, common-mode voltage, leakage current, total harmonics distortion, and steady-state error. The results illustrate that the produced common-mode voltage and leakage current of the FCS-MPC are lower than PI in all cases. Even though FCS-MPC results in higher THD and steady-state error, they were maintained within acceptable limits. The three inverters and case studies were carried out to verify the performance of the controllers via the PSCAD/EMTDA software package.

Keywords: Model predictive control (MPC); Grid-connected inverters; Power electronics control; Common-mode voltage (CMV); Leakage current

1 Introduction

Renewable energy resources (RERs) become attractive alternatives over conventional power generation such as coal and natural gas since they are eco-friendly and fuel-free energy resources. However, some RERs especially wind and solar are not continuously available and their output power may vary considerably during the day [1]. The variation of renewable generation output causes some challenges to integrate them into the existing grid. For instance, the output direct current (DC) of a PV system must be converted to alternative current (AC) to be able to connect it to the grid. To this end, a controlled inverter is used to convert the DC to AC and regulate the grid's current. Such control ensures the stability and reliability of the PV power integration to the gird. In other words, power electronics devices (PEDs) facilitate merging these renewable energy resources with any electrical power system [2]. Nevertheless, integrating renewable energy resources via power electronics devices requires sophisticated control techniques [3]. Different control strategies have been proposed for controlling either grid-connected or islanded PEDs [4-9]. Among all, the most common and popular control strategy is the classical proportional-integral (PI) control method. The PI control uses the feedback mechanism to produce the error as the difference between the measured and reference values. The error then fed to the proportional and integral gains which produce the controlling signals. A tremendous number of literary works have proposed different PI control techniques for grid-tied PEDs [3]. A robust strategy for regulating the grid current entering a distribution network from a three-phase inverter system connected via an LCL filter was presented in [10]. The authors in [11] demonstrated different structures such as dq, stationary, and natural frame of PI control for the grid-side converter. A novel controller optimization algorithm using particle swarm optimization (PSO) in [12] for inverter output controllers.

Another widely used control strategy for PEDs is model predictive control (MPC) [13]. The goal of the MPC controller is to minimize the cost function considering the system constraints. It uses the system model to predict the step ahead of the controlled parameters. MPC applies the feedback mechanism to update the system in each time step for future disturbances. Only the first control action is implemented in each time step, and the rest is discarded. Therefore, MPC can predict the state's evolution over the prediction horizon. MPC can be classified into two groups for controlling PEDs, which are continuous control set MPC (CCS-MPC), and finite control set MPC (FCS-MPC) [14]. CCS-MPC utilizes the average model of the PED to perform the optimization process by minimizing the error between the predicted and reference values [15]. CCS-MPC generates continuous control signals and employs a modulator to generates the appropriate switching signals to the PED. Since this type of control uses an external modulator, it produces a fixed switching frequency, and that is considered the main advantage of CCS-MPC [16]. However, using CCS-MPC for controlling PEDs presents a very complex formulation of the optimization problems. This complex formulation can be reduced considerably by reformulating the optimization problem as a finite moving-horizon optimal control problem. In other words, FCS-MPC formulates the optimization problem based on the discreet nature of the PED, and that does not require an external switching signals generator. FCS-MPC evaluates each switching state of the PED and chooses the state that produces the lowest error. As a result, the computational time for solving the optimization problem is reduced significantly [17].

Many FCS-MPC algorithms have been proposed in the literature for controlling grid-connected PEDs. An early work of the authors in [18] presented the implementation of the FCS-MPC for controlling the output current of a three-phase two-level inverter. This work verified that FSC-MPC avoids the use of external modulators, and the drive signals are generated internally. Moreover, it showed that the control method effectively manages the load currents and provides a satisfactory dynamical response. The authors claim that FCS-MPC is a very powerful tool that offers new possibilities for PEDs control and it can be used for different types of PEDs. A novel model-free predictive current control was proposed in [19] for a three-phase rectifier. The method eliminates the usage of the system model, multiplication operations, and tuning parameters. The technique slightly improved the current control than the original FCS-MPC algorithm. A multi-objective FSC-MPC was proposed for controlling a grid-connected and islanded three-phase inverter in [20]. This work showed that FCS-MPC can perform multiple control actions at the same time such as current control and switching frequency reduction. The multiple control actions are included in the cost function, and a weighting factor is used to choose the priority of the controlled parameter. A reduced computational time FSC-MPC for a modular multilevel converter (MMC) was proposed in [21]. A grouping-sorting-optimized model predictive control with several modules was used for each arm of MMC current control. The method reduces the computational load of the FCS-MPC algorithm for MMC by considering a cascaded two-stage MPC. Reference [22] illustrated a proposed FCS-MPC for a five-level bidirectional converter. The authors confirmed that FCS-MPC improves the performance of the five-level converter in terms of efficiency and grid current total harmonics distortion. Using an extended state observer, [23] proposed an FCS-MPC of a three-phase inverter with a constant switching frequency. The method improves the current control and dynamic response of FCS-MPC for the three-phase inverter.

Numerous research works have compared the performance of the FCS-MPC with PI control for grid-tied PEDs [24- 29]. A comparison between FCS-MPC and PI for three-phase inverter current control was presented in [24, 25]. Both works compare the two controllers' performance in terms of step-change response, steady-state error (SSE), and total harmonics distortion (THD). The authors in [26, 27] proposed the FCS-MPC for controlling the output current of a quasi-Z-source inverter. The step-change response and THD of the FCS-MPC were compared with the PI control in [26] while the comparison was done based on the resultant THD and switching frequency (f_S) in [27]. A model predictive control with a delay compensation method was proposed for a three-phase four-leg grid-tied inverter in [28]. The method was also compared with the PI control depending on step-change response and THD. An FCS-MPC was proposed for a new grid-tied three-IGBTs inverter in [29]. The authors compared the control strategy with PI and hysteresis control methods. They formed a comparison based on the step-change response and THD. All the aforementioned research works have compared FCS-MPC with the classical PI strategy for different types of grid-connected

PEDs based on only the produces total harmonics distortion, switching frequency, or step-change response. However, to the best of our knowledge, there is no existing work that compares the performance of the FCS-MPC with PI control based on the common-mode voltage (CMV) and leakage current adding to that total harmonics distortion, switching frequency, and steady-state error. Therefore, this paper presents a comprehensive comparison of the FCS-MPC versus classical PI control for different grid-connected PEDs. The study was done on three PEDs, which are three-phase two-level, single-phase full-bridge, and H5 inverters. The two control methods' performances were compared based on:

1) the switching frequency,
2) total harmonic distortion,
3) steady-state error,
4) common-mode voltage, and
5) leakage current.

The rest of the paper is organized as follows: Section 2 explains the system modelling. Section 3 describes the principle of the finite control set model predictive control and proportional-integral control strategies. Section 4 discusses the common-mode voltage and leakage current in power electronics devices. The case studies are presented in Section 5, and Section 6 concludes the paper.

2 Systems Modeling and Description

Several power electronic devices are used to transform the direct current to sinusoidal alternative current. Examples of these devices are the single-phase full-bridge, H5, and three-phase two-level inverters (Figure 1). Since each device has a distinct topology, a different control scheme is used to minimize the error between the measured and reference values. Control strategies that are typically applied to control power electronics devices are PI and FCS-MPC. PI controller acts only when the error between measured and reference values has occurred. On the other hand, the FCS-MPC controller can predict the error before it occurs, which makes the MPC more robust compared to other controllers. This work examines the performance of both PI and FCS-MPC controllers on three distinct power electronic devices, which are the single-phase full-bridge, the H5, and the three-phase two-level inverters (Figure 1). Each device has a different configuration in terms of the number of switches and legs. The three-phase two-level inverter in Figure 1(a) consists of three legs with two switches in each leg while the full-bridge in Figure 1(b) is a single-phase inverter consists of two legs with two switches in each leg.

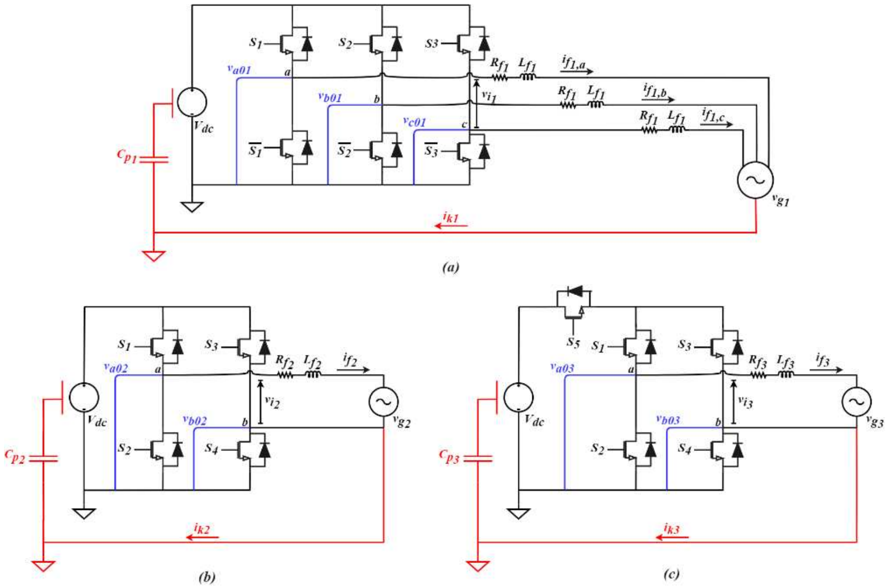


Figure 1

Power electronics devices used in this study: (a) three-phase two-level inverter, (b) single-phase full-bridge inverter, and (c) H5 inverter

The H5 inverter in Figure 1(c) is a modified version of the full-bridge inverter proposed by SMA Solar Technology [30]. In this modified version, one more switch (S_5) is added to disconnect the DC source (e.g., PV system) from the utility grid during the zero operation modes, which results in reducing the leakage current, i_k [31]. For system modeling, each inverter is connected to a DC source to represent the generator. Also, each inverter is connected to the grid through an inductor filter L_f, as shown in Figure 1.

The system modeling of each inverter (Figure 1) in case of grid-connected mode can be derived as

$$v_i = v_g + v_f + R_f i_f \tag{1}$$

where v_i is the inverter output voltage, v_f is the inductor voltage, and v_g is the grid voltage. i_f is the current flows between the inverter and utility grid. Using the filter inductor current dynamical equation,

$$v_f = L_f \frac{di_f}{dt} \tag{2}$$

the continuous-time state-space model of the systems is

$$\frac{di_f}{dt} = \frac{1}{L_f}\left(v_i - v_g - R_f i_f\right) \tag{3}$$

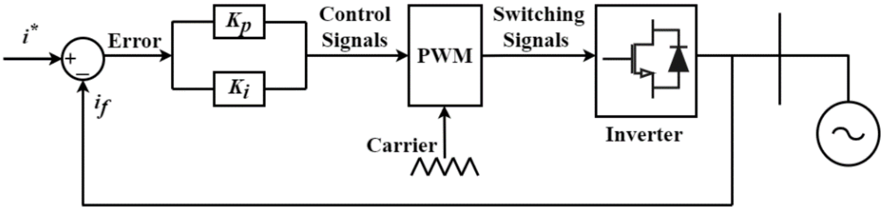


Figure 2
Proportional-integral control block diagram

Whether the inverter is a single- or three-phase, the continuous-time state-space equation can be used for modeling all power electronics devices in Figure 1.

3 Controllers Modeling

3.1 Proportional-Integral Control

The classical pulse width modulation (PWM) PI control method is a well-known technique for controlling the output current or voltage of the power electronics devices [32-34]. Figure 2 shows the basic principle of the PWM PI control for controlling the inverter output current. First, the measured current is used as a feedback in closed-loop control. Then, an error between the measured and reference values is fed to the PI controller, which produces the controlling signals. Finally, a PWM generator is used to produce the appropriate switching signals to the inverter. References [32] and [33] explain the implementation of the PI controller for both the single-phase full-bridge and H5 inverters, respectively. The PI controller using the PWM technique for the three-phase two-level inverter is presented in [34]. In this work, the control techniques in [32-34] are used to implement the PI control for all three inverters (Figure 1).

3.2 Finite Control Set Model Predictive Control

FCS-MPC is a finite moving-horizon optimal control method that uses the system model and local measurements for future values prediction of the controlled parameters. Figure 3 illustrates the FCS-MPC working principle [35]. First, the discrete-time model of the system is obtained. Then, the inverter output measurements are used with the system model to predict the controlled future values. Afterward, the error between the reference and future values is minimized using a cost function. Finally, switching signals are generated by the controller and sent to the inverter. FCS-MPC generates the switching signals internally based on space vector modulation (SVM) technique. Tables 1, 2, and 3 explain the SVM for the three-phase two-level, single-phase full-bridge, and H5 inverters, respectively.

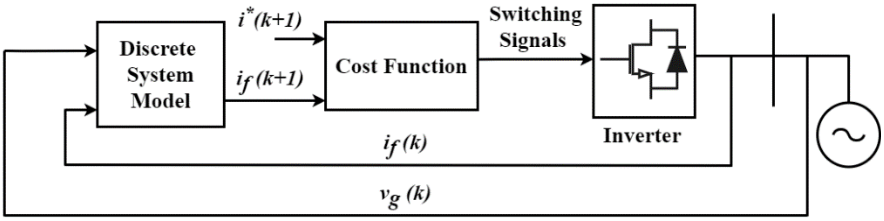


Figure 3
Finite control set model predictive control block diagram

Table 1 shows the upper switches S_1, S_2, and S_3 which control the three-phase two-level inverters while the lower once are switched conversely.

3.2.1 FCS-MPC Current Control for Three-phase Two-level Inverter

FCS-MPC for the three-phase two-level inverter is implemented in *abc* or *αβ* reference frames. Since the latter reduces the mathematical operations performed by the controller [36], FCS-MPC is implemented in the *αβ* reference frame for the three-phase two-level inverter in this work. The transformation from *abc* to *αβ* is obtained using Clarke's transformation as

$$\begin{bmatrix} x_\alpha \\ x_\beta \end{bmatrix} = \frac{2}{3} \begin{bmatrix} 1 & \frac{-1}{2} & \frac{-1}{2} \\ 0 & \frac{\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} x_a \\ x_b \\ x_c \end{bmatrix} \tag{4}$$

using (3) in (4), yields

$$\frac{di_{f_{1,\alpha}}}{dt} = \frac{1}{L_{f_1}} \left(v_{i_{1,\alpha}} - v_{g_{1,\alpha}} - R_{f_1} i_{f_{1,\alpha}} \right) \tag{5a}$$

$$\frac{di_{f_{1,\beta}}}{dt} = \frac{1}{L_{f_1}} \left(v_{i_{1,\beta}} - v_{g_{1,\beta}} - R_{f_1} i_{f_{1,\beta}} \right) \tag{5b}$$

where the notation 1 refers to the three-phase two-level inverter parameters (Figure 1). The discrete-time model of (5a) and (5b) can be obtained using the Euler forward method to approximate the derivative [37].

$$\frac{dx}{dt} \approx \frac{x(k) - x(k-1)}{T_s} \tag{6}$$

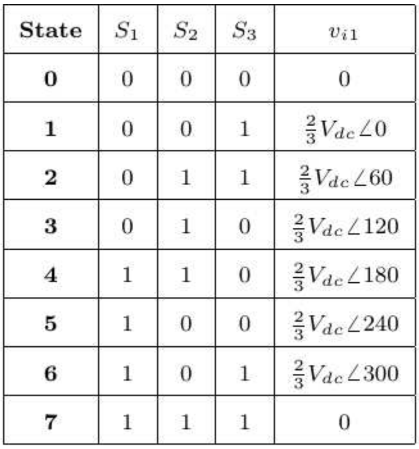
Where $x(k)$ is the present value, $x(k-1)$ is the previous value, and T_s is the sampling time. Applying (6) to (5a) and (5b), the discrete-time model of the system is

$$i_{f_{1,\alpha}}(k+1) = i_{f_{1,\alpha}}(k) + \frac{T_S}{L_{f_1}}\left(v_{i_{1,\alpha}}(k) - v_{g_{1,\alpha}}(k) - R_{f_1} i_{f_{1,\alpha}}(k)\right) \tag{7a}$$

$$i_{f_{1,\beta}}(k+1) = i_{f_{1,\beta}}(k) + \frac{T_S}{L_{f_1}}\left(v_{i_{1,\beta}}(k) - v_{g_{1,\beta}}(k) - R_{f_1} i_{f_{1,\beta}}(k)\right) \tag{7b}$$

Table 1

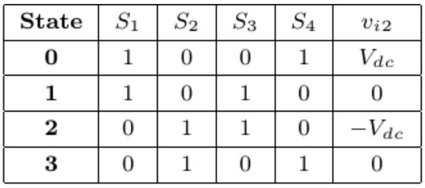
Space Vector Modulation of Three-phase Two-level Inverter



State	S_1	S_2	S_3	v_{i1}
0	0	0	0	0
1	0	0	1	$\frac{2}{3}V_{dc}\angle 0$
2	0	1	1	$\frac{2}{3}V_{dc}\angle 60$
3	0	1	0	$\frac{2}{3}V_{dc}\angle 120$
4	1	1	0	$\frac{2}{3}V_{dc}\angle 180$
5	1	0	0	$\frac{2}{3}V_{dc}\angle 240$
6	1	0	1	$\frac{2}{3}V_{dc}\angle 300$
7	1	1	1	0

Table 2

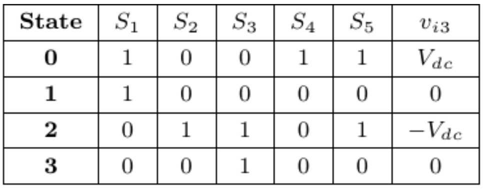
Space Vector Modulation of Single-phase Full-bridge Inverter



State	S_1	S_2	S_3	S_4	v_{i2}
0	1	0	0	1	V_{dc}
1	1	0	1	0	0
2	0	1	1	0	$-V_{dc}$
3	0	1	0	1	0

Table 3

Space Vector Modulation of H5 Inverter



State	S_1	S_2	S_3	S_4	S_5	v_{i3}
0	1	0	0	1	1	V_{dc}
1	1	0	0	0	0	0
2	0	1	1	0	1	$-V_{dc}$
3	0	0	1	0	0	0

The output current future value of the three-phase two-level inverter is predicted by (7a) and (7b). The present values measurements $v_{g_1}(k)$ and $i_{f_1}(k)$ along with the inverter voltage $v_{i_1}(k)$ are used to predict the output current future value. Since $v_{i_1}(k)$ can be one of the eight values (Table 1), that will result in eight future values for the predicted current. Therefore, the cost function (8) is used to investigate each voltage vector and select the one that produces the lowest error between the reference current ($i^*(k+1)$) and the predicted values. Once the optimal vector is selected, the related switching signals are sent to the inverter. This control procedure occurs in every sampling period T_s.

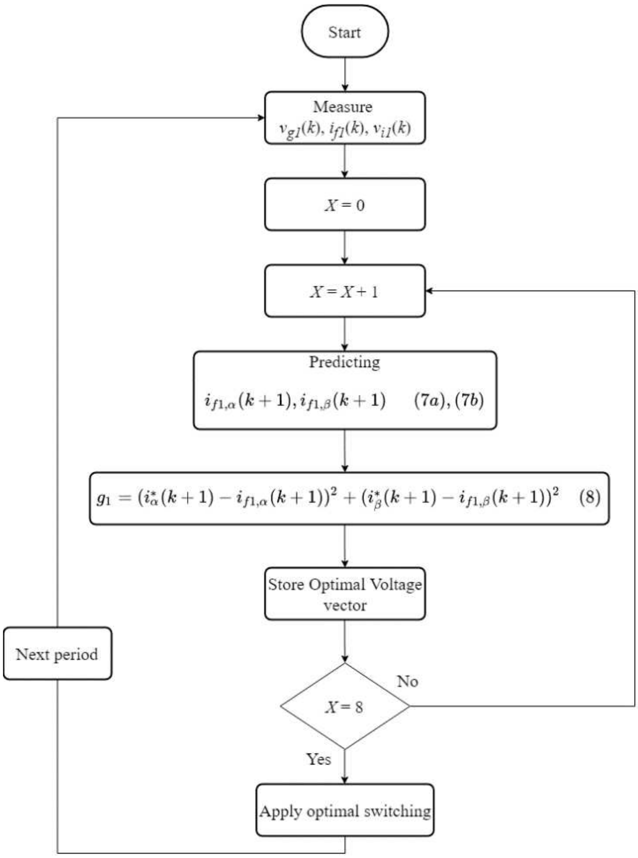

Start
Measure
$v_{g1}(k), i_{f1}(k), v_{i1}(k)$
$X = 0$
$X = X + 1$
Predicting
$i_{f1,\alpha}(k+1), i_{f1,\beta}(k+1)$ (7a), (7b)
$g_1 = (i^*_\alpha(k+1) - i_{f1,\alpha}(k+1))^2 + (i^*_\beta(k+1) - i_{f1,\beta}(k+1))^2$ (8)
Store Optimal Voltage vector
$X = 8$
No
Yes
Apply optimal switching
Next period

Figure 4
Flowchart of Finite control set model predictive for Three-phase Two-level Inverter

Figure 4 illistrates the algorithm process of the FCS-MPC for Three-phase Two-level Inverter.

$$g_1 = (i^*_\alpha(k+1) - i_{f_{1,\alpha}}(k+1))^2 + (i^*_\beta(k+1) - i_{f_{1,\beta}}(k+1))^2 \quad (8)$$

3.2.2 FCS-MPC Current Control for Single-Phase Full-Bridge and H5 Inverters

Since full-bridge and H5 are single-phase inverters, there is no need for using any transformation process. Therefore, the discrete-time model for both inverters can be directly derived from (3) using (6) which result in

$$i_{f_{2,3}}(k+1) = i_{f_{2,3}}(k) + \frac{T_s}{L_{f_{2,3}}}\left(v_{i_{2,3}}(k) - v_{g_{2,3}}(k) - R_{f_{2,3}} i_{f_{2,3}}(k)\right) \tag{9}$$

where thenotations 2 and 3 refer to full-bridge and H5 inverters parameters, respectively (Figure 1). Since the full-bridge and H5 inverters have four states as it is illustrated in Tables 2 and 3, there will be four future values of the predicted current for both inverters. Smellier to (8), the cost function (10) is used to determine the optimal voltage vector.

$$g_{2,3} = (i^*(k+1) - i_{f_{2,3}}(k+1))^2 \tag{10}$$

3.2.3 Second Step Prediction

As mentioned previously, FCS-MPC is optimal control and requires solving a large number of mathematical equations. As a result, a time-delay might occur while the controller solves the optimization problems and performs the control actions within one sampling period [38]. Therefore, a second step prediction $x(k+2)$ is preferred over the first step $x(k+1)$. To predict the second step of the controlled variables, the first step prediction should be obtained first (11a). Then, the first step prediction is used to predict the second step (11b).

$$i_f(k+1) = i_f(k) + \frac{T_s}{L_f}\left(v_i(k) - v_g(k) - R_f i_f(k)\right) \tag{11a}$$

$$i_f(k+2) = i_f(k+1) + \frac{T_s}{L_f}\left(v_i(k) - v_g(k+1) - R_f i_f(k+1)\right) \tag{11b}$$

As the grid frequency is much smaller than the sampling frequency, it can be considered [39].

$$v_g(k+1) = v_g(k) \tag{12}$$

Therefore, (11b) can be written as

$$i_f(k+2) = i_f(k+1) + \frac{T_s}{L_f}\left(v_i(k) - v_g(k) - R_f i_f(k+1)\right) \tag{13}$$

and the cost functions (8) and (10) are modified as

$$g_1 = (i_\alpha^*(k+2) - i_{f_{1,\alpha}}(k+2))^2 + (i_\beta^*(k+2) - i_{f_{1,\beta}}(k+2))^2 \tag{14}$$

$$g_{2,3} = (i^*(k+2) - i_{f_{2,3}}(k+2))^2 \tag{15}$$

4 Common-Mode Voltage

CMV is the potential between the source and load neutral point. In the case of using a grounded DC source, the common-mode voltage can cause a large leakage current i_k (Figure 1) flows between the DC source and the grid through a parasitic capacitor, which causes safety hazards and reduces the overall efficiency [31]. The common-mode voltage for the three-phase two-level inverter is calculated as in (16) while (17) is used to determine the common-mode voltage for both full-bridge and H5 inverters [40] (Figure 1).

$$v_{cm_1} = \frac{v_{ao_1} + v_{bo_1} + v_{co_1}}{3} \tag{16}$$

$$v_{cm_{2,3}} = \frac{v_{ao_{2,3}} + v_{bo_{2,3}}}{2} \tag{17}$$

5 Simulation Results and Case Studies

Three case studies were conducted to compare FCS-MPC and PI control performances on three different PEDs which are the single-phase full-bridge, H5, and three-phase two-level inverters (Figure 1). The three systems of were simulated using PSCAD/EMTDC platform. The comparison of the controllers' performance is based on five indicators:

- switching frequency (f_s),
- total harmonic distortion (THD),
- steady-state error (SSE),
- common-mode voltage (CMV), and
- leakage current (i_k).

Based on these five indicators, case study 1 investigates the performance of the two controllers on the single-phase full-bridge inverter while case studies 2 and 3 investigate the controllers' performance on H5 and three-phase two-level inverters, respectively. As it is known that the FCS-MPC average switching frequency ($f_{s,MPC}$) is not constant. Therefore, $f_{s,MPC}$ was set to be lower or equal to the PI controller switching frequency ($f_{s,PI}$) to present fair comparisons. In other words, the average switching frequency of FCS-MPC does not exceed the PI controller switching frequency at any operating point. The parameters values of the three systems are provided in Table 4. The PI switching frequency was set to 3.6 kHz. The sampling time values of FCS-MPC for three-phase, single-phase full-bridge, and H5 inverters were set to 50, 40, and 45 µsec, respectively.

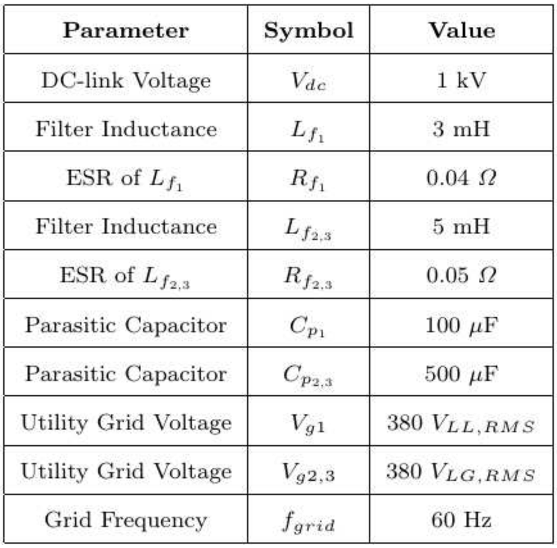
Since the grid voltage was set to 380 V, the acceptable THD at the grid side should not exceed 8% [41].

5.1 Case Study 1: Single-Phase Full-Bridge Current Control

In this case study, PWM PI and FCS-MPC control strategies were implemented to control the single-phase full-bridge inverter's output current. Figure 5 shows the two controllers' resultant switching frequency (f_s), common-mode voltage (CMV), and leakage current (i_k) to the reference current (i^*) where the horizontal axis represents i^*, and the vertical axis represents the responses of f_s, CMV, and i_k. It should be noted that the reference current value increased from 0 to 1 p.u.

Table 4

Systems Parameters



Parameter	Symbol	Value
DC-link Voltage	V_{dc}	1 kV
Filter Inductance	L_{f_1}	3 mH
ESR of L_{f_1}	R_{f_1}	0.04 Ω
Filter Inductance	$L_{f_{2,3}}$	5 mH
ESR of $L_{f_{2,3}}$	$R_{f_{2,3}}$	0.05 Ω
Parasitic Capacitor	C_{p_1}	100 μF
Parasitic Capacitor	$C_{p_{2,3}}$	500 μF
Utility Grid Voltage	V_{g1}	380 $V_{LL,RMS}$
Utility Grid Voltage	$V_{g2,3}$	380 $V_{LG,RMS}$
Grid Frequency	f_{grid}	60 Hz

The average switching frequency of the FCS-MPC was set to be similar or lower than the classic PI controller switching frequency. Figure 5 illustrates that the $f_{s,MPC}$ does not exceed $f_{s,PI}$ at all operating points, resulting in a fair comparison in the other four performance indicators, which are CMV, i_k, THD, and SSE. Since the CMV may increase the i_k flowing in the system, reducing CMV mitigates the i_k effect. In Figure 5, the FCS-MPC shows superiority over PI control in terms of CMV and i_k. As shown in Figure 5, the resultant common-mode voltage of the FCS-MPC is maintained at a lower point around 0.4 p.u. compared to the PI controller at 0.5 p.u. It is clear that FCS-MPC reduces the leakage current by more than 100% compared to the linear controller at law reference value reaching around 50% of leakage current reduction at 1 p.u. This performance improvement using FCS-MPC occurred since the controller eliminates the PI regulators and external modulators, and instead, it implements the SVM technique by considering a finite number of controlling vectors. Regarding THD, both FCS-MPC and PI show an almost similar response at high reference values while there is a very slight increase in the FCS-MPC's THD at lower operating points. This is closely related to the decrease of $f_{s,MPC}$ at low reference values. However, PI control shows a better SSE performance by almost 0.1% reduction at all operating points (Figure 6).

5.2 Case Study 2: H5 Current Control

In this case study, the system was reconfigured with H5 inverter to assess the FCS-MPC performance compared to the linear PI controller. The reference current had increased from 0 to 1 p.u. to examine the system response by assessing f_s, CMV, i_k, SSE, and THD.

It is clear that H5 inverter reduces the leakage current compared to the single-phase full-bridge (Figures 5 and 7) since H5 prevent the freewheeling current in the zero modes operation.

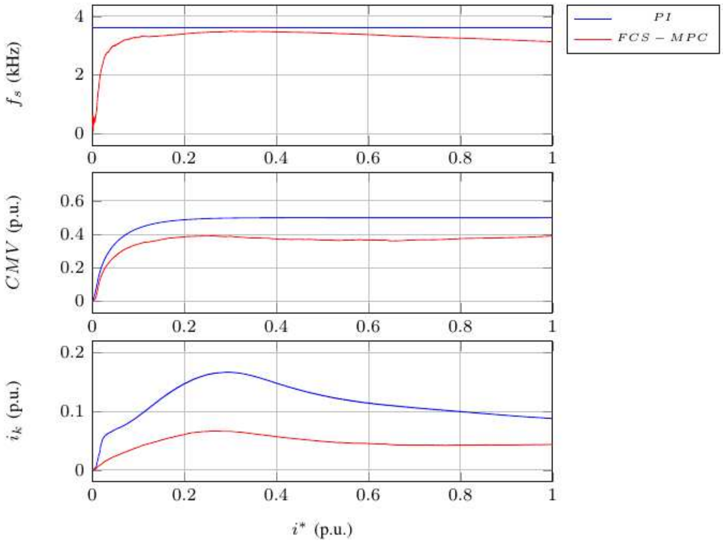


Figure 5

Single-phase full-bridge current control results: switching frequency (f_s), common-mode voltage (CMV), and leakage current (i_k)

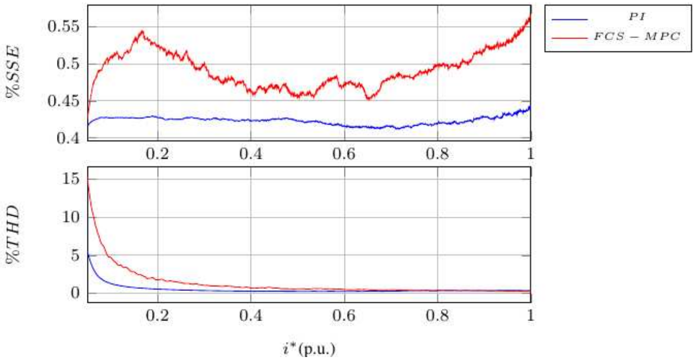


Figure 6
Single-phase full-bridge current control results: steady-state error (SSE), and total harmonics distortion (THD)

Figure 7. Shows that $f_{s,MPC}$ is lower than $f_{s,PI}$ during the whole simulation period. Figure 7 shows that the system performance is enhanced with FCS-MPC, there is a noticeable decrease in both common-mode voltage, and leakage current of the FCS-MPC compared to the classic controller. At lower reference values, FCS-MPC is capable of reducing the CMV by around 0.1 p.u. while it reduces the leakage current by 0.6 p.u. At higher operating points, both controllers result in almost similar CMV and leakage current.

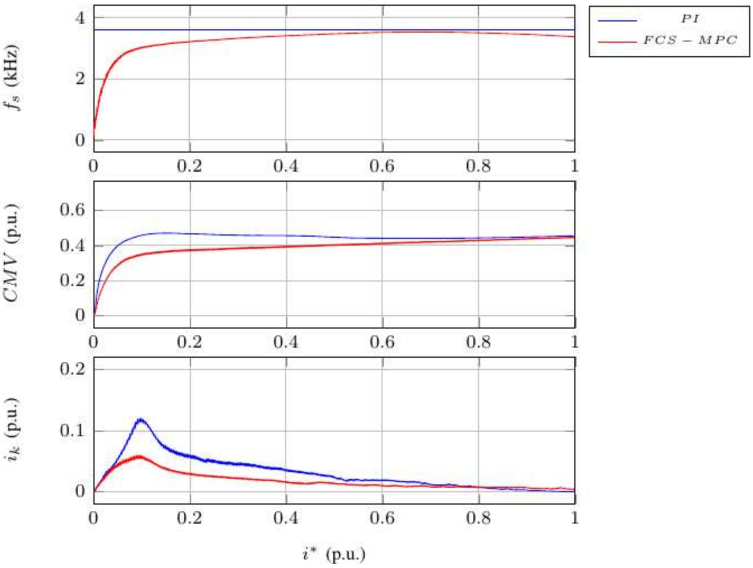


Figure 7
H5 current control results: switching frequency (f_s), common-mode voltage (CMV), and leakage current (i_k)

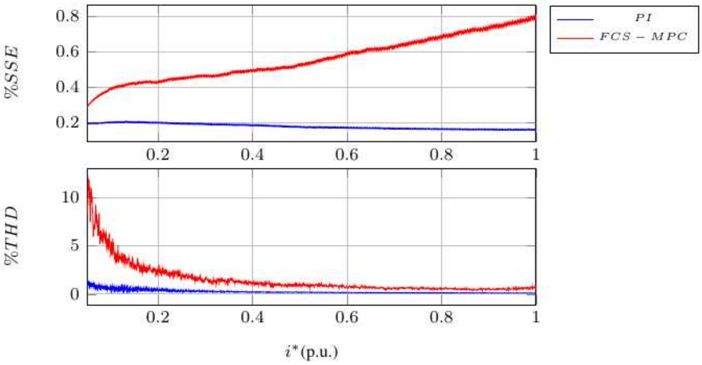


Figure 8

H5 current control results: steady-state error (SSE), and total harmonics distortion (THD)

On the other hand, PI provides a slight less THD by 0.4% compared to 2% of FCS-MPC at small reference values while both controllers result in almost similar THD at large reference values (Figure 8). In addition, PI has a fixed SSE at 0.2% compared to the FCS-MPC's SSE, which is increase from 0.4% at low operating points reaching 0.8% at 1p.u. reference value.

5.3 Case Study 3: Three-Phase Two-Level Inverter Current Control

A three-phase full-bridge inverter was implemented in the following scenario to study the performance of the FCS-MPC compared to the classic PI controller by increasing the reference current significantly from 0 to 1 p.u. It is clear from Figure 9 that the system performance is enhanced by FCS-MPC, where the leakage current is significantly reduced by 50% of the PI leakage current. Another improvement is presented when the common-mode voltage is declined by 0.2 p.u. compared to PI at low operating points as shown in Figure 9. It can be observed from the Figures 9 and 9 the system response has improved with FCS-MPC compared to PI controller; However, there is an acceptable increase in the THD and SSE in both cases FCS-MPC Figure 10.

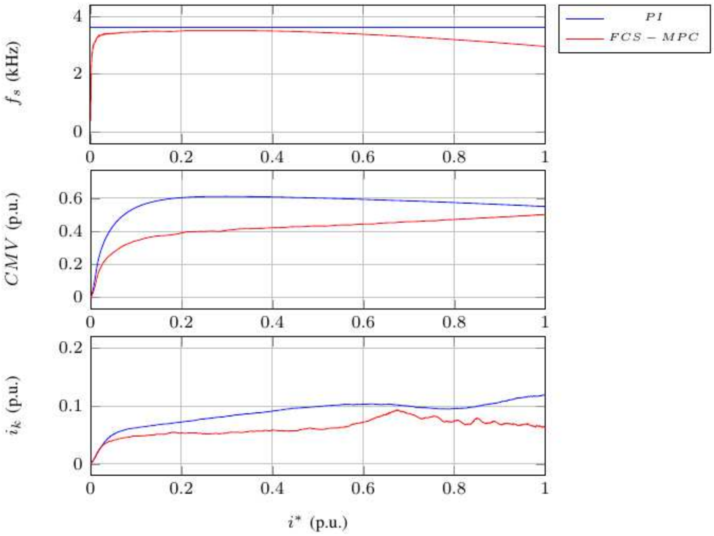


Figure 9

Three-phase two-level Inverter current control results: switching frequency (f_S), common-mode voltage (CMV), and leakage current (i_k)

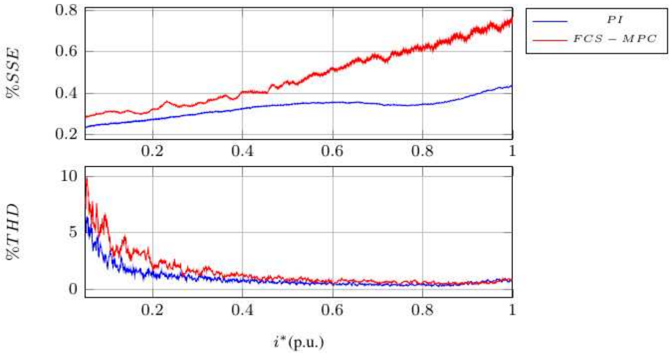


Figure 10

Three-phase two-level Inverter current control results: steady-state error (SSE), and total harmonics distortion (THD)

Conclusions

Among all control techniques, FSC-MPC getting more attention lately for controlling power electronics devices. In this paper, finite control set model predictive control has been compared with the proportional-integral pulls width modulation control for controlling grid-connected power electronics devices. The average switching frequency of the FSC-MPC is adjusted to be below the PI switching frequency for making a proper comparative study. It is evidenced that

FSC-MPC strategy is capable of controlling the grid-tied inverters with superior performance compared to the PI control. It is found that FCS-MPC produces lower common-mode voltage, especially at lower operation points. Moreover, a significant reduction of the leakage current comes to more than 100% in some cases using FCS-MPC. On the other hand, PI control results in lower SSE and THD in all cases, however, FCS-MPC produces SSE and THD with acceptable limits.

Acknowledgement

The authors acknowledge the financial support received from the Scientific Research Deanship Project number 1-440-6178, Taif University, Saudi Arabia.

References

[1] Rehmani, Mubashir Husain, Martin Reisslein, Abderrezak Rachedi, Melike Erol-Kantarci, and Milena Radenkovic. "Integrating renewable energy resources into the smart grid: Recent developments in information and communication technologies." *IEEE Transactions on Industrial Informatics* 14, No. 7 (2018): 2814-2825

[2] Blaabjerg, Frede, Zhe Chen, and Soeren Baekhoej Kjaer. "Power electronics as efficient interface in dispersed power generation systems." *IEEE transactions on power electronics* 19, No. 5 (2004): 1184-1194

[3] Olivares, Daniel E., Ali Mehrizi-Sani, Amir H. Etemadi, Claudio A. Cañizares, Reza Iravani, Mehrdad Kazerani, Amir H. Hajimiragha et al. "Trends in microgrid control." *IEEE Transactions on smart grid* 5, No. 4 (2014): 1905-1919

[4] Tan KK, Zhao S, Xu JX. "Online automatic tuning of a proportional integral derivative controller based on an iterative learning control approach". *IET Control Theory & Applications*. 2007 Jan 1;1(1):90-6

[5] Roman RC, Precup RE, Bojan-Dragos CA, Szedlak-Stinean AI. Combined model-free adaptive control with fuzzy component by virtual reference feedback tuning for tower crane systems. *Procedia Computer Science*. 2019 Jan 1;162:267-74

[6] Precup RE, Roman RC, Teban TA, Albu A, Petriu EM, Pozna C. "Model-free control of finger dynamics in prosthetic hand myoelectric-based control systems". *Studies in Informatics and Control*. 2020 Dec 1;29(4):399-410

[7] Preitl Z, Precup RE, Tar JK, Takács M. "Use of multi-parametric quadratic programming in fuzzy control systems". *Acta Polytechnica Hungarica*. 2006 Sep;3(3):29-43

[8] Haidegger T, Kovács L, Precup RE, Preitl S, Benyó B, Benyó Z. "Cascade control for telerobotic systems serving space medicine". *IFAC Proceedings Volumes*. 2011 Jan 1;44(1):3759-64

[9] Turnip A, Panggabean J. “Hybrid controller design based magneto-rheological damper lookup table for quarter car suspension”. *Int. J. Artif. Intell.* 2020 Mar;18(1):193-206

[10] Twining, Erika, and Donald Grahame Holmes. "Grid current regulation of a three-phase voltage source inverter with an LCL input filter." *IEEE transactions on power electronics* 18, No. 3 (2003): 888-895

[11] Blaabjerg, Frede, Remus Teodorescu, Marco Liserre, and Adrian V. Timbus. "Overview of control and grid synchronization for distributed power generation systems." *IEEE Transactions on industrial electronics* 53, No. 5 (2006): 1398-1409

[12] Chung, Il-Yop, Wenxin Liu, David A. Cartes, Emmanuel G. Collins, and Seung-Il Moon. "Control methods of inverter-interfaced distributed generators in a microgrid system." *IEEE Transactions on Industry Applications* 46, No. 3 (2010): 1078-1088

[13] Rodriguez, Jose, and Patricio Cortes. Predictive control of power converters and electrical drives. Vol. 40. *John Wiley & Sons*, 2012

[14] Vazquez, Sergio, Jose Rodriguez, Marco Rivera, Leopoldo G. Franquelo, and Margarita Norambuena. "Model predictive control for power converters and drives: Advances and trends." *IEEE Transactions on Industrial Electronics* 64, No. 2 (2016): 935-947

[15] Garayalde, Erik, Iosu Aizpuru, Unai Iraola, Iván Sanz, Carlos Bernal, and Estanis Oyarbide. "Finite control set mpc vs continuous control set mpc performance comparison for synchronous buck converter control in energy storage application." In 2019 *International Conference on Clean Electrical Power (ICCEP)*, pp. 490-495, IEEE, 2019

[16] Guzman, Ramon, Luis Garcia de Vicuña, Antonio Camacho, Jaume Miret, and Juan M. Rey. "Receding-horizon model-predictive control for a three-phase VSI with an LCL filter." *IEEE Transactions on Industrial Electronics* 66, No. 9 (2018): 6671-6680

[17] Vazquez, Sergio, Jose I. Leon, Leopoldo G. Franquelo, Jose Rodriguez, Hector A. Young, Abraham Marquez, and Pericle Zanchetta. "Model predictive control: A review of its applications in power electronics." *IEEE industrial electronics magazine* 8, No. 1 (2014): 16-31

[18] Rodriguez, Jos, Jorge Pontt, Csar A. Silva, Pablo Correa, Pablo Lezana, Patricio Cortés, and Ulrich Ammann. "Predictive current control of a voltage source inverter." *IEEE transactions on industrial electronics* 54, No. 1 (2007): 495-503

[19] Lai, Yen-Shin, Cheng-Kai Lin, and Fu-Pao Chuang. "Model-free predictive current control for three-phase AC/DC converters." *IET Electric Power Applications* 11, No. 5 (2017): 729-739

[20] Hu, Jiefeng, Jianguo Zhu, and David G. Dorrell. "Model predictive control of inverters for both islanded and grid-connected operations in renewable power generations." *IET Renewable Power Generation* 8, No. 3 (2013): 240-248

[21] Rashwan, Ahmed, Mahmoud A. Sayed, Youssef A. Mobarak, Gaber Shabib, and Tomonobu Senjyu. "Predictive controller based on switching state grouping for a modular multilevel converter with reduced computational time." *IEEE Transactions on Power Delivery* 32, No. 5 (2016): 2189-2198

[22] Monteiro, Vítor, João C. Ferreira, Andrés Augusto Nogueiras Meléndez, and Joao Luiz Afonso. "Model predictive control applied to an improved five-level bidirectional converter." *IEEE Transactions on Industrial Electronics* 63, No. 9 (2016): 5879-5890

[23] Song, Zhanfeng, Changliang Xia, and Tao Liu. "Predictive current control of three-phase grid-connected converters with constant switching frequency for wind energy systems." *IEEE Transactions on Industrial Electronics* 60, No. 6 (2012): 2451-2464

[24] Young, Héctor, and Jose Rodriguez. "Comparison of finite-control-set model predictive control versus a SVM-based linear controller." *In 2013 15th European Conference on Power Electronics and Applications (EPE)*, pp. 1-8, IEEE, 2013

[25] Young, Hector A., Marcelo A. Perez, and Jose Rodriguez. "Analysis of finite-control-set model predictive current control with model parameter mismatch in a three-phase inverter." *IEEE Transactions on Industrial Electronics* 63, No. 5 (2016): 3100-3107

[26] Mosa, Mostafa, Robert S. Balog, and Haitham Abu-Rub. "High-performance predictive control of quasi-impedance source inverter." *IEEE Transactions on Power Electronics* 32, No. 4 (2016): 3251-3262

[27] Ayad, Ayman, Petros Karamanakos, and Ralph Kennel. "Direct model predictive current control strategy of quasi-Z-source inverters." *IEEE Transactions on Power Electronics* 32, No. 7 (2016): 5786-5801

[28] Chen, Qihong, Xiaoru Luo, Liyan Zhang, and Shuhai Quan. "Model predictive control for three-phase four-leg grid-tied inverters." *IEEE Access* 5 (2017): 2834-2841

[29] Luo, Yixiao, Chunhua Liu, and Feng Yu. "Predictive current control of a new three-phase voltage source inverter with phase shift compensation." *IET Electric Power Applications* 11, No. 5 (2017): 740-748

[30] Victor, Matthais, Frank Greizer, Sven Bremicker, and Uwe Hübler. "Method of converting a direct current voltage from a source of direct current voltage, more specifically from a photovoltaic source of direct

current voltage, into a alternating current voltage." *U.S. Patent* 7,411,802, issued August 12, 2008

[31] Babqi, Abdulrahman J., Zhehan Yi, Di Shi, and Xiaoying Zhao. "Model Predictive Control of H5 Inverter for Transformerless PV Systems with Maximum Power Point Tracking and Leakage Current Reduction." *In IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, pp. 1860-1865, IEEE, 2018

[32] Cherati, S. M., N. A. Azli, S. M. Ayob, and A. Mortezaei. "Design of a current mode PI controller for a single-phase PWM inverter." *In 2011 IEEE Applied Power Electronics Colloquium (IAPEC)*, pp. 180-184, IEEE, 2011

[33] Agarwal, Nikunj, Md Waseem Ahmad, and Sandeep Anand. "Condition monitoring of dc-link capacitor utilizing zero state of solar PV H5 inverter." *In 2016 10th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG)*, pp. 174-179, IEEE, 2016

[34] Schauder, Colin, and Harshad Mehta. "Vector analysis and control of advanced static VAR compensators." *In IEE Proceedings C (Generation, Transmission and Distribution)*, Vol. 140, No. 4, pp. 299-306, IET Digital Library, 1993

[35] Babqi, Abdulrahman J., and Amir H. Etemadi. "MPC-based microgrid control with supplementary fault current limitation and smooth transition mechanisms." *IET Generation, Transmission & Distribution* 11, No. 9 (2017): 2164-2172

[36] Vazquez, S., J. I. Leon, L. G. Franquelo, J. M. Carrasco, E. Dominguez, P. Cortes, and J. Rodriguez. "Comparison Between FS-MPC Control Strategy for an UPS inverter application in α-β and abc frames." In 2010 IEEE *International Symposium on Industrial Electronics*, pp. 3133-3138, IEEE, 2010

[37] Kouro, Samir, Patricio Cortés, René Vargas, Ulrich Ammann, and José Rodríguez. "Model predictive control—A simple and powerful method to control power converters." *IEEE Transactions on industrial electronics* 56, No. 6 (2008): 1826-1838

[38] Cortes, Patricio, Jose Rodriguez, Cesar Silva, and Alexis Flores. "Delay compensation in model predictive current control of a three-phase inverter." *IEEE Transactions on Industrial Electronics* 59, No. 2 (2011): 1323-1325

[39] Cortés, Patricio, José Rodríguez, Daniel E. Quevedo, and Cesar Silva. "Predictive current control strategy with imposed load current spectrum." *IEEE Transactions on power Electronics* 23, No. 2 (2008): 612-618

[40] Barater, Davide, Giampaolo Buticchi, Emilio Lorenzani, and Carlo Concari. "Active common-mode filter for ground leakage current reduction

in grid-connected PV converters operating with arbitrary power factor." *IEEE Transactions on Industrial Electronics* 61, No. 8 (2013): 3940-3950

[41] Duffey, Christopher K., and Ray P. Stratford. "Update of harmonic standard IEEE-519: IEEE recommended practices and requirements for harmonic control in electric power systems." *IEEE Transactions on Industry Applications* 25, No. 6 (1989): 1025-1034

A Novel Risk Assessment Methodology – A Case Study of the PRISM Methodology in a Compliance Management Sensitive Sector

Ferenc Bognár, Petra Benedek

Department of Management and Business Economics
Budapest University of Technology and Economics
Magyar tudósok körútja 2, H-1117 Budapest, Hungary
bognar.ferenc@gtk.bme.hu, benedek.petra@gtk.bme.hu

Abstract: The paper introduces the PRISM methodology built on the critical characteristics of the traditional failure mode and effect analysis (FMEA) and the risk matrix (RM) risk assessment methodologies. The authors create a new definition in the risk assessment process, which is introduced as partial risk. The paper focuses on assessing the compliance risks, the risks of organizational wrongdoing, and legal non-compliance. A real-life case study from the banking sector shows the risk assessment process based on the PRISM method.

Keywords: risk assessment; FMEA; risk matrix; compliance management

Introduction

The current global pandemic and economic crisis have directed the public and legislative focus on risk management and risk prevention. How an organization manages uncertainty can have crucial effects on the customer experience, reputation, competitiveness, and sustainability. Compliance management is a few decades-old business approach to keeping up with fast-changing legal and business requirements. [1] By definition, the purpose of internal controls is to ensure compliance with laws and regulations, the efficiency and effectiveness of the operations, and the credibility of the financial reports. Over the last 20 years, new regulations are being created to such an extent and quantity that compliance with them has become an independent task. This phenomenon has given rise to the organizational function of compliance management. The core definition of compliance is obeying various pieces of internal and external legislation. More recently, a more comprehensive perspective incorporates following the letter and the spirit of the legislation. *Compliance management* is a support function that aims to manage or minimize the risks of organizational wrongdoing and legal non-compliance. Like data loss or information privacy, IT compliance issues affect every department of any organization's daily procedures.

This paper focuses on the presentation of a novel risk assessment methodology via the evaluation of compliance risks. We assume that a combined Failure Mode and Effects Analysis (FMEA) and Risk Map (RM) method can be applied to assess and monitor different kinds of risks, like compliance-related risks. Using the previously mentioned method, organizations would formulate measures for the organization's current, individual operation to reduce error modes' frequency or improve failure and error detection.

This research examines how suitable are the new combined FMEA and RM method in the risk assessment and evaluation of financial service companies' compliance organizations.

The first part of the article is an overview of compliance management, focusing on compliance risk assessment. The essence of the compliance concept is a social and economic interpretation, a novelty that assesses non-compliance as a risk. This risk consists primarily of two factors: regulatory risk and reputational risk. In this part, we introduce the relevant standards like ISO 31000: 2018, IEC 31010: 2019, ISO / IEC 27005: 2018, and ISO 19600: 2014 guidelines.

In the second part, the traditional concept of FMEA and Risk Map is presented. The first method aims to identify the existing or possible failures and their cause, estimating the failures' risks. Risk matrices apply two rating factors, which are used to estimate the "probability" and the "impact" dimensions.

In the following part, we present the concept of partial risk and the new PRISM method based on a unique combination of FMEA and risk maps. Later, a case study from the banking sector shows the practical implementation of the newly proposed method. A discussion and further research hypothesis close this paper.

1 Risk Evaluation in Compliance Management

The first significant compliance management publications describe the relations among transparency, business ethics, and compliance. [2, 3] The post-millennium scandals brought the relevant thematic boom. Standing out of the many was the Turner Review [4], which analyzed the management theory of the global banking crisis, Silverman's comprehensive organizational Compliance Management [5] and the Governance, Risk and Compliance Handbook [6].

The US is serving with the most prominent examples of expectations of corporate compliance systems. The Federal Sentencing Guidelines for Organizations last amended in 2018 [7, 8], the Sarbanes-Oxley Act from 2002 [9, 10], and the COSO Internal Control–Integrated Framework [11, 12] stand as guidelines for the minimum requirement for today's compliance systems. Each organization can tailor its use of the above to its business, and other standards related to the professional profile may also be relevant.

Major international organizations (e.g. UN, World Bank), as well as national governments, have also developed and published several general and thematic directives and best practice recommendations, such as the updated OECD Principles on Corporate Governance (2015), the Corporate Responsibility to Respect Human Rights (2003), UN Global Compact (2000), UN Principles for Responsible Investment (2006).

All business activities are risky to some extent, and these risks can be measured, analyzed, reduced, managed, i.e. kept below a certain level. The task of risk management is to keep the probability of possible effects occurring at some conscious level. Compliance mainly focuses on legal and regulatory requirements. According to a strict approach, legality is not a matter of consideration but merely a requirement. According to the standard approach, compliance manages unique compliance risks. In many cases, the interpretation of legislation gives decision-makers a degree of freedom so that discretion does not appear at the level of taking or rejecting a particular risk but at how 'compliant' is any given solution [13].

Compliance risk consists primarily of two factors: penalties for non-compliance and reputational risk. *Regulatory risks* are assessed based on the potential penalty and the likelihood of falling. There is a relatively straightforward risk level above which compliance officers vetoe the risky decision or product in question. On the one hand, the regulations, requirements, and legislation changes that apply to the organization must be monitored. The tasks, risks, and responsibilities associated with the given legislation or change must be defined.

On the other hand, all other compliance activities provide information on where the organization is facing deficiencies or errors concerning its objectives and how risk management can be continuously improved. The goal of compliance is not to build a bottom-up system of legal references but vice versa. Based on international practices and experiences, each organization defines the relevant compliance risks. Compliance fundamentally incorporates developing a risk management methodology and planning and implementing internal controls related to the specific compliance risks.

Reputational risk is different in different markets. On the one hand, it is a reputational risk that customers become unloyal due to an incident. More importantly, if the organization becomes risky, it can lose its partners, which is a severe threat to its operations. For example, in the spring of 2018, the Latvian bank ABLV was liquidated weeks after it was suspected of connecting to North Korea's weapons development program [14]. All market participants reacted to the news by closing the partnership. If the information, data, customer due diligence, or anything is unreliable and laundered money comes in, that is unacceptable.

Reputational risk is a powerful motivation to operate a robust compliance function.

There is a worldwide effort to define a quality assurance framework for compliance by standards. We would like to highlight ISO 31000: 2018, IEC 31010: 2019, ISO / IEC 27005: 2018, and ISO 19600: 2014 guidelines.

ISO 31000: 2018 guides how to manage the risks faced by organizations. Every organization tailors these guidelines to its environment and operation in practice. The guidelines help in any activity, including decision-making, at all levels. Based on the guidelines' foreword, we would like to focus on two main changes from the 2009 version.

1) "highlighting of the leadership by top management and the integration of risk management, starting with the governance of the organization;
2) greater emphasis on the iterative nature of risk management, noting that new experiences, knowledge, and analysis can lead to a revision of process elements, actions, and controls at each stage of the process. " [15]

The importance of iterative risk management returns in several places in the text; the idea of continuous improvement of Total Quality Management (TQM). The emphasis on leadership and commitment, as well as inclusive responsibility ("Everyone in an organization has responsibility for managing risks." [16]), is also in line with the TQM philosophy.

IEC 31010: 2019 guides the selection and application of risk assessment techniques that help make decisions when there is uncertainty [17]. The 2019 edition of the guidelines contains summaries of an increased number of techniques, referring to other documents which describe the methods and techniques in more detail. "The standard is useful both as part of a process to manage risk and when comparing options and opportunities so that decisions are based on a good understanding of risk," said Professor Jean Cross [18].

ISO 27005: 2018 is designed to help implement information security based on a risk management approach [19]. This standard applies to any organization that seeks to address risks that could compromise its information security.

ISO 19600: 2014 Compliance Management Systems is currently one of the most critical international recommendations for business compliance management, which describes the cooperation between compliance assurance and risk management [20]. The "AS 3806 - Compliance Programs" standard established in the Australian financial sector in 1998, updated in 2006, is the document's predecessor. The document's cited sources show that this directive relates to ISO 9001, the ISO 10002 complaint handling standard, and the social responsibility guidelines (ISO 26000). The 19600: 2014 guidelines for compliance management systems are close to the ISO 31000 risk management standard. Table 1 shows the comparison of the processes in two documents.

Table 1

Management processes in ISO standards, not exhaustive

ISO 31000: 2018	ISO 19600: 2014
Communication and consultation	Communication
Creating the context (Scope, context, criteria)	Creating the context (Scope, context, criteria) and Developing a Compliance Management System
Risk identification	Identification of compliance obligations and related compliance risks
Risk analysis	Risk analysis - the probability and impact of non-compliance
Risk evaluation	Risk evaluation - prioritization
Risk treatment	Risk treatment - planning and implementation of controls
Recording and Reporting	Performance Evaluation and Compliance Reporting

Source: own editing based on ISO 31000: 2018 and ISO 19600: 2014

Every organization is unique. Therefore, compliance systems differ depending on the industry and specific risks. At the same time, good practices outlined in ISO 19600:2014 cover specific areas of ethical corporate operation and serve as guidelines for organizations. According to ISO 19600:2014, integrity and compliance could be considered an opportunity for developing a successful and sustainable organization.

The ISO 19600:2014 standard facilitates the design, implementation, evaluation and maintenance of the compliance system. In the flowchart (Figure 1), the modified PDCA cycle's first step is to identify compliance obligations and evaluate compliance risks. The second step is to address these risks and set measurable objectives related to them. Planning is followed by operation and control of the compliance risks. Per the logic of the PDCA cycle, implementation is followed by performance evaluation and reporting. The outcome of performance evaluation is getting a systematic overview of the strengths and the weaknesses of the system, highlighting the areas for possible development. The fourth step is the management of non-compliance and the continual improvement of the system. Similar to ISO 9001:2015, leadership is a critical factor that ensures all the other flowchart elements cooperate properly.

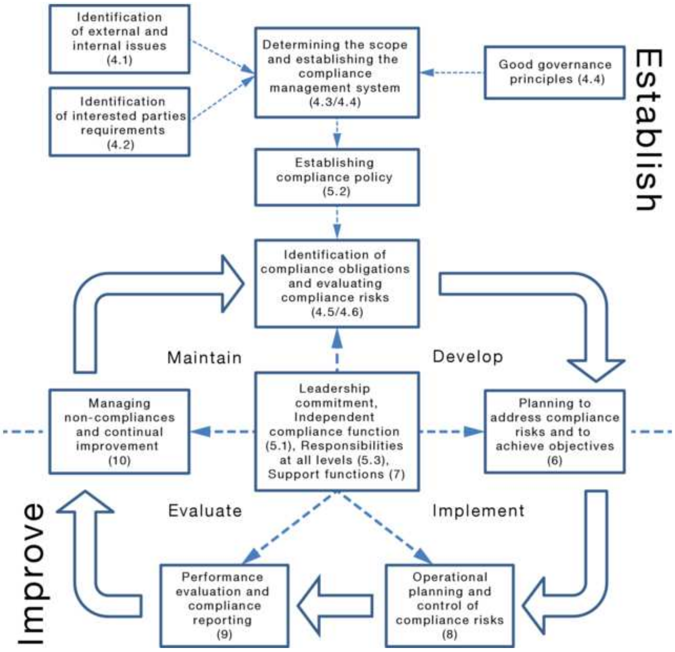

Identification of external and internal issues (4.1)
Identification of interested parties requirements (4.2)
Determining the scope and establishing the compliance management system (4.3/4.4)
Good governance principles (4.4)
Establish
Establishing compliance policy (5.2)
Identification of compliance obligations and evaluating compliance risks (4.5/4.6)
Maintain
Develop
Managing non-compliances and continual improvement (10)
Leadership commitment, Independent compliance function (5.1), Responsibilities at all levels (5.3), Support functions (7)
Planning to address compliance risks and to achieve objectives (6)
Evaluate
Implement
Improve
Performance evaluation and compliance reporting (9)
Operational planning and control of compliance risks (8)

Figure 1

Flowchart of a compliance management system

Source: ISO 19600: 2014 [21]

Compliance management is aligned with risk management in the general sense. The standard recognizes the risk-based approach to compliance. It is also familiar with the concept of complex risk appetite, the extent to which an investor or organization is willing to take risks. In the following chapters of this paper, a new methodology is presented to provide a practical approach to compliance risk evaluation.

2 Risk Assessment Based on the Failure Mode and Effect Analysis (FMEA) and the Risk Matrix (RM)

Failure mode and effect analysis (FMEA) is a widely applied and developed methodology. The methodology is performed continuously in most manufacturing industries and developed by researchers in numerous research papers. [22] Nowadays, the most dominant research development field of FMEA is based on multi-criteria decision making (MCDM) methodologies. [23]

2.1 Failure Mode and Effect Analysis

The FMEA methodology is applied to assess the risks of potential or existing failures in particular objects and prevent these failures from occurring. FMEA can significantly improve the reliability of different complex systems from technology-based services to all production fields. In the last years, many case studies were published related to the development of FMEA in connection with the IT sector. Case studies introduce the application of FMEA in highly IT-relevant fields, just like internet banking services [24] and healthcare systems. [25]

The traditional concept of FMEA is to identify the existing or possible failures and their cause, estimating the risks of the failures and reducing the risk of the failure. The target field of the analysis is traditionally a product or a process. First, a cross-functional team is set up to identify the relevant existing or possible failures using creative techniques. Identifying the failures can be a long process, depending on the nature, complexity, and size of the particular product or process. Once the cross-functional team identifies the failures, the team performs the risk analysis phase of the methodology.

The most crucial goal of the risk analysis is to determine the resultant value of each failure risk. This value is typically interpreted as a Risk Priority Number (RPN) and calculated using three rating factors. The value of occurrence (O), severity (S) and detection (D) is generally applied in the assessment process of the RPN. We calculate RPN as follows

$$RPN = OxSxD \tag{1}$$

where O is the probability of failure, S is the severity of the failure effect, and D is the probability of non-detecting the failure. The value of these rating factors can be estimated using numerous ways. For obtaining the RPN of a specific failure mode, the three risk factors are evaluated using different ten-point scales.

There has been a broad consensus in the research community on which scales should we evaluate each rating factor in recent decades. [26-30] However, in practice, the scales are often transformed to meet the analyzed product or process's measurement or estimation requirements. Based on the literature review of Liu [22], we apply Tables 2-4 to evaluate the three rating factors.

Table 2

Ratings for the occurrence [22]

Probability of failure	Possible failure rates	Rank
Extremely high: failure almost inevitable	≥ in 2	10
Very high	1 in 3	9
Repeated failures	1 in 8	8
High	1 in 20	7
Moderately high	1 in 80	6

Moderate	1 in 400	5
Relatively low	1 in 2000	4
Low	1 in 15000	3
Remote	1 in 150000	2
Nearly impossible	≤ 1 in 1500000	1

Table 3

Ratings for the severity [22]

Effect	Criteria: severity of the effect	Rank
Hazardous	Failure is hazardous and occurs without warning. It suspends the operation of the system or involves non-compliance with government regulations.	10
Serious	Failure involves hazardous outcomes or non-compliance with government regulations or standards.	9
Extreme	The product is inoperable with a loss of primary function. The system is inoperable.	8
Major	Product performance is severely affected but functions. The system may not operate.	7
Significant	Product performance is degraded. Comfort or convince functions may not operate.	6
Moderate	Moderate effect on product performance. The product requires repair.	5
Low	Small effect on product performance. The product does not require repair.	4
Minor	Minor effect on product or system performance.	3
Very minor	Very minor effect on product or system performance.	2
None	No effect.	1

Table 4

Ratings for the detection [22]

Detection	Criteria: the likelihood of detection by the design control	Rank
Absolute uncertainty	Design control does not detect a potential cause of failure or subsequent failure mode, or there is no design control.	10
Very remote	Very remote chance that the design control will detect a potential cause of failure or subsequent failure mode.	9
Remote	Remote chance that the design control will detect a potential cause of failure or subsequent failure mode.	8
Very low	Very low chance that the design control will detect a potential cause of failure or subsequent failure mode.	7
Low	Low chance that the design control will detect a potential cause of failure or subsequent failure mode.	6
Moderate	Moderate chance that the design control will detect a potential cause of failure or subsequent failure mode.	5
Moderately high	Moderately high chance that the design control will detect a potential cause of failure or subsequent failure mode.	4

High	High chance that the design control will detect a potential cause of failure or subsequent failure mode.	3
Very high	Very high chance that the design control will detect a potential cause of failure or subsequent failure mode.	2
Almost certain	Design control will almost certainly detect a potential cause of failure or subsequent failure mode.	1

The higher the factor-related risk of a particular failure mode, the higher the rating factor's value. The higher the overall risk of a particular failure mode, the higher the RPN value. Based on the RPN value, the failure modes can be prioritised to find the riskiest failure modes. If it is necessary, the prioritisation can be applied based on a specific rating factor as well. This step is essential since there are not enough resources to reduce all the possible risks in a product, machine or process. Thus, based on the prioritisation, the focus can be on the most important – so on the riskiest – failure modes. The riskiest failure modes are being placed under investigation for reducing the risk by proper actions. After the corrective actions, the rating factors' values are estimated again, so the iteration starts again.

2.2 Risk Matrix (RM)

Risk matrices represent another widely applicable group of risk assessment methodologies. Similar to the FMEA methodology, the risk matrix is built up by rating factors developed to assess a particular object's risk. [31] While FMEA applies three rating factors, risk matrices apply only two rating factors, which are usually used to estimate the "occurrence" and the "severity" dimensions. [32] Thus, the risk assessment tool's general structure is a matrix, as visible in Figure 2.

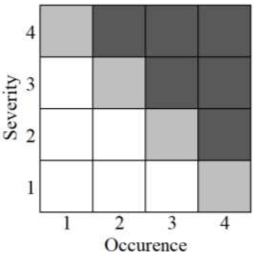


Figure 2

An example of the structure of the risk matrix

In general, the methodology estimates the risk on a 1-3 or 1-4 or 1-5 scale. Similarly to the FMEA, the higher the factor-related risk of a specific failure mode, the higher the rating factor's value. It often occurs that the rating factors of the risk matrix have different scale lengths. Thus, the risk matrix has a non-equal number of rows and columns.

The risk assessment is based on the score of the “occurrence” and “severity” assessment factors. If both the rating factors have high values, the associated risk will be judged high, while when the rating factors have low values, they will be interpreted as low. As shown in Figure 2, the darker the colour of a matrix cell, the higher the associated risk of the failure mode.

As for the risk assessment result, different action categories are available for further steps, aiming to reduce the determined risk level. Based on the matrix cell's colour and the risk level, different actions can be launched, from the "no action needed" category to the "immediate intervention" action. Thus, the methodology classifies the failure modes into different groups, while the groups have ranks and failure modes have only group belonging identifications.

Both methodologies are powerful risk assessment tools that focus on developing the given product, process, or system. In the following part of the paper, we describe a methodology that builds on both the FMEA and the risk matrix's strengths, creating a new, more robust, and practical methodology.

3 The Definition of Partial Risk and the PRISM Methodology

There are failures in the business processes, systems, and products, which have strong connections with the compliance management system. These failures have relatively higher risk content than those processes, systems, and products, which are not directly compliance sensitive. In those sectors, where the compliance management systems have to be highly developed and linked to the organizational business processes, risk estimation has a more critical role than in other operational fields. In this chapter, a novel risk evaluation methodology is described, based on a combination of the failure mode and effect analysis and the risk matrix. Since both FMEA and RM have significant risk evaluation abilities, the new methodology is designed to build on the synergies of these abilities.

FMEA helps rank the risk of different failure modes and effects, and the methodology generally focuses on the value of the RPN. The problem is that multiplication can mask the detailed information held by each rating factor. A failure can have a low RPN value, while the failure's partial risk can be relatively high. Table 5 shows detailed examples of partial risk cases.

Table 5

Examples for partial risks

Case	Occurrence (O)	Severity (S)	Detection (D)	RPN
Case 1	1	10	5	50
Case 2	1	7	7	49
Case 3	10	4	1	40

As shown in Table 5, all the cases have a relatively low risk based on the RPN value. Nevertheless, a relatively small increase of the Occurrence rating factor value can significantly raise the RPN value at “Case 1” as well as at “Case 2”, while a slight increase of the “Detection” rating factor value of “Case 3" results in a significant increase in the RPN value. When the result of a multiplication of two rating factors is high, while the third rating factor's value is relatively low, the case of partial risk emerges.

A three-time risk matrix evaluation can amend the failure mode and effect analysis for the detailed risk estimation of failure modes. Risk matrices can evaluate the partial risks based on three different contexts: “occurrence vs. severity” and “occurrence vs. detection”, and “severity vs. detection”. All three analyses should be performed at the same time for gathering all the necessary information on the possibly existing partial risks. Figure 3 shows the map of the three different, partial analyses.

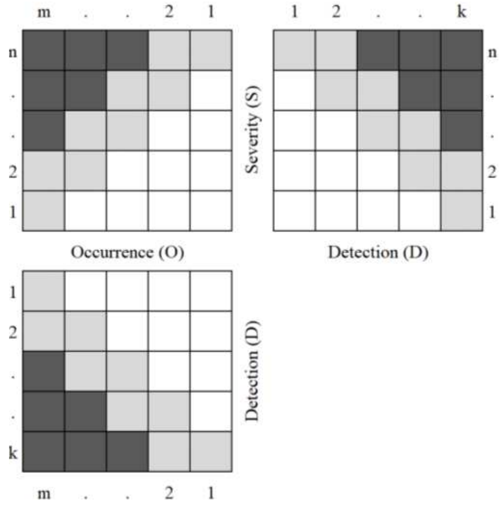


Figure 3

The general model of the PRISM (Partial Risk Map) risk evaluation methodology

In the general model, rating factors have the same scale length, so “k”, “n”, and “m” values are equal to each other. However, the scale length could be different if the practical case requires that. Furthermore, all the "k", "n" and "m" values can be different.

The colourings of the map are similar to the traditional risk matrix. Thus, the darker the colour of a matrix cell, the higher the failure mode's hidden risk. The map's colourings are changeable related to the practical problem and the application field.

According to the PRISM methodology, a failure mode could be determined as a potentially risky failure mode if any of the forthcoming criteria are fulfilled:

(1) the RPN value reaches a specific indicator value which the experts previously set;

(2) based on the values of the occurrence and severity rating factors, the failure mode position is inside that part of the O vs. S matrix, which was set to be risky;

(3) based on the values of the occurrence and detection rating factors, the failure mode position is inside that part of the O vs. D matrix, which was set to be risky;

(4) based on the values of the severity and detection rating factors, the failure mode position is inside that part of the S vs. D matrix, which was set to be risky.

If criterion (1) is fulfilled without fulfilling any other criteria, the failure mode could be considered risky because of the overall RPN value. If any of the criteria (2), (3), or (4) is fulfilled without fulfilling criterion (1), the failure mode could be considered risky because of partial risk.

4 A Case Study from the Banking Sector

In 2021, after several discussions with compliance experts from the Hungarian retail banking sector, a workshop was organized to test the above-proposed PRISM methodology's usability on actual data. Based on real-life non-compliance cases given by the bank experts, researchers have proposed the first version of the scales of the assessment of FMEA factors and a list of the selected compliance incidents.

Based on the workshop discussion, the proposed scales of the assessment were modified, and participants have come to a common understanding. The resulting 4-grade scales in all three rating factors (occurrence, severity, and detection) are shown in Tables 6, 7 and 8.

Table 6

Ratings for the occurrence

Probability of failure	Possible failure rates	Rank
High	weekly	4
Moderate	monthly	3
Low	yearly	2
Remote	less often than once a year	1

Table 7

Ratings for the severity

Effect	Criteria: severity of the effect	Rank
Major	Severe financial, reputational or legal consequences.	4
Significant	Significant financial loss or reputational impact, legal consequences.	3
Moderate	Small financial loss, slight negative reputational impact.	2
Low	No or minor financial loss, no reputational impact.	1

Table 8

Ratings for the detection

Detection	Criteria: the likelihood of detection by the design control	Rank
Absolute uncertainty	Design control does not detect a potential cause of failure or subsequent failure mode, or there is no design control.	4
High	Internal control detects the potential cause of failure or subsequent failure mode	3
Moderate	The second line detects the event.	2
Low	Management control detects the potential cause of failure or subsequent failure mode	1

Experts set that corrective actions have to be launched if the RPN value reaches 20 points of the maximum amount of 64 points. In the next step, the experts determined that two significant outcomes can be proposed based on the risk matrices, as shown in Figure 4. The matrices' grey cells indicate the necessity of corrective actions since the partial risk is high; the white cells indicate low partial risk, so no corrective action is required.

During the workshop, three compliance experts have rated six compliance events individually. All the chosen compliance risks represent human risks. In each case, the bank clerk does not make the right decision in a given situation. By doing so, there is a compliance risk as a result of a wrong decision. Based on a discussion, following the individual ratings, a joint rating was created, as shown in Table 9.

Table 9
Compliance risks in FMEA

Case	Function/ Process step	Potential failure mode	Potential effects of failure	S	Potential causes of failure	O	Current process controls	D
A	cash withdrawal in a bank branch	a young person accompanies the elderly customer	client losing wealth	3	the client is forced to withdraw cash	2	make sure of the client's intentions	4
B	looking into client accounts	checking acquaintance's account after a phone call on business mobile	protocol violation	1	negligence or ignorance of protocols	4	random call controls, managerial controls of account lookups	3
C	replying to a customer inquiry about account abuse	customer misinformation, lack of reporting to bank security	client losing wealth, security incident	2	negligence or ignorance of internal procedure	2	employee training	4
D	cash withdrawal, account closing	the legal representative of a minor client withdraws the full amount and closes the account	minor client losing wealth	2	negligence or ignorance of internal procedure, incomplete internal procedure	1	protocols for checking personal documents	4
E	new account opening	Bank clerk opening a new account for a family member	conflict of interest, protocol violation	1	negligence or ignorance of protocols	3	managerial control	2
F	offering travel insurance	lack of reporting foreign card use	credit card abuse	2	missing protocol	4	none	4

The six cases' risk can be ranked by the RPN value, based on the multiplication of the occurrence, severity and detection values, as shown in Table 10. The higher the RPN value of a particular failure mode, the lower the ranking value.

Table 10
Risk Priority Number values of the compliance cases

Case	Occurrence (O)	Severity (S)	Detection (D)	RPN	Rank
A	3	2	4	24	2
B	1	4	3	12	4
C	2	2	4	16	3
D	2	1	4	8	5
E	1	3	2	6	6
F	2	4	4	32	1

The risk matrices in Figure 4 display the partial risks based on the three different contexts: “occurrence vs. severity”, and “occurrence vs. detection”, and “severity vs. detection”.

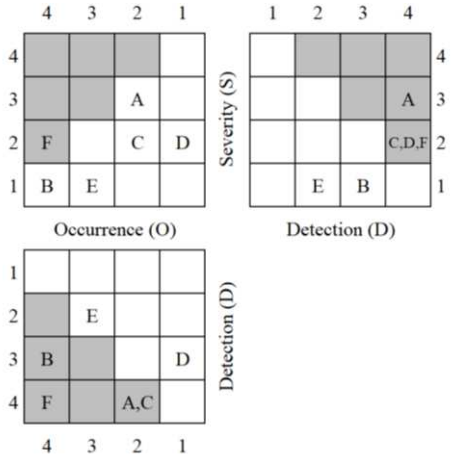


Figure 4
The PRISM pattern of the cases

Based on the RPN values and the PRISM pattern, a detailed risk assessment can be executed. As a result of the assessment, it is evident that "Case F" is the riskiest case since it has the RPN value above the previously set limit. At the same time, and it is represented three times in the PRISM pattern. "Case A" also reaches the previously set limit of the RPN value, and it is represented two times in the PRISM pattern. Though "Case C" does not reach the RPN limit, it still appears two times in the PRISM pattern, so it is necessary to launch corrective action in this case as well. "Case B" and "Case D" are under the RPN limit, but both of them are represented in the PRISM pattern once, so corrective action has to be performed in their case as well. "Case E" is under the RPN limit, and it has no appearance in the PRISM pattern. Therefore, this is the only case where no corrective action is needed.

Table 11
The detailed results of the PRISM analysis

Case	Corrective action required based on the RPN value	Corrective action required based on the PRISM pattern
A	x	x
B		x
C		x
D		x
E		
F	x	x

Table 11 summarises the required corrective actions for reducing the risk level in each compliance case. After the corrective actions were applied, a new risk assessment is performed to identify the failure modes' risk reduction.

Discussion

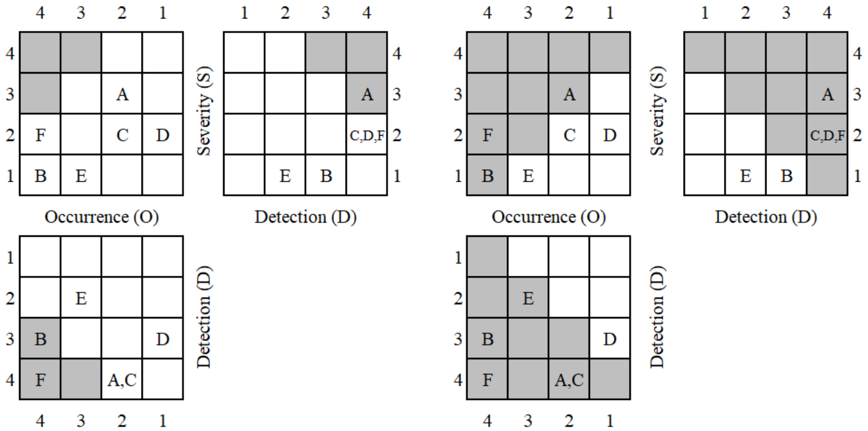
In the PRISM methodology, the traditional RM is applied to estimate the partial risks related to the failure modes' occurrence and severity. Simultaneously, two modified RM is also applied to estimate the occurrence and detection-based partial risks and the severity and detection-based partial risks. It is unequivocal that the three rating factors of the traditional FMEA applied in the PRISM methodology. Additionally, PRISM can also create an RPN-based ranking of different failure modes.

PRISM methodology can be interpreted as a combination and extension of the RM and FMEA. The methodology aims to describe the partial risks, which would stay hidden if only the FMEA or RM were applied. Thus, the methodology gives a more efficient and detailed view of the risk assessment result, which can be necessary for compliance sensitive and safety requiring systems. Based on the RPN values and the possibly existing partial risks, risk reductive action plans can be designed and launched.

Users can customize the PRISM methodology to the assessment's objective area, and PRISM can be useful when the corrective actions' focus has to be defined. Since partial risks can be identified as a result of the assessment, a more detailed risk-reduction action can be formed.

The PRISM methodology is a hybrid methodology that builds on the essential characteristics of the FMEA and the RM methodologies. Based on the parametrization, PRISM can be applied as a methodology that builds more to the RPN value during the risk assessment or focuses more on partial risks. Thus, the methodology can be extensively customizable to fulfil the user needs.

For example, Figure 5 shows customization options for the RPN focused risk assessment (a) and for the partial risk-focused risk assessment (b) as well.



(a) (b)

Figure 5
Customisation of the PRISM methodology

As visible in Figure 5, the customization's critical factor is the grey cells' pattern, indicating high partial risk. Case (a) shows an example where the grey cells are involved only around the rating factors' highest values. In this case, the RPN based risk assessment has more impact on the PRISM analysis. Case (b) shows an example where the grey cells have a significantly more extensive pattern than in case (a). In this case, the partial risk-based assessment has more impact on the PRISM analysis result.

The direction of the corrective actions can be set based on the position of a particular risk. The fact that the partial risk reaches the threshold at one or more part of the PRISM indicates the direction of the action plan. The most significant advantage of PRISM over the FMEA or RM is that PRISM directs the focus on those cases where the partial risk is high, but the entire RPN is low. In this case, a relatively small increase of a relatively small value of an assessment factor can result in a dangerously high overall risk, resulting in serious outcomes, especially in risk-sensitive sectors.

It is unequivocal that the PRISM methodology can apply more and different corrective action categories as well. In this case study, only two resulting categories were defined: "necessity of corrective action" and "no corrective action required". However, in other cases, more warning labels can support a detailed and more sensitive action plan.

Authors note that PRISM analysis can be performed without taking into account the RPN values. In this case, the added value of the FMEA methodology during the risk assessment process is that the PRISM methodology uses the “detection” rating factor of the traditional FMEA.

Conclusions

As a result of ever-changing external regulations and internal development, compliance nowadays appears as a specific problem. In the latest guidelines, like ISO 19600:2014, risk management and compliance management integration is highlighted.

The traditional FMEA methodology is applied to identify and assess potential or existing failure modes' risks, estimating the severity, occurrence, and ease of detecting specific failure modes. Based on the RPN value, the failure modes can be prioritised to find the riskiest ones. Corrective actions based on FMEA aim to reduce the risks. Furthermore, risk matrices apply two rating factors, which usually estimate the "occurrence" and the "severity" dimensions. Both FMEA and RM methodologies are powerful risk assessment tools.

This paper has introduced the notion of partial risk and the new PRISM methodology that combines and exceeds both the FMEA and the risk matrix's strengths, creating a new, more robust, and practical methodology. Partial risk maps can lead to a better understanding of partial risks and serve as a basis for preventive and corrective actions.

In this paper's case study, a list of compliance incidents was rated on three factors: the occurrence, severity, and ease of detection. The traditional scales have been tailored based on discussion with financial sector compliance experts.

The PRISM methodology gives a more efficient and detailed view of the risk assessment, which can be necessary for compliance-sensitive and safety-requiring systems. Based on the parametrization, PRISM can be easily customized to focus more on the RPN value or focus more on partial risks. Users can apply more and different corrective action categories (like installing alarms, training personnel, updating processes) to support a more detailed and sensitive action plan.

References

[1] Kecskés, A. (2010): Tendencies of Corporate Governance Development, Concepts of Regulation in Europe and the United Sates, PhD Thesis, https://ajk.pte.hu/sites/ajk.pte.hu/files/file/doktori-iskola/kecskes-andras/kecskes-andras-vedes-tezisek.pdf, pp. 20-21, 22/01/2021

[2] Paine, L. S. (1994): Managing for Organizational Integrity, Harvard Business Review, v72 n2 p106-17 Mar-Apr 1994

[3] Trevino, Weaver, Gibson, Toffler (1999): Managing Ethics and Legal Compliance, what works and what hurts, California Management Review, Vol. 41, No. 2, Winter 1999, pp. 131-151

[4] The Turner Review, a regulatory response to the global banking crises, Financial Services Authority, March 2009, http://www.fsa.gov.uk/pubs/other/turner_review.pdf, p.79-80

[5] Silverman, M. (2008): Compliance management for Public, Private, and Nonprofit Organizations, Mc Graw Hill

[6] Tarantino, A. (2008): Governance, Risk and Compliance Handbook, John Wiley & Sons

[7] Murphy, D. E. (2002): The Federal Sentencing Guidelines for Organizations: A Decade of Promoting Compliance and Ethics, Iowa Law Review, 87, 697-719

[8] The Federal Sentencing Guidelines for Organizations, chapter 8, https://www.ussc.gov/guidelines/2018-guidelines-manual/annotated-2018-chapter-8, 14/02/2021

[9] Sarbanes-Oxley Act (2002), Public Law 107–204—July 30, 2002, https://www.govinfo.gov/content/pkg/PLAW-107publ204/pdf/PLAW-107publ204.pdf, 14/02/2021

[10] Commission Guidance Regarding Management's Report on Internal Control Over Financial Reporting Under Section 13(a) or 15(d) of the Securities Exchange Act of 1934, 2007, https://www.sec.gov/rules/interp/2007/33-8810.pdf, 14/02/2021

[11] COSO Internal Control – Integrated Framework, 2013, https://www.coso.org/Documents/990025P-Executive-Summary-final-may20.pdf, 14/02/2021

[12] McNally, J. S. (2013): The 2013 COSO Framework & SOX Compliance, https://www.coso.org/documents/COSO%20McNallyTransition%20Article-Final%20COSO%20Version%20Proof_5-31-13.pdf, 14/02/2021

[13] Benedek, P. (2019): Compliance menedzsment a pénzügyi szolgáltatásokban, Munkaügyi Szemle, 62 : 4 pp. 41-51, 11 p.

[14] Coppola, F. (2018): Why The U.S. Treasury Killed A Latvian Bank, Forbes, https://www.forbes.com/sites/francescoppola/2018/02/28/why-the-u-s-treasury-killed-a-latvian-bank/?sh=76bd5d627adc, 22/01/2021

[15] ISO 31000:2018 Risk management — Guidelines, foreword, 2018, https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en, 14/02/2021

[16] ISO 31000:2018 Risk management — Guidelines, foreword, 2018, https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en, 14/02/2021

[17] IEC 31010:2019 Risk management — Risk assessment techniques, 2019, https://www.iso.org/obp/ui/#iso:std:iec:31010:ed-2:v1:en,fr, 14/02/2021

[18] Naden, C. (2019): Understanding Risk with Newly Updated International Stadard, https://www.iso.org/news/ref2403.html, 08/01/2020

[19] ISO/IEC 27005:2018 Information technology — Security techniques — Information security risk management, 2018, https://www.iso.org/obp/ui/#iso:std:iso-iec:27005:ed-3:v1:en, 14/02/2021

[20] ISO 19600:2014 Compliance management systems — Guidelines, 2014, https://www.iso.org/obp/ui/#iso:std:iso:19600:ed-1:v1:en, 14/02/2021

[21] ISO 19600:2014 Compliance management systems — Guidelines, 2014, https://www.iso.org/obp/ui/#iso:std:iso:19600:ed-1:v1:en, 14/02/2021

[22] Liu, H.C., Liu, L., Liu, N.: Risk evaluation approaches in failure mode and effects analysis: A literature review. Expert Systems with Applications 40 (2013) 828-838

[23] Liu, H. C., Chen, X. Q., Duan, C. Y., Wang, Y. M.: Failure mode and effect analysis using multi-criteria decision making methods: A systematic literature review. Computers & Industrial Engineering 135 (2019) 881-897

[24] Chen, L., Deng Y.: A new failure mode and effects analysis model using Dempster–Shafer evidence theory and grey relational projection method. Engineering Applications of Artificial Intelligence 76 (2018) 13-20

[25] Song, W., Li, J., Li, H., Ming, X.: Human factors risk assessment: An integrated method for improving safety in clinical use of medical devices. Applied Soft Computing Journal 86 (2020) 105918

[26] Chang, K. H.: Evaluate the orderings of risk for failure problems using a more general RPN methodology. Microelectronics Reliability, 49 (2009) 1586-1596

[27] Chang, K. H., Cheng, C. H.: A risk assessment methodology using intuitionistic fuzzy set in FMEA. International Journal of Systems Science, 41 (2010) 1457-1471

[28] Liu, H. C., Liu, L., Liu, N., & Mao, L. X.: Risk evaluation in failure mode and effects analysis with extended VIKOR method under fuzzy environment. Expert Systems with Applications, 39 (2012) 12926-12934

[29] Sankar, N. R., & Prabhu, B. S.: Modified approach for prioritization of failures in a system failure mode and effects analysis. International Journal of Quality & Reliability Management, 18 (2001) 324-336

[30] Seyed-Hosseini, S. M., Safaei, N., & Asgharpour, M. J.: Reprioritization of failures in a system failure mode and effects analysis by decision making trial and evaluation laboratory technique. Reliability Engineering & System Safety, 91 (2006) 872-881

[31] Qazi, A., Shamayleh, A., El-Sayegh, S., Formaneck, S.: Prioritizing risks in sustainable construction projects using a risk matrix-based Monte Carlo Simulation approach. Sustainable Cities and Society, 65 (2021) 102576

[32] Wang, R., Wang, J.: Risk Analysis of Out-drum Mixing Cement Solidification by HAZOP and Risk Matrix. Annals of Nuclear Energy, 147 (2020) 107679

Impact of Different CAM Strategies and Cutting Parameters on Machining Free-Form Surfaces with Ball-End Milling Tools in Terms of Micro and Macro Accuracy

Bálint Varga, Balázs Mikó

Institute of Material and Manufacturing Science, Óbuda University
Népszínház u. 8, H-1081 Budapest, Hungary
e-mail: varga.balint@bgk.uni-obuda.hu, miko.balazs@bgk.uni-obuda.hu

Abstract: The use of free-form surfaces is becoming more common in everyday life. Ergonomic, aesthetic, aerodynamic and fluid dynamics aspects also help their spread. The paper examines the CAM aspects of machining the free-form surfaces with a ball-end milling tool. Different CAM strategies create different toolpaths. This makes it possible to examine the effect of the machining direction on the surface roughness, the shape accuracy and the profile error. Changes in cutting parameters have an effect on these properties too, which have been investigated. During the production of the test pieces, the milling forces were measured, from which important conclusions can also be drawn.

Keywords: free-form milling; ball-end milling; geometric tolerance; surface roughness; cutting force; CAM strategies

1 Introduction

The free-form surface can be encountered in many places in machine and tool design. The forming of car body elements, or injection mould tools contain free form surfaces, where the machining is challenging thanks to the high accuracy and productivity demands.

Based on the work of other researchers, the studied circumstances can be classified into seven different groups in the case of free-form surface milling.

CAD: The first problem is, how to design and describe the free-form surfaces. The development of CAD systems solve this problem, but the different file formats use different standard and non-standard descriptions. Ma et al. [1] presents the problems of construction and reconstruction of free form surfaces. Investigate the difference between the points of the real and the theoretical surface, the geometric error.

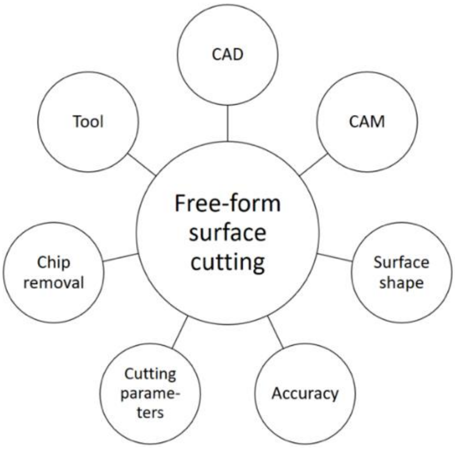


Figure 1
Investigation of milling of free-form surfaces

CAM: The free-form surfaces can be produced by using CAM systems. Different toolpaths can be created that affect the machining time and the quality of the machined surface. This aspect examines the possibilities offered by CAM systems. Different CAM strategies result different toolpaths, which defines productivity and the accuracy. Fountas and Vaxevanidis present the effect of different toolpaths on free-form surfaces [2]. Three different tool path strategies were compared, and the effect of the tool diameter and the stepover parameter were optimised considering the machining time and the surface deviation. Ižol et al. [3] compared different milling CAM strategies when milling free-form surfaces by examining the surface roughness and the machining time. Varga and Spisak study the effect of the different milling strategies to the form accuracy of a circular pocket in the case of aluminium alloy [4]. The value of the roundness deviation depends on the tool path strategy and the place (depth) of the measuring too. The different milling strategy has an effect not only on the cutting force and the surface roughness but on the tool wear also [5].

Surface shape: The geometry of the surface can be convex or concave which have influences on the machining process and the surface quality and accuracy. Huo et al. [6] determines the ideal toolpath based on the shape of the surface and its normal vector, for which it uses a fluid dynamics method. Käsemodel et al. [7] applies a CAD / CAM algorithm interface to study the shape of the free-form surface. This reduces cutting force, roughness and machining time.

Accuracy: The accuracy of the machined surface is examined from three aspects. The first is the dimensional accuracy: what is the size of the feature; the second is the geometric accuracy: how does the machined surface match the theoretical; and the third is the surface roughness. Fountas et al. studied shape accuracy while examining free-form surfaces [8]. There is a lot of research on mathematically predicting surface roughness. Seculic et al. [9] also investigated this using genetic and optimization algorithms.

Cutting parameters: The most frequently observed cutting parameters in the case of milling technology are the spindle speed (n), the cutting speed (v_c) the feed speed (v_f) and the feed rate (f), the depth of cut (a_p), and the width of cut (a_e). Wojciechowski et al. [10] examines the effect of variable input parameters through cutting force measurement. Abainia and Bey [11] optimizes the feed rate based on the expected value of the cutting forces in order to obtain the best possible surface quality.

Chip removal: The examination of the cutting force, the chip shape, belongs to this aspect. It is important to examine the milling forces that arise during chip removal by having a detailed examination of the chip removal process. Mou et al. also examine force components arising during cutting, similar to this article [12]. During the test, the three components of the cutting forces were examined. The magnitude of the cutting force also provides important information about the machining of the free-form surface. Beňo et al. [13] the factors influencing the conditions of chip separation were investigated using the Khattree-Niak multivariate method.

Tool: The coating and the number of edges of the tools, the tool wear is also an important aspect. It can greatly affect surface roughness. Scandiffio et al. [14] investigated the working part of the tool during machining. The working section of the tool is constantly changing because of the changing surface inclination. This parameter significantly affects the surface quality.

Based on the classification, several factors affect the characteristics of free-form surfaces. During the design of the manufacturing process, the parameters of the cutting technology must be determined. In addition, the equipment's availability must be taken into account, both in terms of the production and the quality control. Some machine parts are made of difficult-to-machine materials and have complex geometries, which complicate their manufacturing [15, 16]. The characteristics of each type of material have a large effect on the choice of machining methods, conditions, equipment, and the tools.

Geometric tolerances have an increasingly important role in the industry, defined and marked according to standards [17]. Such a standard (ISO 4287, ISO4288) also fixes the surface roughness, which is the micro-scale deviation of the surfaces. It defines a number of specifications and measurement details [18]. It may also be necessary to estimate the expected surface roughness when designing the machining process.

The accuracy of the free form surface has different aspects and several parameters have effect on it. In the current research the geometric, the dimensional accuracy, and the surface roughness are investigated in the case of ball-end milling of free form surfaces. We focus on the process planning and the application of CAM systems. In this paper, the effect of the different milling strategies on the surface roughness and the accuracy of the form surface is presented. Our aim is to support the work of CAM programmers in the selection of the most appropriate milling strategy.

2 Materials and Methods

The size of the test part was 80x80x30 mm, and the main feature is a cylindrical surface with a 45 mm radius, which is connected to a horizontal plane with a 10 mm radius (Figure 2). A concave (CV) and a convex (CX) part were created, in order to compare the effect of the nature of the surfaces. The height and the depth of the profile were 9.2 mm.

The material of the test parts was 42CrMo4 (1.7225; Rm = 1000 MPa) low alloy steel (Table 1), which is one of the chromium, molybdenum, manganese low alloy steel material with high fatigue strength and good low-temperature impact toughness. 42CrMo4 alloy steel is widely used for engineering steel purposes.

The machining was performed by a Mazak 410 A-II CNC machining centre. The tools were always fixed in an EMUGE-FRANKEN SK40 cold shrink clamp (powRgrip). Flooded cooling-rinsing lubrication (Aquamet 40, 6-8% emulsion, yield about 30 l/min) was used for the measurements. The CNC programs were generated by CATIA v5 CAD/CAM system. During the finishing the applied cutting tool was a Fraisa X7450.450 ball-end milling cutter with 10 mm diameter (D_c = 10 mm) and the number of teeth was 4 (z = 4).

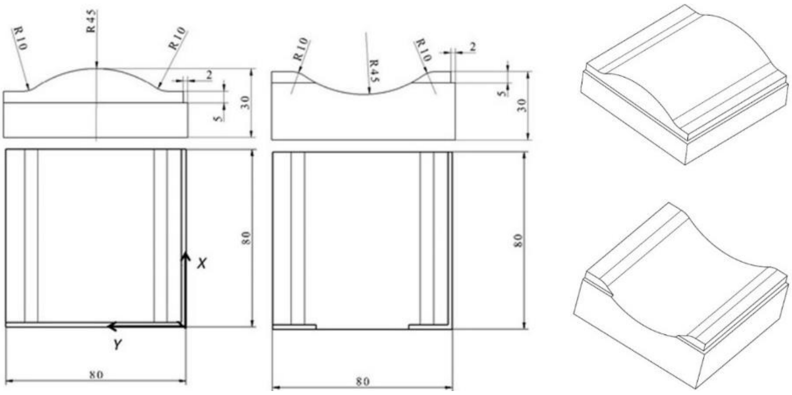


Figure 2
Convex (CX) and concave (CV) test part

The surface roughness was measured by Mahr Perten GD120 test instrument, which works with the contact method.

The Ra and Rz parameters were measured during the test. For a better comparison of the surface roughnesses, the measuring was performed only in x-direction in 7 different positions. Table 1 shows the angles of the surface normal vector.

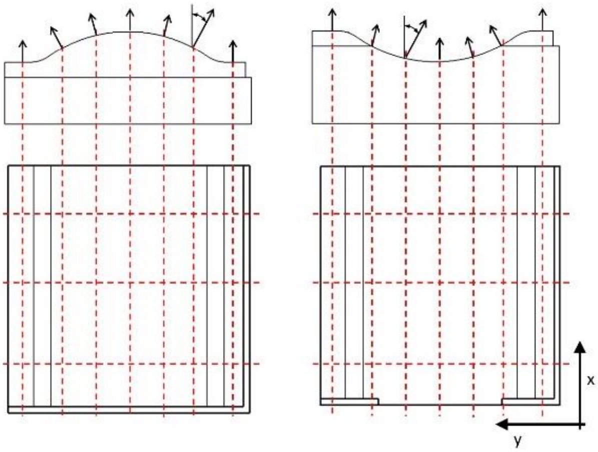


Figure 3
Measuring positions and normal vectors (° / 32.2° / 15.5° / 0° / - 15.5° / - 32.2° / 0°)

Table 1
Angles of the surface normal vector

Measuring points (y)	2	14	26	38	50	62	74
CX angles	0°	32.2°	15.5°	0°	-15.5°	-32.2°	0°
CV angles	0°	-32.2°	-15.5°	0°	15.5°	32.2°	0°
Absolute value of angles	0°	32.2°	15.5°	0°	15.5°	32.2°	0°

The current research focuses on the investigation of geometric tolerances of free-form surfaces created with a ball-end milling. Two types of geometric tolerances were studied, cylindricity (Cyl) and surface profile (SPE) tolerance. Based on the ISO 1101 standard the cylindricity error is the radial distance of the two cylinders (Figure 4), which have the same axis and covers the investigated surface. So the cylindricity describes the accuracy of the geometric shape, ignoring its position. The cylindricity tolerance can be applied only in the case of cylindrical surfaces. The surface profile error can also be applied to non-cylindrical surfaces. The surface profile tolerance zone is limited by two parallel envelope surfaces, which are parallel with the theoretical surface too.

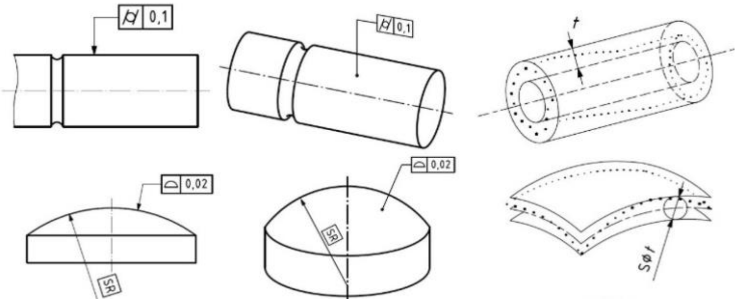


Figure 4
The definition of cylindricity and surface profile error based on ISO 1101

Surface points were measured by a Mitutoyo Crysta-Plus 544 coordinate measuring machine. 49 points were measured on the surface along a 7x7 grid. The evaluation of the geometric error was determined based on measured points by Evolve Smart Profile v6 software (Figure 5), which can compare the measured point to the theoretical CAD model of the test part. The nominal radius of the cylindrical surface was also measured during the test by the Evolve Smart Profile v6 software.

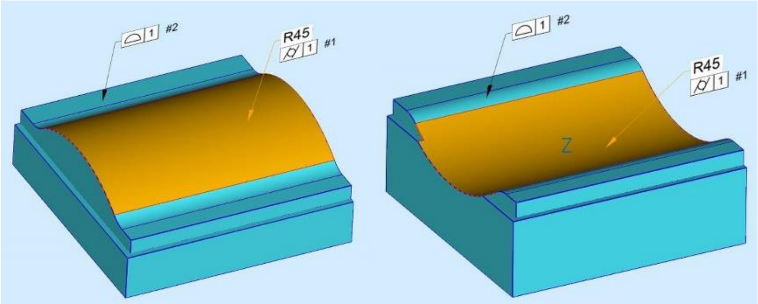


Figure 5
Definitions of geometric tolerances in Smart Profile and test parts

Cutting force measurement was performed during the cutting process of the second stage. KISTLER 5019 type, 3 component force measuring device was used. The data were evaluated by DynoWare software. The force measure was performed by 500 Hz frequency, so data was recorded in every 0.002 seconds. Considering the spindle speed, 0.002 seconds means 61.2° revolution of the tool.

2.1 Investigation of Various Tool Path Strategies

Two different test series were performed. In the first phase, the effect of the tool path strategies was investigated, and the surface roughness and the geometrical errors were compared. In the second phase, the effect of the cutting parameters was studied in the case of the selected toolpath strategy.

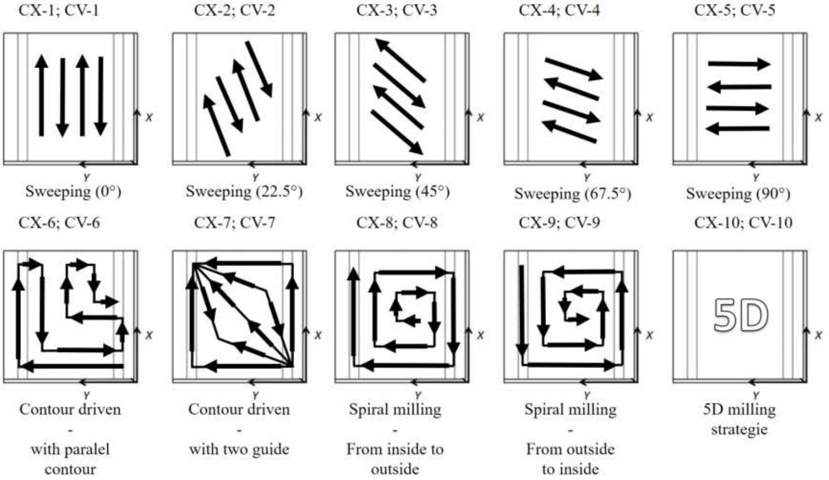


Figure 6
CAM strategies

During the first test series the cutting parameters were the same: cutting speed v_c = 160 m/min; spindle speed n = 5100 rpm; feed per tooth f_z = 0.08 mm; feed speed v_f = 1650 mm/min; the depth of cut a_p = 0.3 mm; the width of the cut a_e = 0.15 mm. The width of cut means the distance between the tool paths, and it has a critical effect on the surface quality.

The milling of the test surfaces was performed by different strategies, as sweeping (with variable sweeping direction), spiral milling (from inside to outside), spiral milling (from outside to inside), contour driven (with two guides), contour driven (with parallel contour) and one 5D milling strategy (Figure 6).

2.2 Investigation of Various Cutting Parameters

In the second phase, the effect of the cutting parameters was investigated. 8 more test pieces were produced, four convex (CX) and four concaves (CV). During the manufacture of the test pieces, the roughing and pre-finishing parameters were same. Table 2 shows the cutting parameters of the test. The feed per tooth and the width of cut were varied. Test parts No4 were add to the comparison. During the production of each new test piece, only one parameter changed, the other manufacturing parameters were the same.

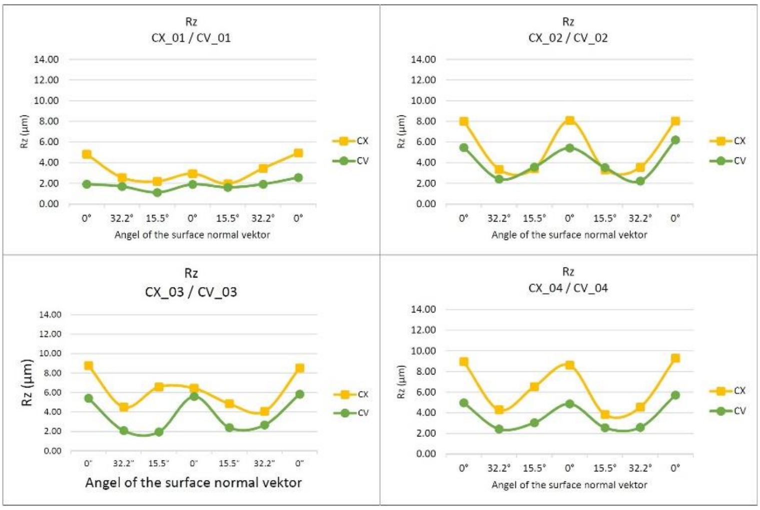
Table 2

Cutting parameters used in the test

Test part id.	CX-15; CV-15	CX-16; CV-16	CX-04; CV-04	CX-17; CV-17	CX-18; CV-18
Cutting speed v_c [m/min]	160				
Spindle speed n [rpm]	5100				
Feed per tooth f_z [mm]	0.08	0.08	0.08	0.12	0.16
Feed speed v_f [mm/min]	1630	1630	1630	2450	3260
Depth of cut a_p [mm]	0.3				
Width of cut a_e [mm]	0.35	0.25	0.15	0.15	0.15

3 Result and Discussion - Investigation of Various CAM Strategies

Based on the results of the presented milling tests, it can be stated that the surface normal vector (inclination of the surface) has an effect on the surface roughness. Surface roughness is worse on horizontal or near-horizontal surfaces. These surfaces are located at both ends and in the middle of the test pieces as the diagrams show (Figure 7). At these places, the value of the normal vector is 0 °. A "W" shape is drawn for each diagram. In some cases, the differences are smaller, but the character can be recognized. The measuring direction modifies the values because of the nature of the contact measuring method. The highest values are located where the milling direction is perpendicular to the measurement direction.



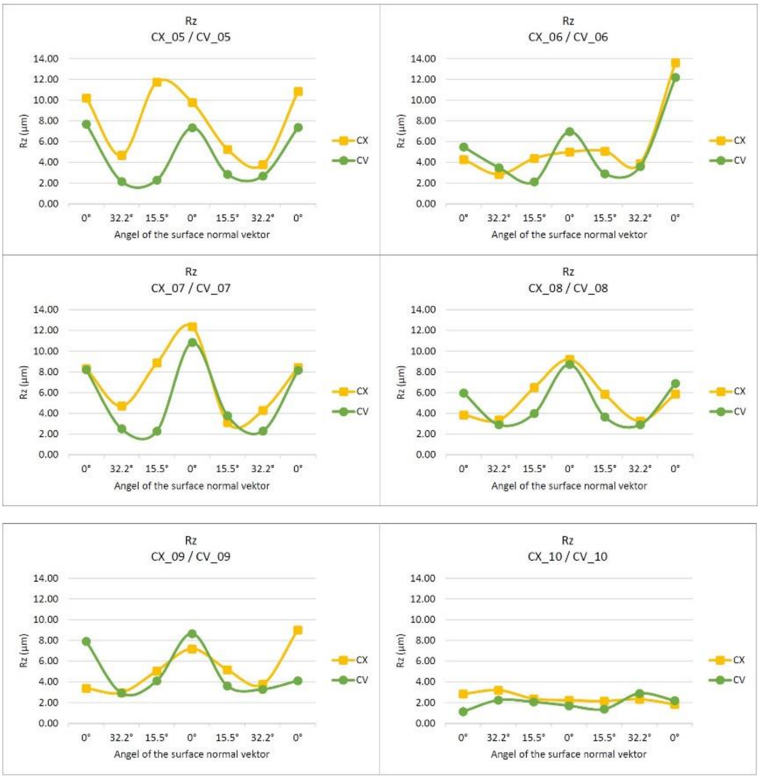


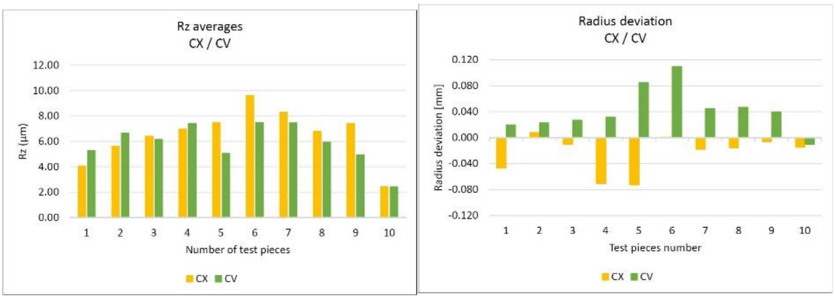
Figure 7

Rz surface roughness in different normal vectors (CV-CX 1-10)

The average surface roughness was calculated from all measured values. Figure 8a shows the average surface roughness of the test pieces. The diagram shows that for the first five test pieces that the change in sweeping direction affects the magnitude of roughness. Based on this, it can be stated that the direction of a sweeping strategy influences the surface roughness. From x-milling direction (0°) to y-direction (90°) the surface roughness increases in the case of concave test pieces. This observation wasn’t made for the concave pieces. It can also be seen that better roughness values are obtained for test piece No10. In this test piece, the concave and convex roughness values almost coincide.

The measured roughness value is influenced by a number of factors, such as the test piece geometry, the tool, the cutting parameters, the nature of the toolpath, the properties of the manufacturing machine, and the method of measuring the roughness.

The milling strategies have an effect on dimensional accuracy also. The dimensional accuracy is the part of macro-scale errors and means the error of size. Figure 8b shows the deviations of the machined surface from the nominal radius value (45 mm). Based on the obtained results, it can be said that the convex test pieces are below the nominal radius value, while the concave test pieces are above it. The five-axis machining does not fit in line in this case either (CV-10; CX-10). The reason for this is the constant position of the tool relative to the surface.



a) b)

Figure 8

Rz averages and radius for all data (CV1-10 – CX1-10)

The second type of macro scale error is the group of geometric errors. The values of the different geometric errors can be seen in Figure 9. The values of the cylindricity and surface profile error are very similar because of the similarity of the definitions. The first 5 test parts (CX1-5; CV1-5) show the effect of the feed direction. The investigated geometric errors are smaller when the milling is performed in x-direction (1), where the z-coordinate of the surface is the same. In the y-direction (5) the z-level of the tool path changes, other section of the tool is worked and it results in a larger error. The other milling strategies (6-10) result in smaller geometric errors. In all cases, the convex surfaces have better accuracy than the convex. The values of the cylindricity error are smaller because of the less number of freedom of the tolerance zone. The difference of the geometric error between the convex and concave parts is larger in the case of sweeping strategies (1-5). When combined strategies are used (6-9) the differences are smaller, but the 5D milling (10) ensures the smallest difference.

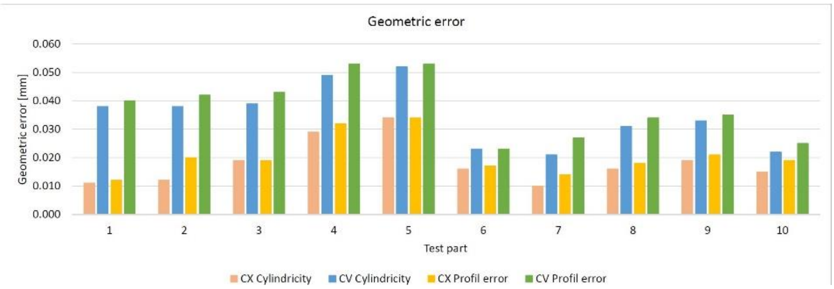


Figure 9

Values of geometric error (CV1-10 – CX1-10)

4 Result and Discussion - Investigation of Various Cutting Parameters

During the second phase, the effect of the cutting parameters were investigated. The CV-4 and CX-4 were the basis of the comparison. In the case of test pieces CV-15, CX15, CV16, CX16, the width of cut was changed (Table 2), and at two test pieces, the feed per tooth was changed (CV17, CX17, and CV18, CX18).

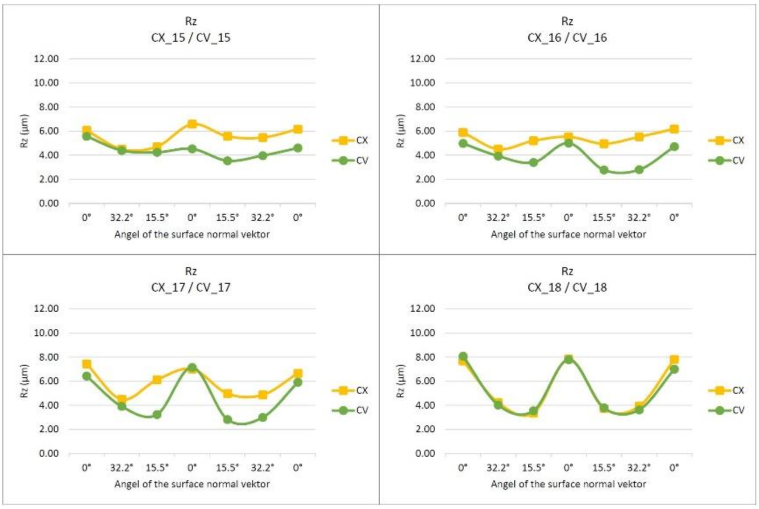


Figure 10

Rz surface roughness in different normal vectors (CV-CX 15; 16; 17; 18)

The "W" character of the diagrams of surface roughness can be recognised (Figure 10). Generally, the concave surfaces have smaller surface roughness because of the better insertion of the ball-end tool and surface radius.

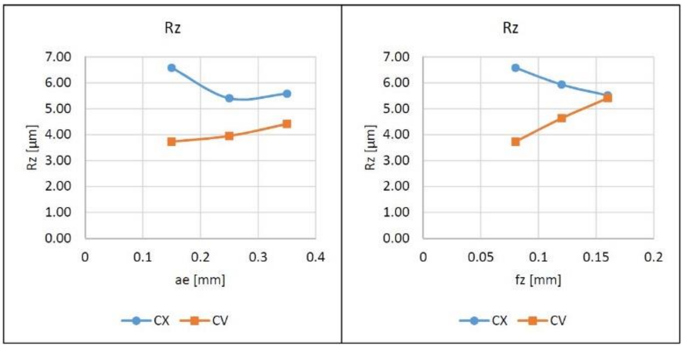


Figure 11

The effect of width of cut and feed per tooth on the surface roughness

The effect of the investigated cutting parameters on the micro and macro accuracy can be seen on the next figures. In the case of surface roughness (the Rz parameter is presented only) the parameters have an inverse effect in the case of convex and concave test surfaces (Figure 11). The surface roughness is increasing in the case of a concave surface parallel with the width of cut and feed per tooth. However, the Rz value of the convex surface decreases. This observation is remarkable. The concave surfaces have better roughness.

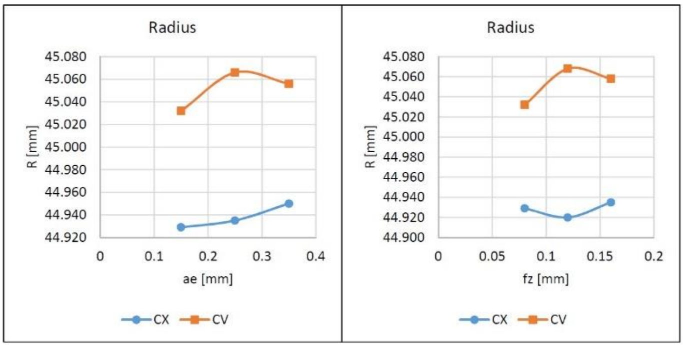


Figure 12
The effect of width of cut and feed per tooth on the radius

This inverse effect can be observed in the case of the value of the radius (Figure 12). The convex surfaces have a smaller radius and the width of cut increases. In the case of a concave surface, the largest a_e decreases the value. The feed per tooth initially decreases, later increases the radius of the convex surface. In the case of a concave surface, the value of the radius is larger, and it shows the inverse change. The changing of the radius due to the parameters is very small (0.02-0.03 mm).

The geometric errors show a very similar look. As Figure 13 shows, the cylindricity error is smaller in the case of a convex surface, but the difference decreases when the width of cut and the feed is larger. The cylindricality shows that concave test pieces are more inaccurate than convex ones especially in the case of test part No4 (ae = 0.15 mm; fz=0.075 mm). However, there is no significant difference between the other parts.

At the profile error (Figure 14) the parameters show similar tendencies, but changing of the error is smaller in the case of convex surfaces. The profile error values show that the error also decreases when the feed rate increases. However, the width of the cut does not significantly affect it. The value of the change is 0.01 mm in case of the convex parts and 0.03 mm in case of concave surfaces.

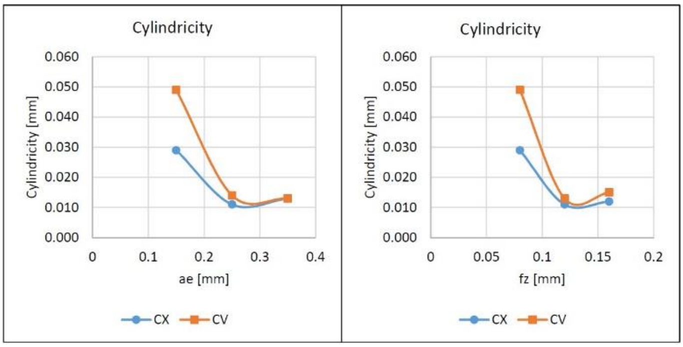


Figure 13

The effect of width of cut and feed per tooth on the cylindricity

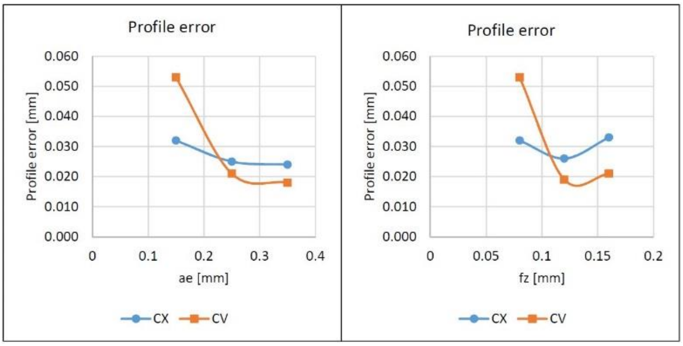


Figure 14

The effect of width of cut and feed per tooth on the profile error

The source of the geometric error is the load of the cutting tool and the deformation of the system. Three force components were measured based on the direction of the coordinate system, and the resultant cutting force was calculated (Figure 15).

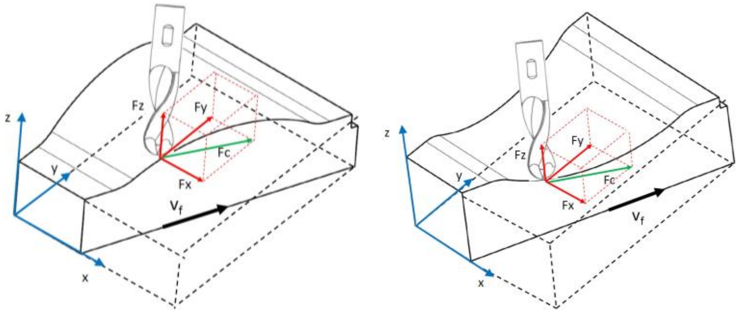
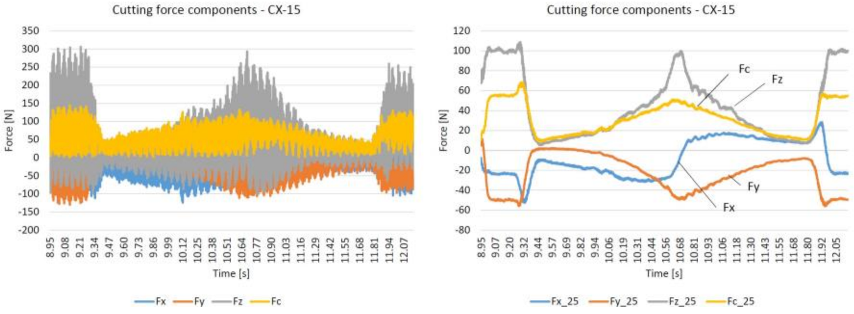


Figure 15

The cutting force and the force components in case of the convex and the concave parts

The measuring frequency was 500 Hz, which means 0.002 time resolution, but it means 61.2° revolution of the tool. Therefore, the changing of the values and the range of the force components are very large and fuzzy (Figure 16a). The 61.2° resolution means that in some cases there are no connection between the tool and the workpiece. In order to have better data processing a filtering function was used: the average values of 25 measured data were calculated and marked by the index '_25'. The 25 data cover 0.05 seconds. Figure 16b shows the filtered data of CX-15 part.

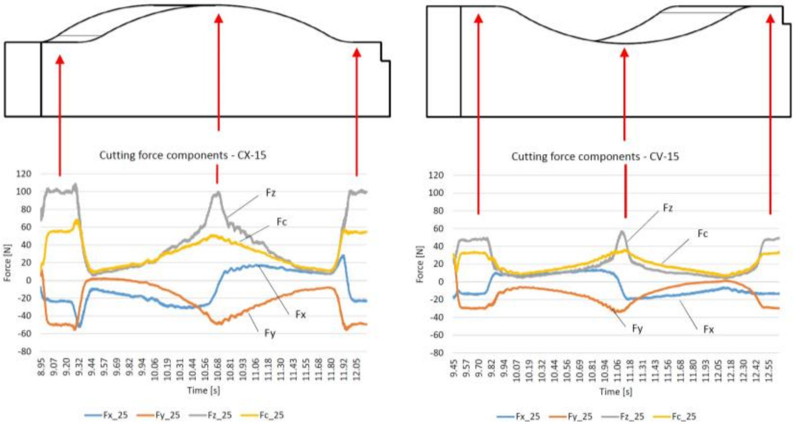


a) b)

Figure 16

The measured and filtered force values in case of CX-15

On Figure 17 the relationship between the force components and the character of the surface can be observed. The milling was performed left to right.



a) b)

Figure 17

The average value of the force components along the surface

The Fz component (axial direction of the tool) has maximum value at the horizontal sections. The decreasing of the Fz is very fast after the initial plate, and then a slow increase can be seen. The Fx component has a negative value at the beginning, and it changes to a positive at the top point in the case of a convex part. In the case of a concave part, the tendency is reversed. The Fy component there is in the negative region, it has a maximum value at the horizontal sections, but the dynamic of the changes is different at the two sides. Hereinafter only the resultant forces (Fc) will be compared, which is the vectorial sum of Fx and Fy components.

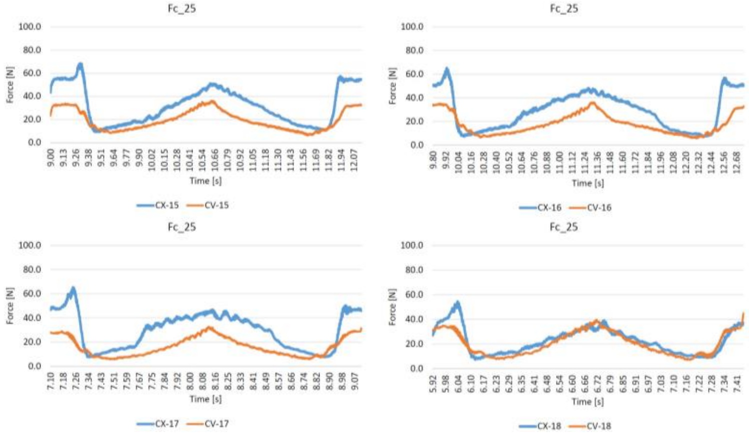


Figure 18

The resultant forces (Fc)

Figure 18 shows the Fc forces based on the average of 25 data in the case of the 8 specimens. The convex parts have a larger force, but the character of the curves is very similar. The average values were calculated of the Fc_25 resultant forces. In the case of concave parts (CV) the effect of the fz and ae is well recognised, the decreasing width of cut (a_e) decreases the Fc a little, but the increasing feed (f_z) increases the force on a larger scale. This tendency is similar to the effect of cutting parameters on the surface roughness. In the case of the CX-17, the average value of the cutting force is larger than expected because of the early force increase in the first section.

The changing resultant cutting force has an effect on the geometric error of the surface, as Figure 19 shows. In the case of a convex part the higher force causes a higher profile error. But the error map shows, that at the sides and in the middle region, where the cutting force has a maximum value, the error is positive, the measured surface there is above the theoretical surface. On the other hand, if the force is smaller, the profile error becomes negative, and the machined surface will be under-milled.

The effect of the sudden decreasing of the resultant force after the initial horizontal section can be seen on the error map too. The dark blue regions indicate large negative errors.

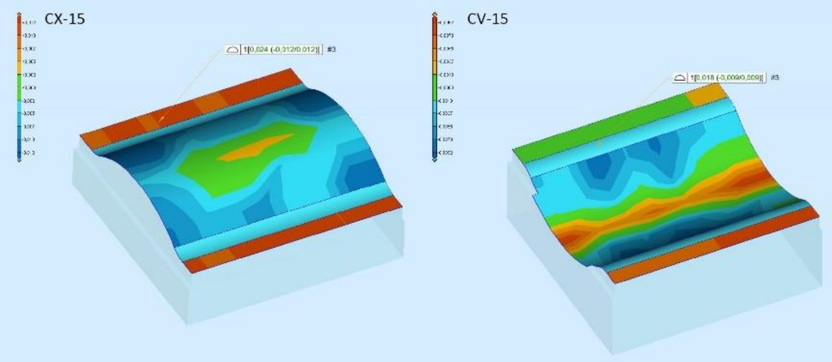


Figure 19

Profile error maps of CX-15 and CV-15

Conclusion

There are several ways to qualify manufactured parts. It can be classified on the basis of surface roughness, shape accuracy, and dimensional accuracy. From an industrial point of view, accuracy requirements are becoming increasingly important. The current research focuses on the examination of the CAM system-generated toolpaths and cutting parameters. The free-form test parts were machined by a ball-end milling tool. In the current paper, we examined the dimensional accuracy, the surface roughness, the geometric error, and the milling force. In the first half of the experiment, convex and concave test parts were machined using different toolpath strategies. In the second half of the experiment, the effect of the width of the cut (distance between the tool paths) and the feed were investigated.

Based on the results, it was found that the surface roughness, the dimensional error, and the geometric error represent different aspects of accuracy, but are not independent of each other.

In case of the test parts No1-5, the direction of the tool path between the x and y axes increased the surface roughness, but it is not affected by the nature of the surfaces (convex or concave). The surface roughness of the test parts No6-9 is similar to the previous five. Test part No10 hangs out of line, but that was to be expected, thanks to the 5-axis machining, when the cutting tool can be tilted during machining, and thus better cutting circumstances can be achieved.

Based on the obtained data, it can be seen that the surface roughness of the concave test parts is better than that of the convex pieces. This tendency was not affected by feed rate or cutting width. The surface roughness value is greatly influenced by the surface normal vector. Where the vector is parallel with the tool axis, there are the worst roughness values because of the ball-end mill and the test piece contact on the smallest surface and working diameter.

The investigation of the feed rate confirmed that the surface roughness deteriorates with increasing feed rate and the change in cutting width has a smaller effect.

The dimensional accuracy was tested by examining the radius of the cylindrical surface. The nominal value was 45 mm. The results showed that the radius of the concave test pieces is larger than the nominal value and in the case of convex test pieces, it was smaller. This is not observed in case of five-axis machining, where both test parts had smaller radius. This is due to the homogeneity of the machining.

The investigation of the cylindricity error suggests that concave pieces are more inaccurate in all cases than convex pieces. This suggests that the nature of the geometry greatly affects cylindricity error. The dimensional accuracy is more affected by milling direction than cutting parameters. The cutting width and the feed rate have no significant effect on the cylindricity.

The profile error was also larger for concave test parts, so the nature of the geometry has a large effect on this type of error. The direction of machining also greatly influences their value. For the convex and the concave test pieces, the deviation can be up to twice. However, it is not particularly affected by the feed rate and the cutting width. Overall, the milling direction has a greater effect on shape tolerance than the cutting parameters.

In the case of convex parts, the cutting force is higher, and at the horizontal surface sections, the force has maximum value. The size of the cutting width has little effect on the force. The change in feed rate requires further research. Examination of the y-direction forces shows that the tool requires a force in the opposite direction when traveling on an uphill or downhill path. The inflection point of the diagrams indicates this. In further research the deeper analyses of the cutting force and the force components is necessary. The effect of tool deformation on shape accuracy also requires further investigation.

The results are summarised as follows:

From the surface roughness point of view, the direction of the surface normal vector is the most influential factor. Five-axis machining is not better than three-axis machining in terms of surface roughness, the average value is higher, but the constant along the surface. It is also found that the surface roughness deteriorates with increasing feed rate.

In terms of shape accuracy, the milling direction has a greater effect on shape accuracy than the cutting parameters. And the convexity of the workpiece has a strong influence on the rollability.

In terms of machining force, it can be stated that higher cutting forces are required to machine convex parts.

References

[1] Ma W., He G., Han J., Xie Q.: Error compensation for machining of sculptured surface based on on-machine measurement and model reconstruction. The International Journal of Advanced Manufacturing Technology 106:3177-3187(2020) doi: 10.1007/s00170-019-04862-0

[2] Fountas N. A., Vaxevanidis N. M.: Intelligent 3D tool path planning for optimized 3-axis sculptured surface CNC machining through digitized data evaluation and swarm-based evolutionary algorithms. Measurement 158(2020) doi:10.1016/j.measurement.2020.107678

[3] Ižol P., Vrabel' M., Maňková I.: Comparison of Milling Strategies when Machining Freeform Surfaces. Material Sience Forum 862:18-25(2016) doi: 10.4028/www.scientific.net/MSF.862.18

[4] Varga J., Spišák E.: Influence of the milling strategies on roundness of machined surface. Acta Mechanica Slovaca 24(3):20-27(2020) doi: 10.21496/ams.2020.001

[5] Mali R. A., Aiswaresh R., Gupta T. V. K.: The influence of tool-path strategies and cutting parameters on cutting forces, tool wear and surface quality in finish milling of Aluminium 7075 curved surface. The International Journal of Advanced Manufacturing Technology 108(5):589-601(2020) doi: 10.1007/s00170-020-05415-7

[6] Huo G., Jiang X., Su C., Lu Z., Sun Y., Zheng Z., Xue D.: CNC tool path generation for freeform surface machining based on preferred feed direction field; International Journal of Precision Engineering and Manufacturing 20:777-790(2019) doi: 10.1007/s12541-019-00084-2

[7] Käsemodel R. B., de Souza A. F., Voigt R., Basso I., Rodrigues A. R.: CAD/CAM interfaced algorithm reduces cutting force, roughness, and machining time in free-form milling. The International Journal of Advanced Manufacturing Technology 107:1883-1900(2020) doi: 10.1007/s00170-020-05143-x

[8] Fountas N. A., Benhadj-Djilali R., Stergiou C. I., Vaxevanidis N. M.: An integrated framework for optimizing sculptured surface CNC tool paths based on direct software object evaluation and viral intelligence; Journal of Intelligent Manufacturing 30:1581-1599(2019) doi: 10.1007/s10845-017-1338-y

[9] Sekulic M., Pejic V., Brezocnik M., Gostimirović M. Hadzistevic M.: Prediction of surface roughness in the ball- end milling process using response surface methodology, genetic algorithms, and grey wolf optimizer algorithm. Advances in Production Engineering & Management; 13(1):18-30(2018) doi: 10.14743/apem2018.1.270

[10] Wojciechowski S., Maruda R. W., Barrans S., Nieslony P., Krolczyk G. M.: Optimisation of machining parameters during ball end milling of hardened

steel with various surface inclinations. Measurement 111:18-28(2017) doi:10.1016/j.measurement.2017.07.020

[11] Abainia S., Bey M.: Feedrate optimization for 3-axis sculptured surfaces finishing using flat-end tool. Journées de Mécanique de l'EMP (JM'11–EMP) (2018)

[12] Mou W., Zhu S., Zhu M., Han L., Jiang L.: A Prediction Model of Cutting Force about Ball End Milling for Sculptured Surface. Mathematical Problems in Engineering 1389718(2020) doi: 10.1155/2020/1389718

[13] Beňo J, Maňková I, Ižol P, Vrabel' M.: An approach to the evaluation of multivariate data during ball end milling free-form surface fragments. Measurement 84:7-20(2016) doi: 10. 1016/j.measurement.2016.01.043

[14] Scandiffio I., Diniz A. E., de Souza A. F.: Evaluating surface roughness, tool life, and machining force when milling free-form shapes on hardened AISI D6 steel. The International Journal of Advanced Manufacturing Technology 82(9):2075-2086(2016) doi: 10.1007/s00170-015-7525-0

[15] Kaya E., Akyuz B.: Effects of cutting parameters on machinability characteristics of Ni-based superalloys: A review. Open Engineering 7(1):330-342(2017) doi: 10.1515/eng-2017-0037

[16] Olufayo O. A., Che H., Songmene V., Katsari C., Yue S.: Machinability of Rene 65 superalloy. Materials 12(12):2034(2019) doi: 10.3390/ma12122034

[17] ISO 1101-2017 Geometrical product specifications (GPS) - Geometrical tolerancing - Tolerances of form, orientation, location and run-out

[18] Farkas G.; Drégelyi-Kiss Á.: Measurement uncertainty of surface roughness measurement. IOP Conference Series: Materials Science and Engineering 448(1):012020(2018) doi:10.1088/1757-899x/448/1/012020

Torque Quality Improvement of Switched Reluctance Motor Using Ant Colony Algorithm

Fahad Al-Amyal[1,2], Mahmoud Hamouda[1,3], László Számel[1]

[1] Budapest University of Technology and Economics, Department of Electric Power Engineering, Egry József utca 18, H-1111 Budapest, Hungary.
[2] Department of Computer Technical Engineering, College of Technical Engineering, The Islamic University, 54001 Najaf, Iraq.
[3] Mansoura University, Electrical Engineering Department, Elgomhouria Street, Mansoura 35516, Egypt

E-mail: fahad.alamyal@vet.bme.hu, m_hamouda26@mans.edu.eg, szamel.laszlo@vet.bme.hu

Abstract: The switched reluctance motors (SRMs) are gaining increasing interest in many industrial applications, including electric vehicles (EVs). However, their main drawback is the high torque ripple and noise. This paper presents an optimization-based method to improve the torque quality of SRM drives. The focus is on reducing torque ripple without complicating the control algorithm. The switching angles are optimized using a multistage ant colony algorithm (MSACA). The multistage algorithm provides a better search capability that fits appropriately with the high nonlinearities of SRMs. The finite element method (FEM) is employed to calculate the magnetic characteristics of the tested 8/6 SRM prototype. These characteristics are used within the MATLAB environment in the form of lookup tables to model the machine. The performance indices are calculated within the simulation model. Series of simulation results are included to show the effectiveness of the proposed control. Besides, experimental verification is also included to verify the theoretical findings.

Keywords: switched reluctance motor; finite element method; optimization; switching angles; ant colony algorithm

1 Introduction

With no windings or magnets on the rotor and concentrated windings on the stator, the switched reluctance motors (SRMs) have the simplest structure of all electrical machines. They can provide reliable, less maintenance, and low-cost variable speed drives. These interesting features made them powerful alternatives to be employed for electric vehicles and the aviation industry [1]-[4]. However, the double salient structure makes their magnetic characteristics highly nonlinear functions of rotor position and current. Besides, the inherited torque ripple is the main blocking factor for the acceptance of SRM drives in high-performance applications.

Many strategies have been carried out to reduce the torque ripple of SRMs. They can be classified into two categories: machine design optimization and the other based on control algorithms. Although the design optimization of SRM, such as geometry-optimization, multiphase-machine, and winding configuration, can reduce torque ripple [2]-[3], but after the machine has been manufactured, it is difficult and complicated to modify its geometric dimensions. The alternative way to improve the torque quality of an existing machine is by using advanced control techniques [4]-[9]. Noting that the more complicated the algorithms are, the high-cost and low-performance the drive is. Therefore, a trade-off should be adopted to have the best overall performance. A simple control method is the best to be chosen. Hence, further improvements have to be proposed for torque ripple reduction.

The control parameters for SRMs involve the switch-on (θ_{on}), switch-off (θ_{off}) angles, and reference current (i_{ref}). The reference current is defined by the outer loop control. Therefore, the switching angles (θ_{on}, θ_{off}) are the dominant parameters for SRM control [10]-[13]. Proper estimation of θ_{on} and θ_{off} must be done accordingly with motor speed and loading torque. This is a very complicated task due to the highly nonlinear characteristics of SRMs. In [14], the switching angles are employed to reduce torque ripple and improve drive efficiency. Online tuning of turn-on angle is proposed, while from the flux linkage waveforms of two neighboring phases, a mathematical formula for the optimal turn-off angle is achieved. This method does not fit for traction application as the operating point is continuously changing. Besides, the formulation uses the linear inductance profile that fits a limited speed range. In [15], efficiency optimization is achieved in steady-state operation by the fine-tuning of firing angles through an algorithm that minimizes the drive's input power. In [16], an online optimization scheme is introduced to determines the optimal turn-on and turn-off angles to provide maximum efficiency. In [17], automatic control of the turn-off angle is presented in the face of automatic turn-on angle control, aiming to maximize the torque per ampere ratio. In [18], The effects of the switch-on and switch-off angles on motor torque, copper loss, and torque ripple are investigated, and a multi-objective function is developed to maximize the average output torque, average torque per RMS value of the phase current, and torque smoothness factor. In [19], analytical formulations are developed for both the switch-on and switch-off angles based on the fitting of phase inductance over the minimum inductance zone. In [20], closed-loop control is introduced for the switch-on angle aiming basically to reduce torque ripples and copper losses. First, closed-loop control is proposed for the optimum switch-on angle. Then, the switch-off angle is optimized using a one-dimensional search algorithm. In [5], a simple structure average torque control is introduced. The switching angles are defined based on a described searching algorithm. The algorithm uses a fixed step resolution that does not guarantee the best outcome. In [21], the switching angles are optimized mainly for torque production improvement, ignoring the generated torque ripple. For EV applications, there are basic requirements that have to be fulfilled by any drive system. These requirements include the minimum torque ripple, especially at low speeds. Besides, minimum losses and high efficiency are also of great interest. Therefore, this paper focuses on achieving the vehicle requirements based on the SRM drive. The focus is to achieve the minimum torque ripple, minimum losses, and maximum drive efficiency. Hence, a trade-off is done to achieve this goal. A modified version of the Ant Colony Algorithm (ACA), called Multistage Ant Colony Algorithm (MSACA), is introduced and implemented to optimize the switching angles. The modification is achieved to fit appropriately with the high nonlinearity in magnetic characterestics of SRMs. The objective is to maximize the average torque (means maximum efficiency), minimize net torque ripple (means minimum torque ripples), and minimize copper losses simultaneously. A simple control algorithm is adopted. Besides, a high-fidelity machine model is built and employed within the optimization problem using the finite element method (FEM). This paper is organized as follows: Section II includes machine modeling. Section III involves the formulation of the optimization problem and the proposed MSACA. Also, the simulation and the experimental results are presented in Sections IV and V, respectively. Lastly, the conclusions.

2 Machine Modelling

Due to the doubly salient structure of SRMs, their magnetic characteristics are functions of both the current (i) and the rotor position (θ). The voltage and torque equations are given by equations (1) and (2), respectively [1].

$$V = iR + \frac{d\lambda(i,\theta)}{dt}, \quad \therefore \lambda(i,\theta) = \int (V - iR)\, dt \tag{1}$$

$$T = \frac{1}{2} i^2 \frac{dL(i,\theta)}{d\theta} \tag{2}$$

where V is the phase voltage, R is the phase resistance. $\lambda(i,\theta)$ is the flux linkage. $L(i,\theta)$ is the phase inductance.

As the analytical model [22]-[24] does not provide an efficient representation of machine characteristics, the FEM is employed to accurately model the highly nonlinear magnetic characteristics of the tested 8/6 SRM prototype. The dimensional parameters of SRM are given in Table 1. The FEM obtained torque, flux-linkage, and inductance characteristics of the SRM are presented in Figure 1. For verification purposes, these data are presented along with their corresponding experimentally measured data. As seen, an excellent agreement is observed, and therefore, the FEM-calculated can be adopted to build a trusted SRM simulation model.

Table 1

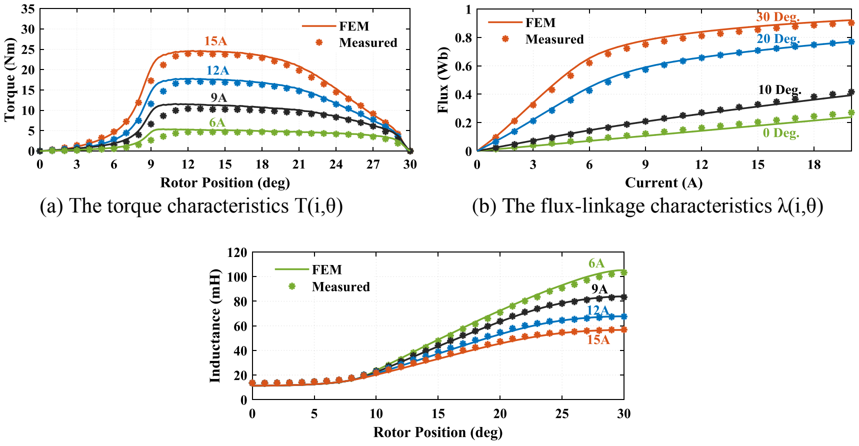
The design data of 8/6 SRM

Parameter	Value	Parameter	Value
Stator/ Rotor poles	8/6	Stator outer diameter	179.5 mm
Number of phases	4	Shaft / Bore diameters	36/96.7 mm
Phase resistance	0.642 Ω	Rotor/stator pole arc	21.5°/20.45°
Output power	4 kW	Height of rotor/stator pole	18.1/29.3 mm
Rated speed	1500 rpm	Air-gap length	0.4 mm
Turns per pole	88	Stack length	151 mm

The measurement is done first by applying a pulsed DC voltage to one phase winding of SRM at a known rotor position (θ). Then, the phase voltage and current are measured and recorded to calculate the phase flux linkage using equation (3) [25].

$$\lambda(i,\theta) = \int \left(v_{ph} - Ri_{ph}\right) dt \tag{3}$$

where v_{ph}, i_{ph}, and R are the phase-voltage, phase-current, and phase-resistance, respectively. $\lambda(i,\theta)$ is the calculated flux-linkage.
The torque is measured directly by the torque transducer. The measurement procedure is repeated several times at different rotor positions.



(a) The torque characteristics T(i,θ)

(b) The flux-linkage characteristics λ(i,θ)

(b) The inductance curves $L(i,\theta)$

Figure 1

The FEM-calculated and the corresponding measured magnetic characteristics

The complete, steady-state model of one phase of SRM is shown in Figure 2. The inputs are the reference current, motor speed, and switching angles. The outputs are the phase current and torque. A hysteresis controller is used to control the phase current.

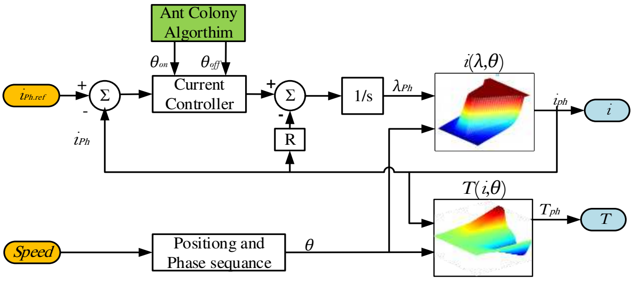


Figure 2

One phase model of 8/6 switched reluctance motor

The performance indices are estimated within the simulation model. They involve the average torque (T_{av}), torque ripple (T_r), average supply current (I_{av}), root mean square of supply current (I_{RMS}), and efficiency (η). These indices are calculated as follows:

$$T_{av} = \frac{1}{\tau}\int_0^{\tau} T_e(t)dt \tag{4}$$

$$T_r = \frac{T_{max}-T_{min}}{T_{av}} \tag{5}$$

$$I_{av} = \frac{1}{\tau}\int_0^{\tau} i_s(t)dt \tag{6}$$

$$I_{RMS} = \sqrt{\frac{1}{\tau}\int_0^{\tau} i^2(t)dt} \tag{7}$$

$$\eta = \frac{\omega\, T_{av}}{V_{DC}\, I_{av}} \tag{8}$$

where τ is the time of one electric cycle, T_{max} and T_{min} are the maximum and minimum values of instantaneous torque. i_s is the instantaneous supply current. ω is the motor speed.

3 The Optimization Problem

The optimization problem consists of two control variables, a multi-objective function, and two constraints. The SRM Simulink model in Figure 2 will be utilized within the optimization problem. Where, after specifying the desired input speed and current, the remaining inputs are the switch-on, and the switch-off angles will be considered as the control variables of the optimization problem. On the other hand, from the output torque and current waveform of the Simulink model, the performance indices in (4), (5), and (7) can be calculated. The proposed objective function combines two criteria: the average output torque (4) and the net torque ripple (5). A maximum allowed RMS value of the phase current (7) is specified as a constraint in this problem in order to maintain the motor efficiency with an acceptable range. The upper and lower boundary of the conduction angle is also constrained because, in 8/6 SRM, each phase should contact at least 15°.

Furthermore, a too wide a conduction angle may result in a considerable amount of negative torque generation. The focus is to locate optimal switch-on and switch-off angles to make the motor generate the highest possible average output torque with the minimum torque ripple at each operating point of the torque/speed profile. And therefore, improve the torque quality of the machine in a wide speed range.

3.1 Problem Formulation

The formulation of the optimization problem is as shown:

Find the best value of the control variables: θ_{on} and θ_{off}

In order to:

$$\text{Maximize: } F_{obj} = T_{ave} - w\,(T_{max} - T_{min}) \tag{9}$$

subjected to

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \theta_{on} \\ \theta_{off} \end{bmatrix} \leq \begin{bmatrix} -\frac{60^\circ}{m} \\ \frac{60^\circ K}{m} \end{bmatrix}, \tag{10}$$

$$I_{RMS} \leq max(I_{\mathrm{RMS}}) \tag{11}$$

where F_{obj} is the proposed objective function, T_{av} is the average torque (4). w is a constant weighting factor used to magnify the net torque ripple value (T_{max} -T_{min}). m is the number of phases, which is four in our case, and 60° mechanical degree corresponds to a complete 360° electrical cycle, K is a constant between (1 and 1.85) for adjusting the upper boundary of the conduction angle (θ_{off} - θ_{on}), different values of K are used at different operating points.

Generally, many algorithms have been presented and carried out by previous researchers to solve optimization problems in different fields of study. These algorithms can be categorized as conventional or meta-heuristics algorithms. The traditional algorithms like the interior point method, newton-raphson method, linear programming, or lambda iterations are complex and require high-level mathematical formulations, especially if the optimization problems are highly nonlinear. On the other hand, the meta-heuristics algorithms are inspired by nature and considered powerful tools to solve multi-objective optimization problems. Among these algorithms: the genetic algorithm (GA), particle swarm optimization (PSO), migration algorithm (MA), grey wolf optimizer (GOA), hybrid swarm algorithm, conflict monitoring optimization, slime mold algorithm (SMA), and the ant colony algorithm (ACA) [26]-[32]. These algorithms are easier to implement than the conventional analytical-based techniques and have given very good results over the years. In this paper, a modified version of the ACA is adopted to solve our problem since it can provide outstanding solution quality with short computational time and guaranteed convergence. However, in ACA, the convergence time is uncertain, but since the optimization results are obtained offline, the convergence time does not matter. The ACA and the modified MSACA are illustrated in the following sections.

3.2 Ant Colony Algorithm (ACA)

Regardless of their members' simplicity, ant colonies are exceptionally cooperative societies that can accomplish challenging tasks that are long way past single ant capacities [33]. The ACA is influenced by the cooperative methodology of the actual ant colonies, where a kind of indirect contact helps them to find the best path from the home to any food source. The ACA method can be exhibited by the guide of the graph that appeared in Figure 3.

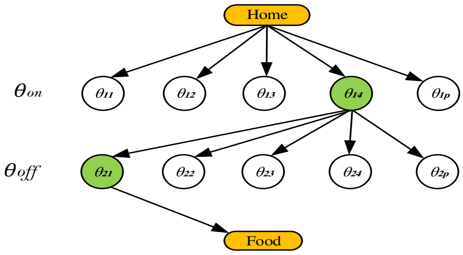


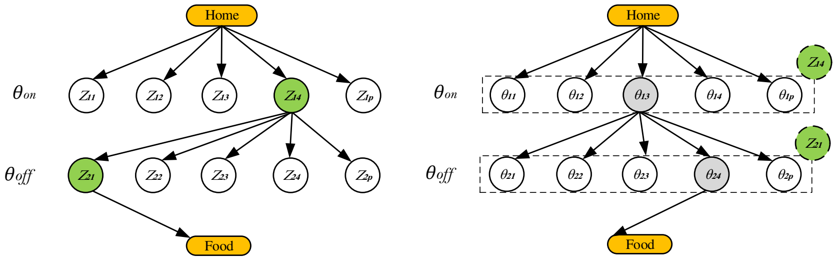
Figure 3

Graphical representation for the ACA

Every raw in the graph corresponds to a controlled variable within the optimization problem. In our case, the first raw corresponds to the switch-on angle while the second raw corresponds to the switch-off angle. In each raw, there are a set of random values within the allowed domain of the control variables called nodes or particles (p or θ_{ij}). Various N ants in the colony will begin looking for food starting from home, traveling via all rows until they arrive at the food source. Through its journey, each ant selects only one particle among the θ_{ij} particle from each row. The group of particles that each ant has selected will represent a single path or a candidate solution, such as the green path (θ_{14}, θ_{21}). And therefore, at each journey (iteration), the number of candidate solutions equal to the number of ants in the colony. Once each journey is finished, every ant store an amount of pheromone on the pre-selected path. The quantity of the stored pheromone is proportional to the preference of the corresponding path. In other meaning, if the selected path gives a better value of the proposed objective function (9), then the ant will store a higher amount of pheromone on that path and vice versa. Whenever the ants return to the food source and start searching again for food, the stored pheromone at each particle will be updated, either increased for particles that are frequently selected by the ants, or evaporated by pre-defined rate from the particles belonging to unvisited paths. Consequently, the probability of these particles being chosen by the approaching ants will increment or decrement relying upon the stored pheromone. In the end, the set of particles stored the largest pheromone amount is nothing but the optimal solution. In contrast, the pheromone of other particles will be totally evaporated. The optimization process will be terminated either if all ants select the same path or the maximum allowed iterations number is reached. In this work, a modified version of the ACA is employed.

3.3 Multistage Ant Colony Algorithm (MSACA)

A large number of particles (*p*) is essential to make a precise selection of the switching angles. However, increasing *p* will badly affect the search capability of the ACA. Because before the start of the iteration cycle, all particles share the same stored amount of pheromone and, therefore, an equal selection probability. As a result, the chance of choosing the optimal particle amongst a large number of particles will decrease, and due to the evaporation rate of the pheromone, the probability of selecting the optimal solution will further decrease during the optimization procedures, consequently, in some cases, the ants will probably fail to find the optimal solution. Since an accurate selection of the optimal switching angles is the main goal of this paper, a modified version of the ant colony algorithm is necessary. In this version, the simple ACA is repeated in two completely separated stages. In the first stage, the algorithm uses zones (Z) rather than discrete particles (*p*). After locating the optimal zones for the switch-on and switch-off angles in the first stage, the ACA will search only around and within these zones in the second stage to find a more accurate solution. The idea is to eliminate the need for a higher number of *p* by narrowing the search domain in the second stage. Figure 4 (a, b) shows the graphs of the first and second stages of the proposed multistage ant colony algorithm (MSACA). More stages can be used to give higher accuracy. However, the converged time will increase. The proposed MSACA is used in [34] to solve an Economic Emission Dispatch (EED) with six control variables. The results obtained in [34] compared favorably with corresponding results obtained from the GA.



(a) Graph representing the 1st stage (b) Graph representing the 2nd stage

Figure 4
Graphical representation for the MSACA

3.4 The Solution Scenario

The detailed optimization procedures are illustrated in the following steps:

Step 0. Input the desired speed and current commands to the SRM Simulink model and specify the search limit of the switching angles as shown in (12).

$$Limits = \begin{bmatrix} \theta_{on\_min} & \theta_{on\_max} \\ \theta_{off\_min} & \theta_{off\_max} \end{bmatrix} \tag{12}$$

Step 1. Initiate the MSACA parameters: number of ants (*N*), number of particles (*p*), pheromone evaporation rate (ρ), number of stages of the MSACA, and the number of allowed iterations in each stage.

Step 2. Start the first stage of the MSACA and formulate the search domain as in figure 3, within limits specified in step 0.

$$\text{Search domain } = \begin{bmatrix} \theta_{on\_min} & \theta_{12} & \theta_{13} & \dots\dots\dots & \theta_{on\_max} \\ \theta_{off\_min} & \theta_{21} & \theta_{31} & \dots\dots\dots & \theta_{off\_max} \end{bmatrix}_{2\times p} \tag{13}$$

Step 3. Initiate an equal amount of pheromone $\left(\mathcal{F}_{ij}\right)$ in all the θ_{ij} particles.

$$\text{Pheromone} = \begin{bmatrix} \mathcal{F}_{11} & \mathcal{F}_{12} & \mathcal{F}_{13} & \dots\dots\dots & \mathcal{F}_{1p} \\ \mathcal{F}_{21} & \mathcal{F}_{22} & \mathcal{F}_{23} & \dots\dots\dots & \mathcal{F}_{2p} \end{bmatrix}_{2\times p} \tag{14}$$

Step 4. Calculate selection probability of each particle ($P_{\,ij}$) using (15)

$$P_{\,ij} = \frac{\mathcal{F}_{ij}}{\sum_{j=1}^{p} \mathcal{F}_{ij}} \quad \begin{cases} i = 1\text{: number of control variables} \\ \quad j = 1\text{: number of particles} \end{cases} \tag{15}$$

Step 5. Perform roulette wheel selection process for each ant, so that each ant selects only one particle from each raw of the search domain, for example, ant number 1 (N^1) select θ_{17} from the first raw and θ_{23} from the second raw. Save all selected solutions by each ant, as shown in Table 2.

Table 2

Random candidate solutions performed by the ACA

Ant number	N^1	N^2	N^3	N^4	N^5	…………	N^N
Selected θ_{on}	θ_{17}	θ_{16}	θ_{11}	θ_{13}	θ_{15}	…………	θ_{12}
Selected θ_{off}	θ_{23}	θ_{21}	θ_{29}	θ_{28}	θ_{24}	………….	θ_{17}

Step 6. Run the Simulink model in Figure 2 for each of the above columns and obtain and save the output torque and current waveforms for each ant.

Step 7. From the stored torque and current waveform, calculate the phase current's RMS value (7), the average torque (4), the maximum and minimum value of the output torque. Check the constraint in (11) and obtain the objective function value from (9).

Step 8. Mark the ant that gave the best objective function value as N^{best} and update the pheromone $(\mathcal{F}_{ij})$ in all particles using (16)

$$\mathcal{F}^{new}{}_{ij} = \begin{cases} \mathcal{F}^{old}{}_{ij} \times 2 & for\ \ particles\ belong\ to\ \ N^{best} \\ \mathcal{F}^{old}{}_{ij} \times (1-\rho) & for\ other\ particles \end{cases} \tag{16}$$

The pheromone evaporation rate ρ has been chosen between (0.7 and 0.9) in this work.

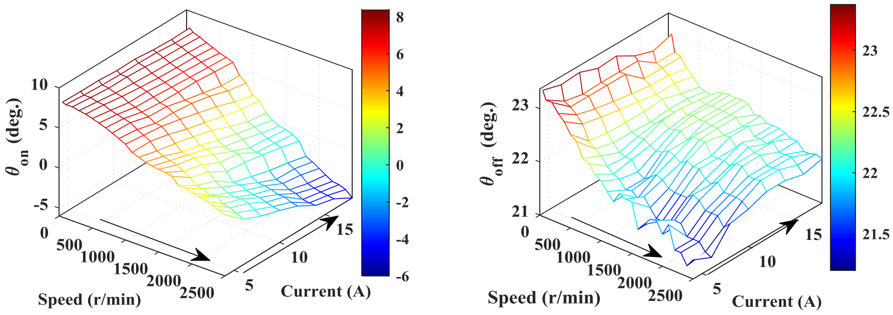
Step 9. Check if all ants selected the same solution, then terminate the first stage of the MSACA and go to step 10. If not, repeat from step 4 until the maximum iteration number has been reached.

Step 10. Mark the optimal results of the first stage as $\theta^{best}{}_{on}$ and $\theta^{best}{}_{off}$ and formulate a new smaller domain around these results by updating the old searching limits as shown in (17)

$$Limits = \begin{bmatrix} \theta^{new}{}_{on\_min} & \theta^{best}{}_{on} & \theta^{new}{}_{on\_max} \\ \theta^{new}{}_{off\_min} & \theta^{best}{}_{off} & \theta^{new}{}_{off\_min} \end{bmatrix} \tag{17}$$

The new minimum and maximum limits in the above search domain are the left and right neighbors of the best particles in the old search domain. In this way, the search space becomes smaller, and the results will be more accurate. Repeat the ACA and search within the new domain by starting from step 3 all the way to step 9.

In this work, the number of ants has been chosen in the range (25-30 ants) while the number of particles in the range (7-11 particles). Two or three stages are adopted. Figure 5 shows the optimized switch-on and switch-off angles for each combination of speed and reference current. As noted in Figure 5 (b), the switch-off angles have a small band of variation compared to the switch-on angles.



(a) Switch-on angles

(b) Switch-off angles

Figure 5

Lookup tables of optimized switching angles

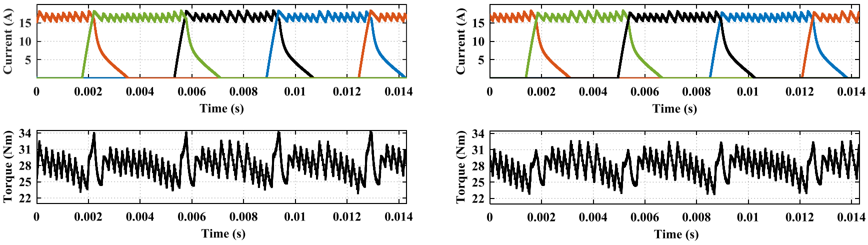
These MSACA-optimized switching angles are stored in lookup tables and adopted as inputs to the Simulink model of the SRM, and the corresponding Simulink results are presented in the next section.

4 Simulation Results and Discussion

In order to show the feasibility and effectiveness of the proposed control method, a comparison with the literature is achieved. Ref [20] is chosen for the comparison because it focuses mainly on reducing torque ripple and copper losses. The results involve steady-state and dynamic comparisons as follows.

4.1 The Steady-State Performance

Figures 6 (b), 7 (b), and 8 (b) show the obtained steady-state current and torque waveforms when applying MSACA-optimized switching angles to the Simulink model in Figure 2 at a speed of 700 r/min, 1500 r/min, and 2400 r/min, respectively. While, Figures 6 (a), 7 (a), and 8 (a) give the corresponding results when applying a switch-on angle that forces an intersection between the last peak of the outgoing phase current and the first peak of the incoming phase current [20]. As noted, the proposed method has a better torque profile with minimum torque ripples.

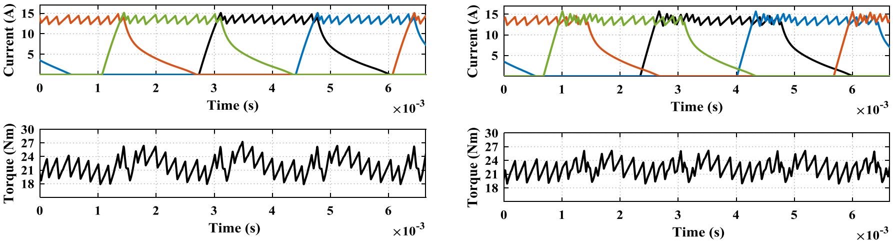


(a) Ref [20]

(b) The proposed

Figure 6

Steady-state simulation results for 700 r/min speed and 16 A commanded current



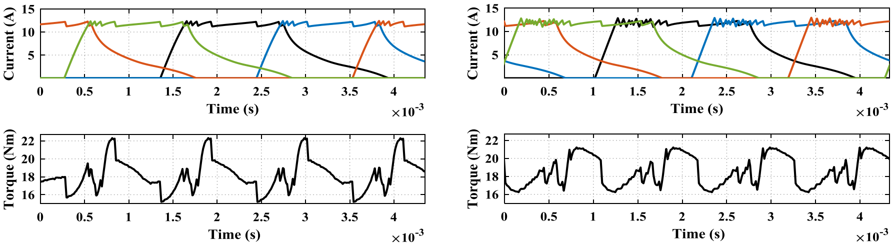
(a) Ref [20]

(b) The proposed

Figure 7

Steady-state simulation results for 1500 r/min speed and 14 A commanded current

The detailed analysis of steady-state results is given in Table 3. As seen in the table, for the same speed and the same reference current, both methods have almost the same average torque and a comparable value of efficiency. The proposed method has a significant reduction in torque ripple. However, it possesses a bit lower efficiency. However, it shows the best overall performance.



(a) Ref [20]

(b) The proposed

Figure 8

Steady-state simulation results for 2400 r/min speed and 12 A commanded current

Table 3

Optimized parameters

Algorithm	ω (r/min)	i_{ref} (A)	θ_{on} / θ_{off} (°)	T_{av} (Nm)	T_{rip} (%)	η (%)
Ref [20]	700	16	5.31 / 22.1	28.13	40.3	92.3
Proposed			5.69 / 22.1	28.01	35.1	93.2
Ref [20]	1500	14	3.62 / 21.95	21.9	40.7	96.2
Proposed			0.11 / 21.95	22.35	31.5	95.5
Ref [20]	2400	12	2.58 / 21.77	18.2	38.17	97.4
Proposed			-2.1 / 21.77	18.6	27.23	96.8

4.2 The Dynamic Performance

For the dynamic performance investigation, the results (in Figure 5) are implemented within the proposed closed-loop speed controller of SRM, as shown in Figure 9.

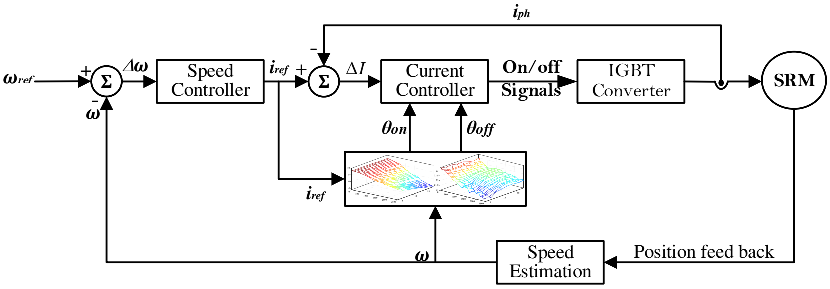
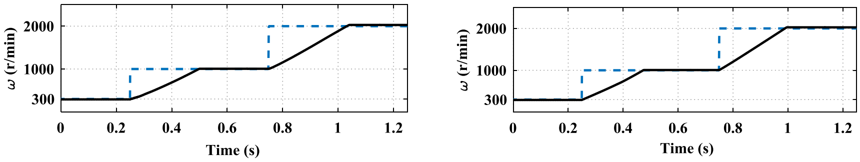


Figure 9

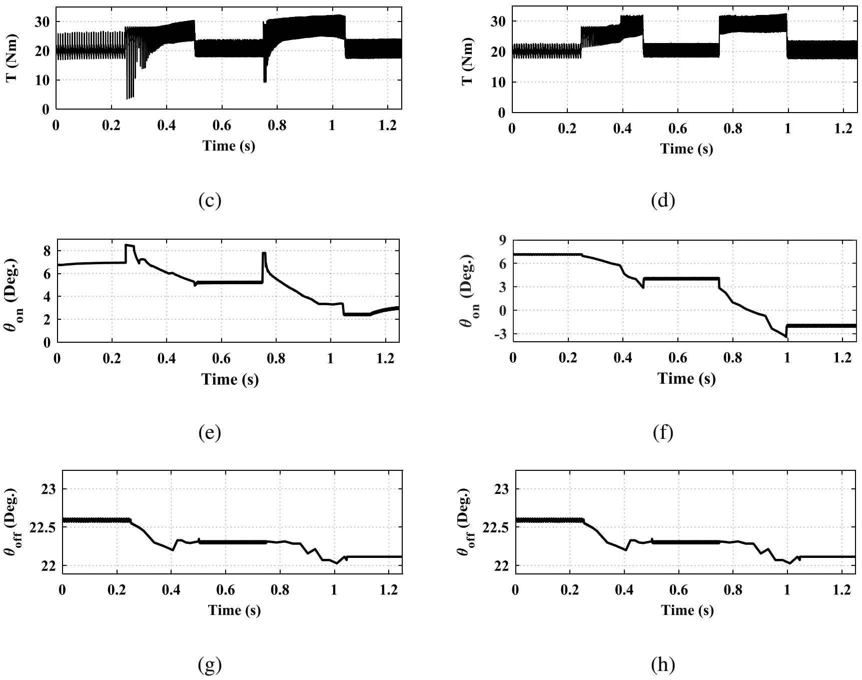
Overview of the proposed closed-loop speed controller for 8/6 SRM

Figure 10 shows the dynamic torque waveform and the variations of the switching angle at constant load torque of 20 Nm and different speeds. It can be seen that the dynamic torque performance in Figure 10 (d) is much superior to in Figure 10 (c), especially at transit conditions. This is because in [20], the switch-on angle is tuned online using a PI controller, and therefore, this controller will require a certain time period before it achieves the final value of the switch-on angle. This delay will badly affect the dynamic performance during transit periods.



(a)

(b)



(c) (d)

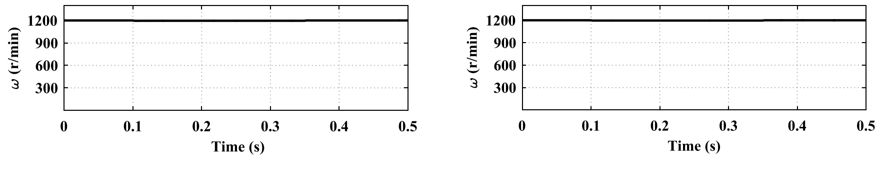
(e) (f)

(g) (h)

Figure 10

The dynamic performance at constant load torque of 20 Nm and different speeds

Figure 11 shows the dynamic torque waveform and the variations of the switching angle when the motor runs at a constant speed (1200 r/min) and subjects to a sudden loading torque of 20 Nm at 0.1 sec, as seen in Figure 11 (c, b). As noted, the proposed method has better torque dynamics with a better profile and reduced ripple, especially at transient moments. Besides, the variation of switching angles is shown in Figure 11(e, f). It is observed that the proposed control changes the switching angles instantaneously according to the operating point.



(a) (b)

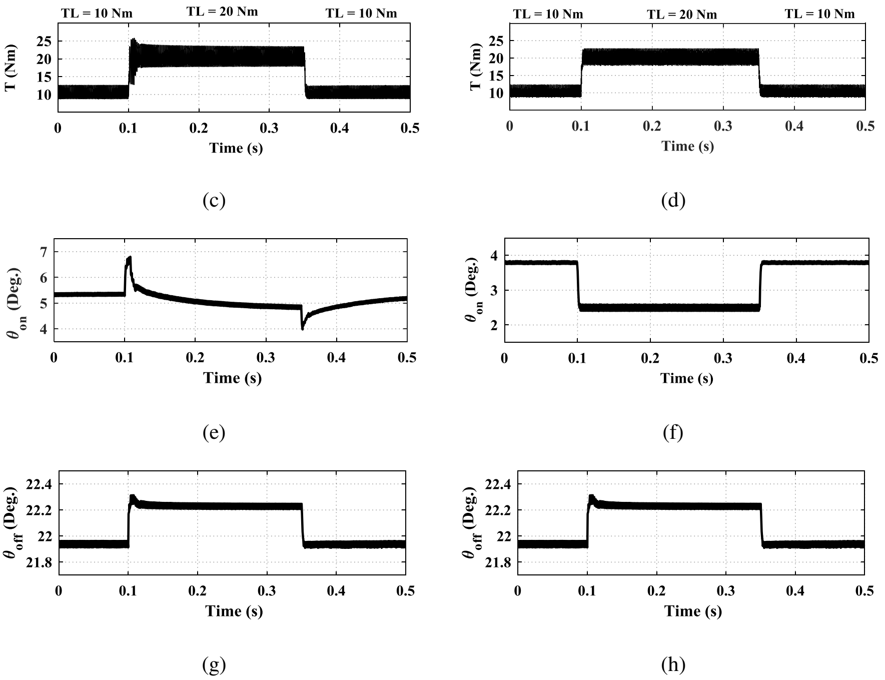


Figure 11

The dynamic performance at 1200 r/min when the motor is subjected to a load torque disturbance from 10 Nm to 20 Nm to 10 Nm

5 Experimental Verification

The experimental test-bed is shown in Figure 12 is constructed to validate the theoretical findings. The SRM is coupled mechanically with an electromagnetic brake. The drive system involves an IGBT asymmetrical bridge converter, current transducers, TMS320F28335 control board, 1024 PPR incremental encoder, data acquisition board (DAQ NI USB-6009).

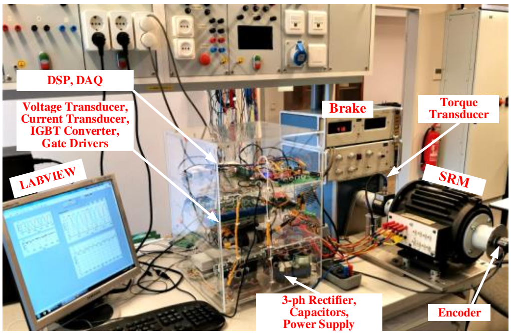
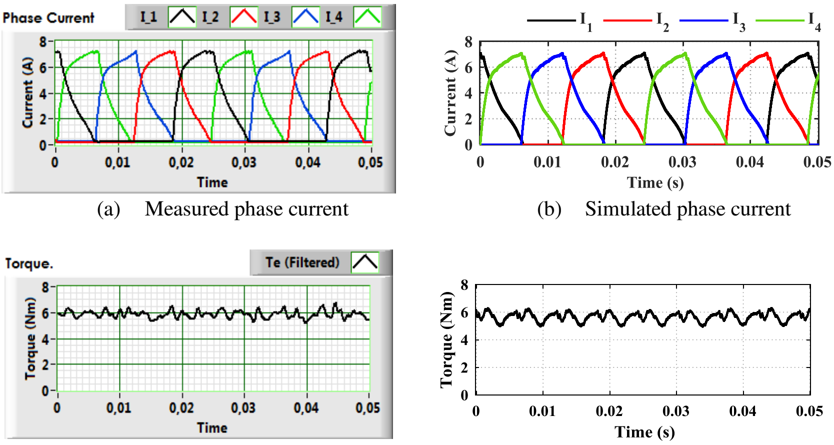

DSP, DAQ
Voltage Transducer, Current Transducer, IGBT Converter, Gate Drivers
LABVIEW
Brake
Torque Transducer
SRM
3-ph Rectifier, Capacitors, Power Supply
Encoder

Figure 12
The experimental testbed

5.1 Model Verification

The accuracy of FEM-calculate results has been previously verified in Figure 1. However, for more demonstrations, the accuracy of the dynamic simulation model is verified here in Figure 13. In this figure, a comparison between the simulated motor waveforms and the experimentally measured ones is given. The comparison involves the dynamic waveforms for phases' current and total electromagnetic torque. An excellent agreement is observed that reflects the model accuracy.



(a) Measured phase current

(b) Simulated phase current

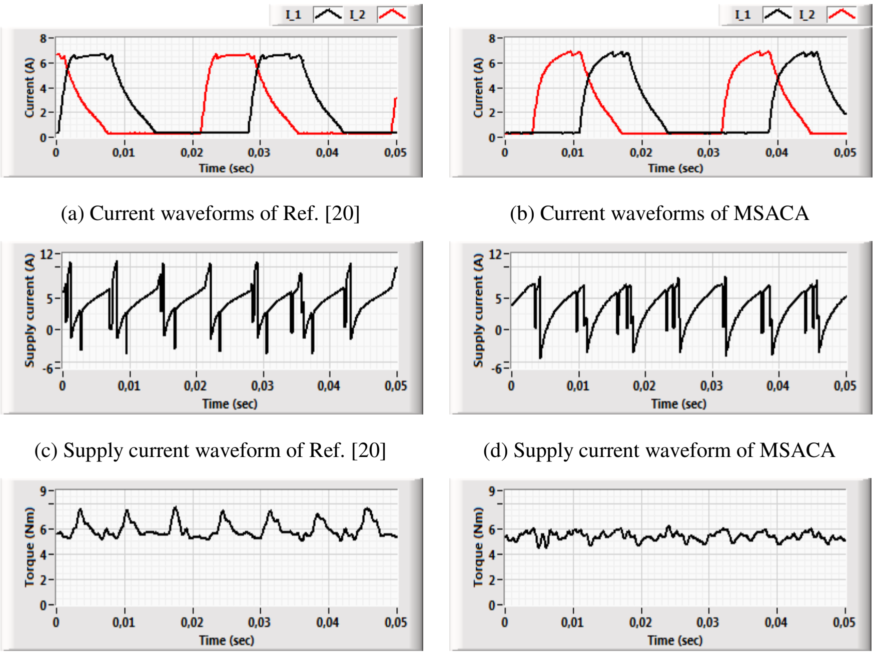
(c) Measured torque waveform

(b) Simulated torque waveform

Figure 13
Comparison between the measured and simulated current and torque curves at a speed of 410 r/min and 7A commanded current

5.2 Experimental Comparative Analysis

Figures 14 and 15 show a quantitative comparison between the proposed control method and Ref. [20]. Figures 14 give the comparison under the low speed of 360 r/min. As noted, the proposed MSACA shows the best torque profile. It also draws a lower supply current and hence lower copper losses. The detailed quantitative analysis is seen in Table 4. The proposed MSACA has the lowest torque ripple of 29.8% compared to 40.3% for Ref. [20]. Besides, the proposed MSACA provides the best efficiency and the higher torque to current ratio.



(a) Current waveforms of Ref. [20]

(b) Current waveforms of MSACA

(c) Supply current waveform of Ref. [20]

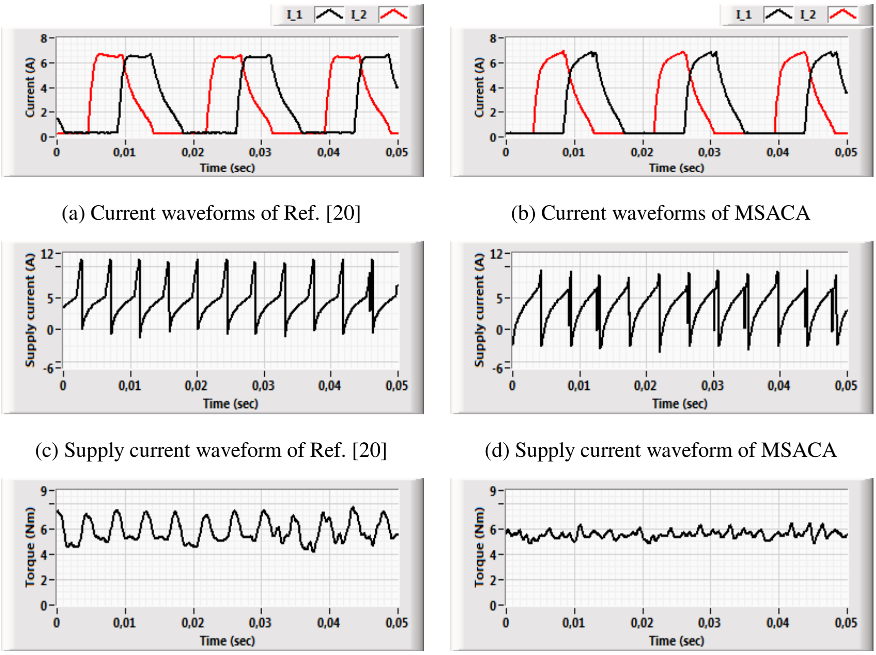
(d) Supply current waveform of MSACA

(e) Torque waveform of Ref. [20]

(f) Torque waveform of MSACA

Figure 14
Experimental results for 360 rpm speed and 5.6 Nm load torque

Figures 15 give the comparison under a higher speed of 570 r/min. As noted, the proposed MSACA shows the best torque profile. The proposed MSACA has the lowest torque ripple of 25.5% compared to 50.1% for Ref. [20]. Besides, the proposed MSACA provides the highest torque to current ratio. However, it possesses a bit lower efficiency.



(a) Current waveforms of Ref. [20]

(b) Current waveforms of MSACA

(c) Supply current waveform of Ref. [20]

(d) Supply current waveform of MSACA

(e) Torque waveform of Ref. [20]

(f) Torque waveform of MSACA

Figure 15
Experimental results for 570 rpm speed and 5.4 Nm load torque

Table 4
A comparison between Ref. [20] and MSACA parameters

Algorithm	Speed (r/min)	i_{ref} (A)	θ_{on}/θ_{off} (Degree)	T_{ave}/I_{RMS} (Nm/A)	T_{ripple} (%)	η (%)
Ref. [20]	360	6.3	5.6/ 22.5	1.07	40.3	65.1
MSACA			7.7/ 23.5	1.16	29.8	68.4
Ref. [20]	570	6.3	5.3/ 23.05	1.245	50.1	79.2
MSACA			6.8/ 23.05	1.285	25.5	77.6

Conclusion

This paper is an optimization-based method to provide the best overall performance of SRM. A trade-off is done to achieve the minimum torque ripple, the highest average torque, and the minimum losses. A multistage ACA (MSACA) is developed to fit properly with the high nonlinearity of SRM. The MSACA is employed to estimate the optimum switching angles over the entire torque-speed profile. The machine model is achieved accurately using FEM. Besides, the FEM-calculated characteristics are validated through a comparison with the measured ones. Moreover, the final dynamic simulation model is verified in a closed-loop control for the best validation. Both the simulation and the experimental results showed the effectiveness of the proposed MSACA based optimization problem. The proposed controller showed the lowest torque ripples and the highest torque to current ratio. It also gave the best efficiency under low speeds and a very good efficiency under high-speed ranges.

References

[1] B. Bilgin, J. W. Jiang, and A. Emadi: Switched Reluctance Motor Drives, CRC Press, 2019, pp. 1-794

[2] Z. Yueying, Y. Chuantian, Y. Yuan, W. Weiyan, and Z. Chengwen: Design and optimisation of an In-wheel switched reluctance motor for electric vehicles, Intelligent Transport Systems, 2019, Vol. 13, No. 1, pp. 175-182

[3] B. Anvari, H. A. Toliyat, and B. Fahimi: Simultaneous Optimization of Geometry and Firing Angles for In-Wheel Switched Reluctance Motor Drive, IEEE Transactions on Transportation Electrification, 2018, Vol. 4, No. 1, pp. 322-329

[4] R. Abdel-Fadil, F. Al-Amyal, and L. Szamel: Torque ripples minimization strategies of switched reluctance motor - A review, International IEEE Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE), Budapest, Hungary, 2019, pp. 41-46

[5] M. Hamouda and L. Szamel: Reduced torque ripple based on a simplified structure average torque control of switched reluctance motor for electric vehicles," International IEEE Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE), Budapest, Hungary, 2018, pp. 109-114

[6] H. Li, B. Bilgin, and A. Emadi: An Improved Torque Sharing Function for Torque Ripple Reduction in Switched Reluctance Machines, IEEE Transactions on Power Electronics, 2019, Vol. 34, No. 2, pp. 1635-1644

[7] M. Hamouda, Q. S. Ullah, and L. Szamel: Compensation of Switched Reluctance Motor Torque Ripple based on TSF Strategy for Electric Vehicle Applications, International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET), Islamabad, Pakistan, 2018, pp. 1-6

[8] R. M. Abdel-Fadil and L. Számel: Direct Instantaneous Torque Control of the Switched Reluctance Motor for Electric Vehicles Applications Using Fuzzy Logic Control, Acta Tech. Jaurinensis, 2019, Vol. 12, No. 2, pp. 101-116

[9] R. Abdel-Fadil and L. Számel: Predictive control of switched reluctance motors for aircraft electrical actuators applications, Acta Polytech. Hungarica, 2020, Vol. 17, No. 5, pp. 209-227

[10] M. Debouza, A. Al-Durra, H. M. Hasanien, S. Leng, and W. Taha:

Optimization of switched reluctance motor drive firing angles using grey Wolf optimizer for torque ripples minimization, 44th Annual Conference of the IEEE Industrial Electronics Society (IECON), Washington, DC, USA, 2018, pp. 619-624

[11] J. W. Jiang, B. Bilgin, and A. Emadi: Three-Phase 24/16 SwitchedReluctance Machine for a Hybrid Electric Powertrain, IEEE Transactions on Transportation Electrification, 2017, Vol. 3, No. 1, pp. 76-85

[12] Z. Yueying, Y. Chuantian, and Z. Chengwen: Multi-Objective Optimization of Switched Reluctance Generator for Electric Vehicles, 21st International Conference on Electrical Machines and Systems (ICEMS), Jeju, Korea (South), 2018, pp. 1903-1907

[13] M. Hamouda, A. Abdel Menaem, H. Rezk, M. N. Ibrahim, and L. Számel: Numerical Estimation of Switched Reluctance Motor Excitation Parameters Based on a Simplified Structure Average Torque Control Strategy for Electric Vehicles, Mathematics, 2020, Vol. 8, No. 8, p. 1213

[14] C. Mademlis and I. Kioskeridis: Performance optimization in switched reluctance motor drives with online commutation angle control, IEEE Transactions on Energy Conversion, 2003, Vol. 18, No. 3, pp. 448-457

[15] B. Blanqué, J. I. Perat, P. Andrada, and M. Torrent: Improving efficiency in switched reluctance motor drives with online control of turn-on and turn-off angles, European Conference on Power Electronics and Applications, Dresden, Germany, 2005, pp. 9 pp.-P.9

[16] I. Kioskeridis and C. Mademlis: Maximum efficiency in single-pulse controlled switched reluctance motor drives, IEEE Transactions on Energy Conversion, Vol. 20, No. 4, pp. 809-817

[17] Y. Sozer and D. A. Torrey: Optimal turn-off angle control in the face of automatic turn-on angle control for switched-reluctance motors, IET Electric Power Applications, 2007, Vol. 1, No. 3, pp. 395-401

[18] X. Xue, K. Cheng, J. Lin, Z. Zhang, K. Luk, T. W. Ng, N. Cheung: Optimal control method of motoring operation for SRM drives in electric vehicles, IEEE Transactions on Vehicular Technology, 2010, Vol. 59, No. 3, pp. 1191-1204

[19] Y. Z. Xu, R. Zhong, L. Chen, and S. L. Lu, "Analytical method to optimise turn-on angle and turn-off angle for switched reluctance motor drives, IET Electric Power Applications, 2012, Vol. 6, No. 9, pp. 593-603

[20] A. Shahabi, A. Rashidi, M. Afshoon, and S. M. Saghaian Nejad: Commutation angles adjustment in SRM drives to reduce torque ripple below the motor base speed, Turkish J. Electr. Eng. Comput. Sci., 2016, Vol. 24, No. 2, pp. 669-682

[21] M. Hamouda and L. Számel: Optimum control parameters of switched reluctance motor for torque production improvement over the entire speed range, Acta Polytech. Hungarica, 2019, Vol. 16, No. 3, pp. 79-99

[22] S. S. Ahmad and G. Narayanan: Linearized Modeling of Switched Reluctance Motor for Closed-Loop Current Control, IEEE Transactions on Industry Applications, 2016, Vol. 52, No. 4, pp. 3146-3158

[23] W. Uddin and Y. Sozer: Analytical Modeling of Mutually Coupled Switched Reluctance Machines under Saturation Based on Design Geometry, IEEE Transactions on Industry Applications, 2017, Vol. 53, No. 5, pp. 4431-4440

[24] D. S. Mihic, M. V. Terzic, and S. N. Vukosavic: A New Nonlinear Analytical Model of the SRM with Included Multiphase Coupling, IEEE Transactions on Energy Conversion, 2017, Vol. 32, No. 4, pp. 1322-1334

[25] M. Hamouda and L. Számel: Accurate magnetic characterization based model development for switched reluctance machine, Period. Polytech. Electr. Eng. Comput. Sci., 2019, Vol. 63, No. 3, pp. 202-212

[26] Z. Abdmouleh, A. Gastli, L. Ben-Brahim, M. Haouari, and N. A. Al-Emadi: Review of optimization techniques applied for the integration of distributed generation from renewable energy sources, Renewable Energy, 2017, Vol. 113, pp. 266-280

[27] J. Vaščák: Adaptation of fuzzy cognitive maps by migration algorithms, Kybernetes, 2012, Vol. 41, No. 3/4, pp. 429-443

[28] R. E. Precup, R. C. David, E. M. Petriu, A. I. Szedlak-Stinean, and C. A. Bojan-Dragos: Grey Wolf Optimizer-Based Approach to the Tuning of PiFuzzy Controllers with a Reduced Process Parametric Sensitivity, IFACPapersOnLine, 2016, Vol. 49, No. 5, pp. 55-60

[29] R. C. Roman, R. E. Precup, C. A. Bojan-Dragos, and A. I. Szedlak-Stinean: Combined Model-Free Adaptive Control with Fuzzy Component by Virtual Reference Feedback Tuning for Tower Crane Systems, Procedia Computer Science, 2019, Vol. 162, pp. 267-274

[30] H. Zapata, N. Perozo, W. Angulo, and J. Contreras: A Hybrid Swarm Algorithm for Collective Construction of 3D Structures, International Journal of Artificial Intelligence, 2020, Vol. 18, No. 1, pp. 1-18

[31] M. Moattari and M. H. Moradi: Conflict Monitoring Optimization Heuristic Inspired by Brain Fear and Conflict Systems, International Journal of Artificial Intelligence, 2020, Vol. 18, No. 1, pp. 45-62

[32] R.-E. Precup, R.-C. David, R.-C. Roman, E. M. Petriu, and A.-I. SzedlakStinean: Slime Mould Algorithm-Based Tuning of Cost-Effective Fuzzy Controllers for Servo Systems, International Journal of Computational Intelligence Systems, 2021, Vol. 14, No. 1, pp. 1042-1052

[33] Marco Dorigo and Thomas Stützle, Ant Colony Optimization. The MITPress, 2004

[34] F. Al-Amyal, K. J. Al-Attabi, and A. Al-Khayyat: Multistage Ant Colony Algorithm for Economic Emission Dispatch Problem, International IEEE Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE), Budapest, Hungary, 2019, pp. 161-166

Master Manipulator with a Limitless Angular Displacement

Paweł Żak

Institute of Machine Tools and Production Engineering, Lodz University of Technology, Stefanowskiego 1/15, 90-924 Lodz, Poland; pawel.zak@p.lodz.pl

Abstract: In the presented paper a survey of currently used popular master manipulators solutions is given, pointing their advantages and disadvantages in the process. It is pointed that all widely available devices of such type are burdened with motion limitations which affect their utilization greatly. As a solution to this problem, a novel construction of a master manipulator device is being proposed. The design process of the device itself is being discussed, but also the description of a control system is presented along with the control algorithm and filtration method (Kalman filter). The publication is being concluded with the description of experiments performed along with results obtained.

Keywords: master manipulator; construction; control system

1 Introduction

In the robotic field, a great number of commonly used solutions are telemanipulators [1], which main feature is that instead of replaying earlier created control program, they need to be controlled by the operator constantly to reproduce his arm motions. Such an approach is mainly to extend human capabilities, e.g. increase strength [2], provide an ability to teach the operator some typical motions by slightly guiding his hand [3], and so on. Another application of such an approach is to replace humans to ensure their safety while working in a hazardous environment, e.g. under high temperature or radiation [4], or in any type of harsh environment [5], or the combat zone [6]. The number of applications is not limited to above of course – a good example of using manipulator are medical applications. Cardio surgical robots, like daVinci [7], or its alternatives – RobIn Heart [8] or HeroSurge [9] can be pointed here. The idea behind using such solutions is to aid a surgeon during the operation – laparoscopic procedures can last for many hours, during which the personnel need to remain standing still, often in some uncomfortable position, while with robotic system help, the surgeon can sit down next to the control console, which obviously provides more comfort. Also, by using such a robot, the surgeon doesn't have to be sited next to the robot – there was an experiment conducted with transatlantic surgery [10], which only shows the potential of using such solutions.

To create a well-working control system that would provide full control of the operation field and would not limit the operator's awareness of the environment, it is necessary to equip the robotized system with a set of sensors to create a feedback loop to the operator, especially force feedback which "transfers" touch sense from the machine to operating hand [11]. Such a system enables the surgeon to, e.g. use the palpation examination method without using his fingers, also the facture, rigidness, etc. of the tissues can be checked by touching it with the robot [12]. Another possibility to use such a system is to slightly guide the surgeon's hand while performing some typical gestures [13], like sewing. It can be either used to control the correctness of performed procedures – on a basis of a database created during several earlier treatments, or it can serve as a simulator and training utility for students during their medical studies [14].

The last thing to be taken care of while creating a telemanipulator system is a proper construction master manipulator, also called a haptic device, which serves as a bridge between the operator and a manipulator. Such device should be characterized by high ergonomics – it should not limit operator's movements at any time, as such limitation breaks the fluency of work [15] – hand's motion is being stopped without any significant (from operator's point of view) reason after what some kind of clutch needs to be pressed to move the handle back to the center of a workspace.

There is a great number of commercially available solutions of haptic devices. Also, the problem of a limited workspace is well-known. Therefore, the manufacturers of master manipulators try to solve it using a variety of methods, yet none of them provide an ultimate solution to the problem. They tend to make the operator unable to reach such unwanted configurations of the haptic devices. There are three commonly used methods the designers use to avoid reaching the haptic device workspace range.

The first method is to create the master manipulator in form of an exoskeleton – perfectly fitted to the operator's arm. The solution uses the operator's motion limitations – he is not able to move anywhere away from his reach, therefore no artificial boundaries appear during the operation.

A clear example of such a device is Able [16]. The device was created by the Haption company. The device is being offered in the market and can be used in a variety of different applications. Another example is Active A-Gear presented in [17]. This device is used for rehabilitation purposes, not to control anything, yet it possesses the features of haptic devices created under this principle. Another example a 7-dof haptic device is described in [18]. Its only purpose is to control medical robots.

All the above examples show that besides the advantage pointed out earlier, such devices also have some disadvantages. The main disadvantage is the lack of the universality of such a solution – each device needs to be customized for the operator before it can be used to provide the assumptions given earlier in this section. Also, such a solution should not be fixed to the operator's back, as it is supposed to be light, therefore, such a system requires a lot of space and cannot be moved easily if needed.

Another possibility is to create a haptic device with a heavily extended workspace which limits are far beyond the operator's reach. Such a solution has been proposed i.e. by Haption in Scale1 device [19].

In this case, the device requires a frame with rails that supports the grip held by the operator. The fact is that he can walk freely around the room as the grip will follow him fluently, therefore, his motions are not restricted in any way.

Another interesting example is iFeel6BH1500 device wider described in [20]. In this case, the user's interface is being located in the middle of the frame and is being held by a number of cables. The cable's ends are coupled with drives fixed to the frame. The motions of the user pull some of the cables, while others are being released. The rotations of motors shafts are being calculated into interface orientation. A similar solution was introduced by Haption in INCA [21].

The fact is, that this solution solves the problem of linear limitations completely, as the operator movements are not limited in any way, yet both of the presented devices have enormous volumes, therefore the application has to be limited to some large rooms. Also, the usage of some of these devices can be hazardous to the operator, as the used motors generate a great torque that might easily break the user's arm, which is possible when adjusting the haptic device to its own applications using API. The other disadvantage is quite obvious when the size of the overall device is taken into account – fixing the device to the ground might even require some heavy equipment. It needs to be added that the problem of orientation limitations remains unsolved in these devices.

A known solution for overcoming orientation limitations in haptic devices is to add an additional degree of freedom to the kinematic chain. Such additional joint extends the maneuverability of the device, enabling the operator to move his hand without boundaries. Such a solution is being used by daVinci robot haptic device.

Such a solution solves the problem stated in previous examples of the large space required to place the device before using it. Yet, there is also a disadvantage connected to such a redundant approach – the analytical solution of such a kinematic chain becomes problematic [22] and computation complexity of the inverse kinematic problem increases [23] and the possibility of reaching the singularity of such kinematic chain remains.

Another possibility to avoid the limitations of any kind is to use an advanced vision system, for example as commercially available Kinect by Microsoft [24] in which the projector generates an Infrared (IR) net covering the user and the rest of the room and connected camera traces this net along with changes in its shape to detect motion. Another popular solution known from commercially available entertainment systems is the Move controller from Sony [25] where 6 DoF (Degrees of Freedom) motion of operator's hand is being detected by an IMU-based sensor system located inside a grip held by a human. Despite the fact, these solutions work well enough for entertainment purposes, they are not likely to be used as an alternative to haptic devices because it is not easy to implement force feedback features while using them.

The examples given above show that there is currently no ultimate solution to the stated problem, as mentioned at the beginning. Yet, such a solution had been found and will be presented in the following sections. It needs to be mentioned that the work will only focus on the orientation module of the haptic device, as this is the element in which singularities might appear. The device should be also as light as possible. The ideal solution would be to make the operator move his arm without feeling that he is using any kind of device at the same time. Yet, such systems do not exist currently. This article provides a solution to a haptic device that is characterized by a limitless workspace.

2 The Device

The development of the device free of drawbacks defined in the Introduction began with its construction design based on an analytical approach and the creation of assumptions for the control system and algorithms used. The description of these parts is divided into two parts presented in the following subsections.

1.1 Design Assumptions

To achieve an unlimited workspace of the manipulator, one needs to change the approach to the construction design and use an element that could move freely and without boundaries. The idea behind the proposed solution can be seen in Fig. 1.

The idea behind the proposed solution is that it consists of a sphere (1) that serves as a user interface (the operator is holding that part). Next to the sphere, there is a permanent magnet (2) that pulls the sphere towards itself. The sphere does not touch the magnet as it lies on three Omni wheels (3) that are mounted on drives (4) shafts. Omni wheels were added to the system to provide the possibility to transfer torque onto the sphere, which is required to enhance the device with a force feedback feature. The sphere is made of ARMCO 03J pure iron, which is a ferromagnetic alloy with low magnetic memory.

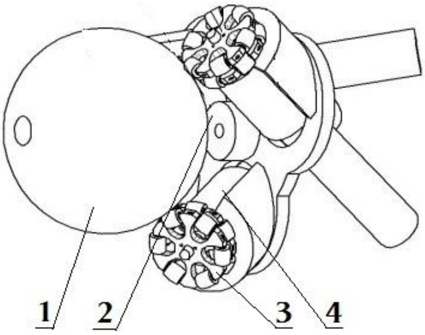


Figure 1
A concept of the device [27]

Also, the sphere is divided into two halves and is hollow. That is because of a fact that it also serves as a housing for the sensor system. The overall device needs to be oriented as shown in Fig. 1, as the working principle of the sensor system requires it [26].

The mentioned sensor system uses an IMU consisting of three triaxle sensors: an accelerometer, a magnetometer, and a gyroscope. The sensor is also equipped with a power source and a Bluetooth interface that provides communication with a control unit. The purpose of the sensor system is to determine the actual orientation of the sphere on a basis of acceleration and magnetic field vectors, while data obtained from the gyroscope serve as feedback for the filtration method. The magnetic field source provided by the magnet is much stronger than Earth's magnetic field and fields generated by the drives. As mentioned before, the magnet's second purpose is to pull the sphere next to itself resulting in its stable position. Although drives are used in the construction are equipped with encoders, they are not being used in the orientation calculation.

The orientation of the sphere is being derived by a combination of two vectors: Earth gravity and the magnetic field of the permanent magnet shown in Fig. 1. As a result of the constant orientation of the presented device, these vectors are always non-parallel, thus a rotation matrix can be defined with Eq. 1:

$$\mathbf{R} = [A_i \times M_i \quad A_i \times (A_i \times M_i) \quad A_i] \tag{1}$$

where:

A_i – acceleration unit vector components,

M_i – magnetic unit vector components.

A matrix presented in the given form fully describes the actual orientation of the sensor unit and a sphere. The derivation process had been thoroughly described in [27]. Data obtained from the sensors serve not only to determine the orientation but also as the input for the Kalman filter developed for this solution to obtain the highest accuracy measurement possible. It needs to be pointed that matrix (1) is used only once in the orientation determination process to obtain preliminary values – after that, it is the filter's duty to estimate the orientation of the system based on mentioned input. The only disadvantage of using a rotation matrix to define the orientation of the object is calculation complexity, as in each iteration all 9 elements need to be taken into account. Therefore, it has been decided to switch to some other orientation representation method. Quaternion algebra has been selected, for orientation described that way still fully represents object's state and preserves rotation matrix main advantage – lack of singularities occurrence, while the number of generated quaternion elements is four. Also, using quaternion algebra to describe the orientation is much more convenient in the case of developing mentioned Kalman filter, for all the elements of the unit quaternion are dependent on each other. Additionally, calculation of quternion elements direvatives is well known. The transformation from matrix to quaternion has been done according to [28].

As mentioned, the user interface is being pulled toward three Omni wheels. To ensure constant one-point contact between the sphere's surface and each of the wheels, their rolling elements shape had been geometrically determined [29]. It was necessary to select a proper material pair for elements being in contact to achieve maximum friction. The material selected for the rollers is Polyamide (PA6), which is characterized by a ca. 0.45 friction coefficient against steel [30]. The created Omni wheel construction can be seen in Fig. 2.

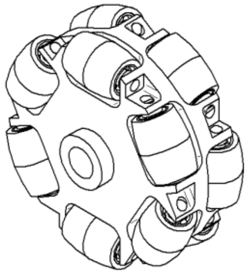


Figure 2
The isometric view of the designed Omni wheel [29]

It was necessary to determine the Jacobian matrix for the proposed device construction to achieve the possibility to control the drives on which Omni wheels are mounted. The result is the presence of a force feedback feature in the device. As the kinematic structure of the device is not typical, the standard Jacobian matrix derivation method has not been used. Instead, the derivation procedure had been based on the geometry of the device and vectors describing the relevant displacements of the drives [31], which are shown in Fig. 3.

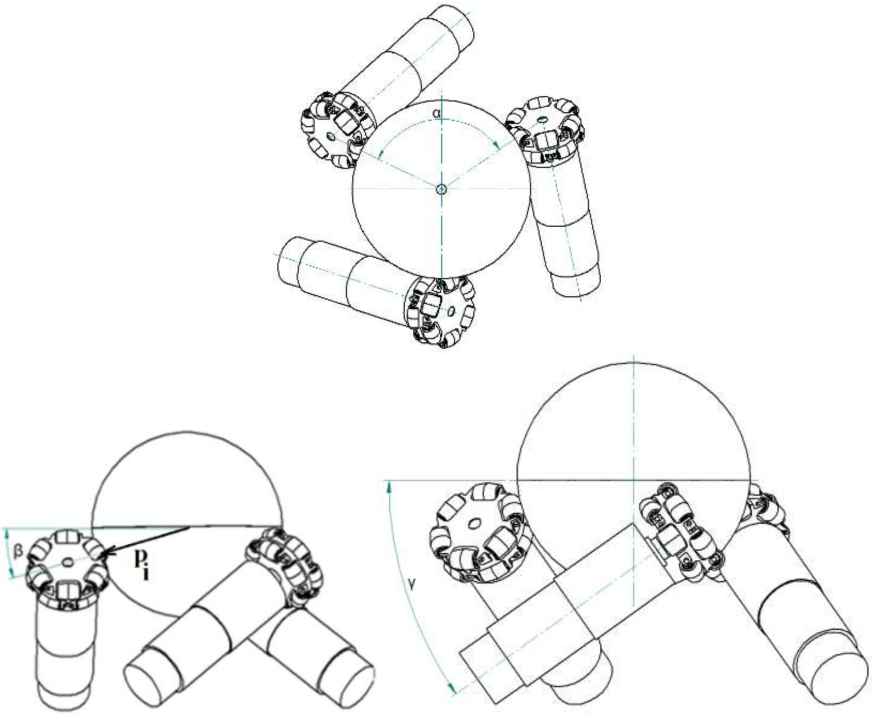


Figure 3
Geometrical dependencies of the device [31]

The outcome Jacobian matrix takes the following form:

$$J = \frac{(J^{-1})^{D}}{det\, J^{-1}} \tag{2}$$

Where

$$J^{-1} = \frac{R}{r} * \begin{bmatrix} \sin\alpha_i * \sin\gamma + \cos\alpha_i * \sin\beta * \cos\gamma \\ -\cos\alpha_i * \sin\gamma + \sin\alpha_i * \sin\beta * \cos\gamma \\ \cos\beta * \cos\gamma \end{bmatrix}^{T} \tag{3}$$

R – radius of the sphere, equal 35 mm,

r – radius of the Omni wheel, equal 15 mm,

α_i – angular position of the individual drives,

and the Jacobian matrix determinant presented in equation (2) is equal

$$detJ^{-1} = c\beta * c\gamma * (c\beta^2 * c\gamma^2 - 1) * (s(\alpha_1 - \alpha_2) - s(\alpha_1 - \alpha_3) + s(\alpha_2 - \alpha_3)) \tag{4}$$

It is worth noticing that all the elements of equation (4) are constant. Therefore, the resulting determinant value differs from zero, which is the ultimate proof that the presented device is free of singularities. The optimum orientation of the individual drives has also been determined using a derived Jacobian matrix.

1.2 Control System

The overall control system loop can be seen in Fig. 4.

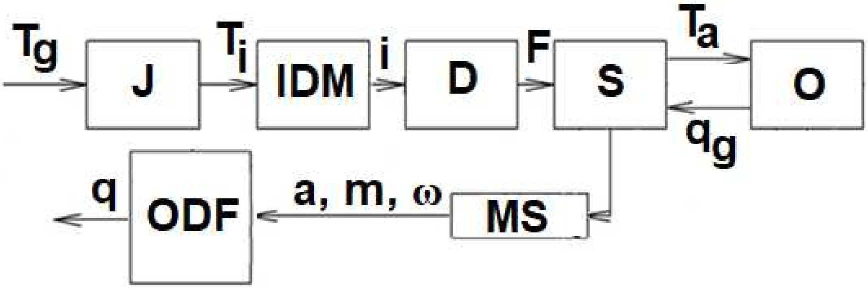


Figure 4
Haptic device control system schematic

The input for the control system is a Torque value (T_g – given torque) that should be generated according to the actual state of the manipulator controlled by the given haptic device. By using the derived Jacobian matrix (J) the input is being calculated to input torques (T_i) for the individual drives. Basing on the Inverse drive model (IDM) the control currents (i) are being calculated and passed to the drives (D). Their motion is being transferred by the friction force (F) to the sphere (S) which is being held by the operator (O), therefore, the outcome torque (T_a – actual torque) generated in the sphere serves as an output of the system. Yet, there is another input – the operator who is in control of the device makes rotation with his hand providing the system with orientation change of the sphere (described using quaternions, q_g – given quaternion). A measurement system (MS) located inside the sphere detects changes in the orientation by measuring acceleration, magnetic flux, and angular velocity vectors (a, m, ω). Based on their values, an algorithm determines the actual orientation of the object and performs the filtration process (ODF – Orientation determination and filtration), providing corrected data on a measured angular displacement of the sphere (q).

The mentioned filtration method is being based on Extended Kalman Filter (EKF) [32]. To make use of the filter possible in this case, it is necessary to correctly select elements of the state vector that describe the actual state of the measured object. It was decided that this vector should be:

$$x = [q_i \quad \omega_i \quad b_i]^T \tag{5}$$

where

q_i - quaternion elements in which q_o is scalar element,

ω_i – angular velocity elements,

b_i – gyroscope's bias elements.

Mentioned bias elements are taken into account in the case of this filter, a sensor on which the measurement unit is being based is a low-cost model and it was assumed that measured values will change over time. A measurements vector of EKF is being constructed with values that are being measured in the process, which is:

$$y = [a_{mi} \quad m_{mi} \quad \omega_{mi}]^T \tag{6}$$

where:

a_{mi} – measured acceleration vector elements,

m_{mi} – measured magnetic vector elements,

ω_{mi} – measured angular velocity vector elements.

As EKF is a recurrent filter type, it requires an object's state in the next time step to be defined. For a given state vector, such prediction takes the following form:

$$x_{t+1} = \begin{bmatrix} q_i + \Delta q_i \\ \omega_i + \Delta\omega_i \\ b_i \end{bmatrix} \tag{7}$$

where:

$$\Delta q = \dot{q} \cdot \Delta t \tag{8}$$

Quaternion derivative from equation (8) can be calculated according to [33] and

$$\Delta\omega_i = \varepsilon_i \tag{10}$$

where ε_i is an angular acceleration value in a given direction. Its value needs to be determined during the filter creation process. In the given case, the standard acceleration value corresponds to the operator's hand motions and for the filter's creation purpose it was defined as equal 2 m/s^2, for exceeding this value seems unlikely in the typical haptic device application.

The last three elements of the state vector (7) describe the bias of the gyroscope. Besides noise data acquired with the used gyroscope is also burdened with bias – the initial values were taken by reading the angular velocity of the fixed device. The reading should be equal, yet they are not and it was assumed that this difference is to change its volume over time. Bias elements were added into the state vector and included in equations describing angular velocity to estimate bias variations and to remove its impact on the readings.

Further creation of the EKF for the given example does not deviate from the standard EKF procedure and the details and overall process description can be found in [34].

2 Experiments

The experimental phase of the created device was divided into two parts. The first part is connected to define mechanical parameters of the device, such as the maximum transferable force and repeatability of force generation by the force feedback feature. Another part was the tests done on created measurement unit to determine its features, like filtration correctness, repeatability of measurement, resolution of the sensor system.

2.1 Torque Transfer Level

One of the assumptions of the described project was to use materials pair providing maximum friction to obtain drive transmission in 3 Degrees of Freedom. As mentioned in Section 1.1, the assumption was that sphere is made of some kind of metal, while the Omni wheels rollers material is polyamide. It became necessary to check what is the maximum transferable torque for the final assembly of created elements and selected materials. To perform this task, an experiment had been conducted. During the process, a 6D force sensor was used. The device had been equipped with a rode which was being pushed toward the force sensor's surface and pressed with an additional external load to make it possible to further press onto the force sensor and release the pressure in the same experiment. The test started with zero torque applied by the drives of the device – it was the zero point of the measurement system. Next, a maximum negative torque of -0.11 Nm was applied and afterward, it was being slightly increased step-by-step until the positive maximum of 0.11 Nm was reached. After that, the process was reversed and repeated until the maximum negative value was reached again. The step value was 0.01 Nm. The experiment was repeated 5 times for each direction of the internal device coordinate system. The result of one data set can be seen in Fig. 5.

The maximum set values named earlier were specified during a preliminary experiment which was to detect a force for which the frictional connection between the sphere and Omni wheels becomes unstable and slippage starts to occur. It was done using the same test stand, yet the procedure was different.

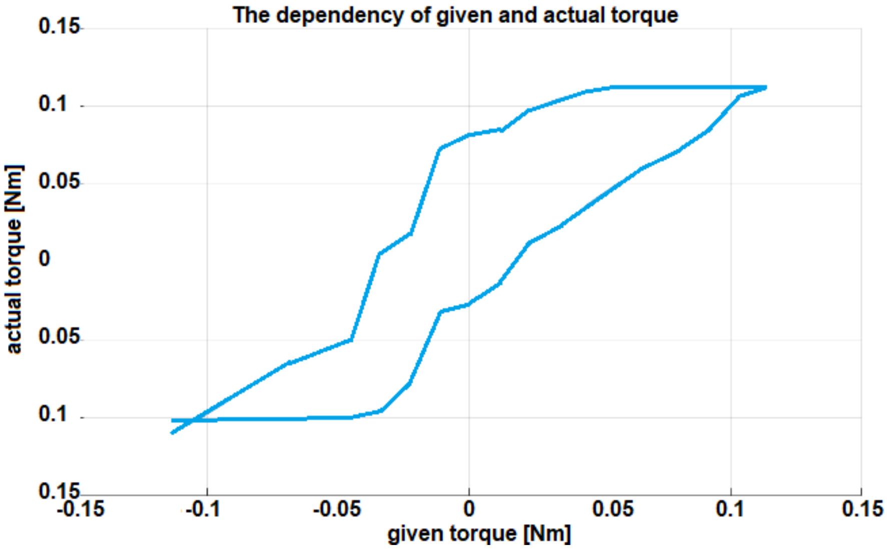


Figure 5
A dependency of set and obtained torque of the user's interface

The torques of the drives were being increased by a small step starting from 0 value, while the outcome torque of the sphere was measured. The torques of the drives were being increased until a slippage occurred. The experiment has shown that the maximum "safe" torque value is 0.11 Nm.

The actual experiment has shown that there are little losses of the torque during the process and measured values correspond to set values in boundary positions. Yet, during the motion of the device, this dependency is not so strict – as expected, hysteresis can be seen. The main purpose of this phenomenon is the presence of the backlash in cooperating elements, especially in Omni wheels rollers mounting system. The knowledge of this fact is important as it is possible to create the correction algorithm that would take into account the sureness level of torque value transfer at a set level to control the given torque value.

2.2 Filtration Corectness

During this experiment, the developed EKF has undergone series of tests to check the correctness of the filtration process. During all of the tests, the sensor unit mentioned in Section 1.1 was being placed inside the user's interface and set to record the data. After the acquisition process, the resulting file was processed. The motion performed by the sphere was done by the operator's hand in given sequences but also in random direction with variating velocity dependent on operator selection. The sampling rate was 100 Hz.

In Fig. 6 a scalar and two imaginary elements of the quaternion are presented. Raw data from the sensor unit was taken to calculate the quaternion value to compare it with the corresponding value estimated by the Kalman filter. This calculation based on raw data is made only for these tests and is not being used anywhere in the algorithm.

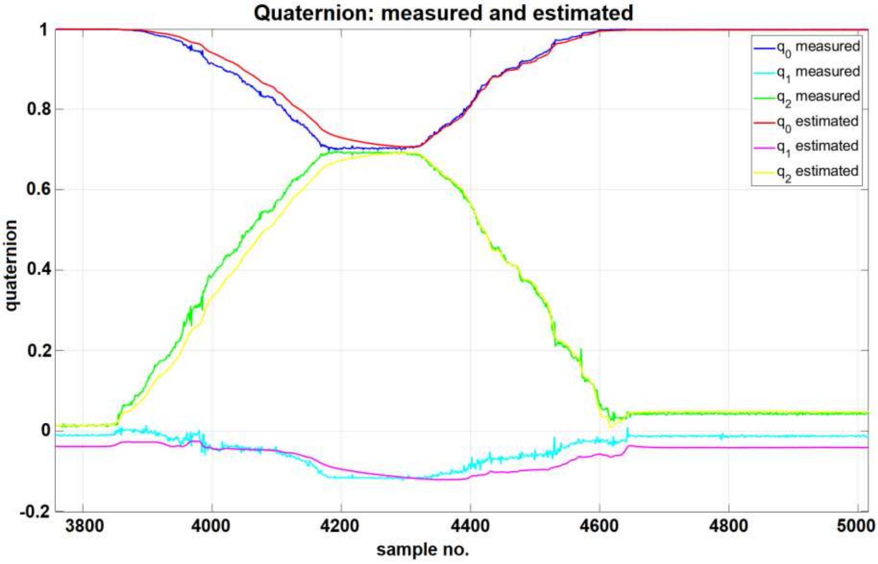


Figure 6
Quaternion filtration

It can be seen that in the stationary phases (e.g. samples 0 – 1500) actual and estimated values are slightly moved from each other. This is an outcome of the fact that estimation based on the data collected from three different sensors and their fusion and is acceptable.

The estimated quaternion values presented in Fig. 6 are based on filtered data obtained from sensors specified earlier. The estimation results of this collection can be seen in Fig. 7, an accelerometer in this case. It needs to be mentioned that during the filtration procedure, only raw data from the sensor are taken into account. Filtered readings are only used to check filtration process correctness.

As it can be seen, the phenomena present in the previous figure, are present in accelerometer estimation as well. The estimated value follows the actual one perfectly, yet corresponding values are slightly moved next to each other. The origin of the displacement has already been analyzed.

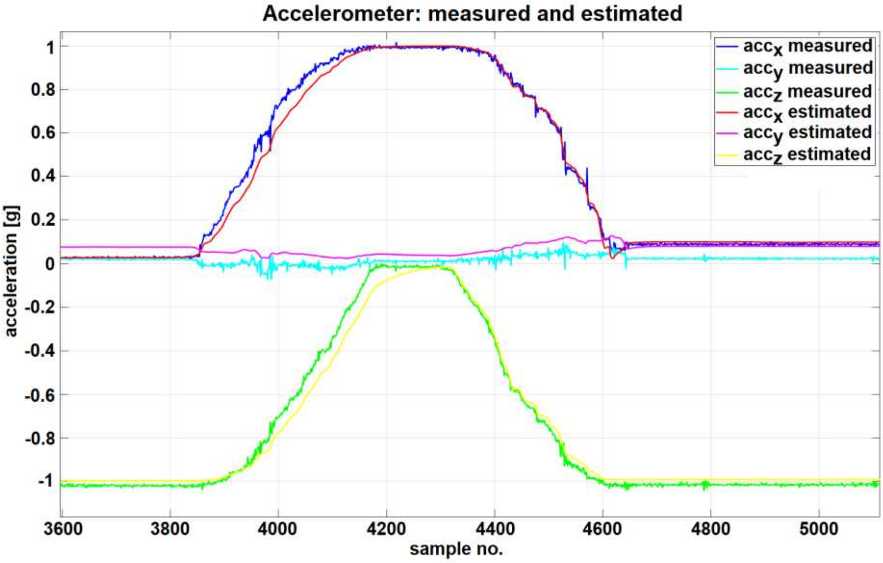


Figure 7
Accelerometer filtration

2.3 Repeatability

In the repeatability of the measurement test, the haptic device had been equipped with a measurement rode which could be pressed against a prismatic base. It resulted in the stable and repeatable orientation of the interface when pressed. During the experiment the rode was being moved away from the prism and pressed again towards it – the process has been repeated 30 times. Each time the rode was being pressed, its orientation was being stabilized and measured using a built-in sensor system. The angle values of the position to be obtained were determined by moving the rode to the prism for the first time, and after waiting a couple of seconds to make self-vibrations of the system damp, the reading has been made.

The point cloud obtained during the experiment can be seen in Fig. 8.

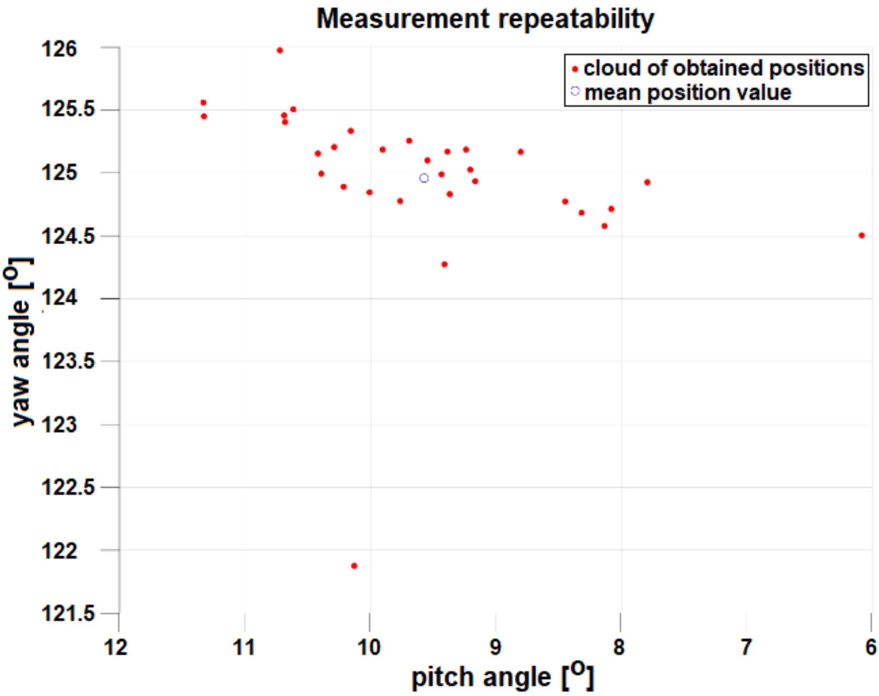


Figure 8

Measured positions in subsequent approaches to the set orientation presented in pitch-yaw plane

The blue point visible in the Figure represents the mean value of all obtained points in the cloud. The obtained point cloud can be used to determine the value of the repeatability parameter with the use of the equation provided in [35]:

$$P_a = \frac{\sum_{j=1}^{n}\sqrt{(\overline{Rl}-Rl_j)^2+(\overline{Pch}-Pch_j)^2+(\bar{Y}-Y_j)^2}}{n} \tag{11}$$

where

$\overline{Rl}$, $\overline{Pch}$, $\bar{Y}$ – mean value of the obtained angle,

Rl_j, Pch_j, Y_j – actual values of obtained angles.

Obtained repeatability for the described measurement system was calculated as $P_a = 1.18^o$.

The creators of commercially available master manipulators seldom share this parameter with the audience, yet it can be found that in the case of Cybergrasp haptic glove it is ca. 1° [36], which shows that the obtained result proves the possible applicability of the created device in this matter.

2.4 Resolution

To specify the minimum detectable displacement that can be sensed by the sensor system, a test on the resolution has been performed. To conduct this experiment, a test stand has been constructed. The idea behind this test was to make the sphere containing a sensor system (1) rotate by a given value. To ensure the maximum precision and repeatability of subsequent rotations, the stand has been equipped with a Maxon Motor servo drive (2). It can be seen in Fig. 9.

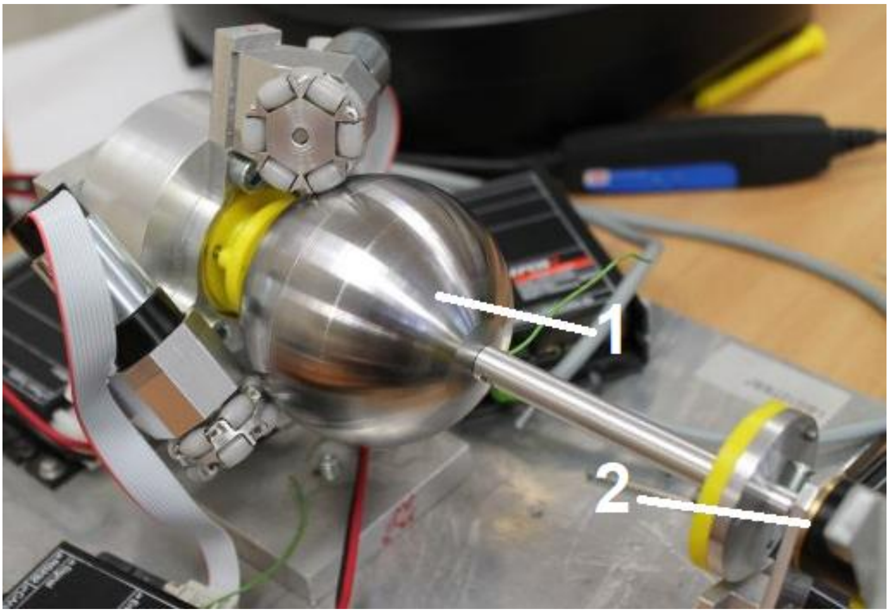


Figure 9
Test stand for resolution measurement

During the test, a set of rotations was done with the drive's shaft resulting in the user's interface and sensor system rotation at the same time. The rotation angle value was being changed between the sets and results read by the sensor were recorded. The purpose was to find displacement value for which it wouldn't be possible to distinguish individual steps. The displacement value range starts with 1° and ends with 0.006°, which was the minimum value possible to obtain by used servo drive.

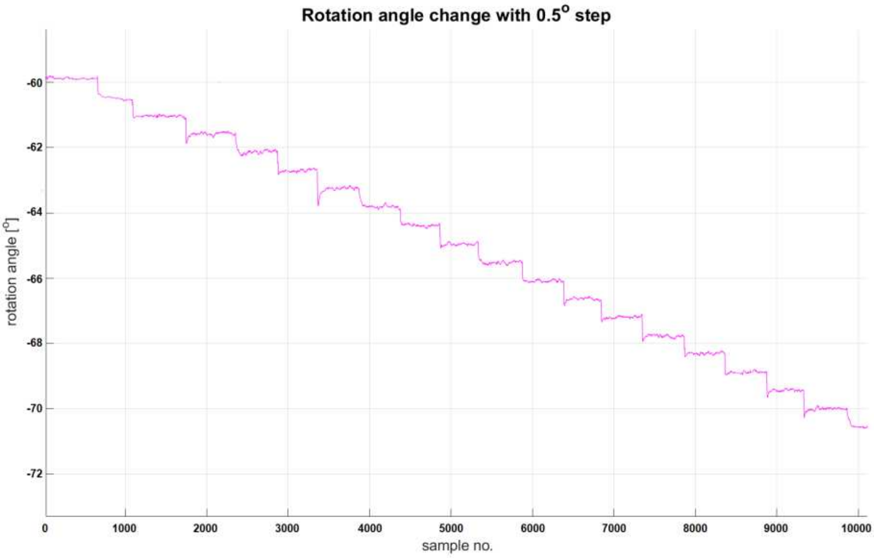


Figure 9
Diagrams showing a set of rotations with 0.5dg step

For the motions with the displacement, steps equal 1° and 0.5° (Fig. 9) the obtained value was big enough to allow a clear reading of a rotation angle. It also showed that measured rotation angle corresponds to a given one, which additionally proves that the developed fusion data algorithm and orientation determination procedure is effective and works correctly.

The rotation angle step has been further decreased to 0.1° value. The sensor system was able to detect such displacements, yet the increment value is lower than the present noise of the reading, therefore the obtained angular value can be only shown with some approximation.

3 Discussion

As the performed experiment results have shown, the proposed device can be used to control manipulators (their orientation) while providing force feedback and quite precise orientation determination while giving no artificial restrictions to the operator's hand motion. Yet, there is still room for improvements, for example, a correction algorithm should be added to the control system which was schematically presented in Fig. 4. It is planned to add correction value to the T_i element to avoid torque losses resulting from backlash shown in Fig. 5. To fully determine the sureness level of torque transfer it is necessary to perform additional experiments in a dynamic setup in which the device would be in motion during force application.

Additional works are planned to be done on the filtration algorithm – it is planned to add additional filtration used during really slow and precise motions to increase resolution value over the values shown in Fig. 9. Also, the dynamics of the created filtration method will be checked to determine how the system reacts to fast and rapidly changing motions of the sphere. Also, the completed control system (with mentioned correction algorithm loop) will be fully described, including its transmittance calculation.

The planned works will be summarized by the creation of a new measurement unit that will be equipped with all the features shown in the paper and listed above, as for now, its only purpose is to collect orientation data from the sensors – as was stated earlier.

Conclusions

This article presented a thorough description of a novel master manipulator device, in which the main feature is a limitless orientation motion and the fact that it is being free of singularities occurrence. The survey of existing commercially available, as well as experimental haptic devices has been done to point advantages and disadvantages of such mechanisms. It showed that all of them try to avoid singularities occurrence by making the user unable to reach such a position, without trying to solve the problem completely. This work presented the creation process of the device that is free of named issues. The idea behind the construction has been explained, the design process has been described, there is also the description of the Jacobian matrix necessary to implement the force feedback feature to assist the operator during his work. Also, a control system of the device was presented and thoroughly described, taking into account the explanation of constructed measurement unit, its work principle, and an algorithm used to calculate the orientation of the connected object. Also, the description of the used filtration method has been presented, along with the explanation of basic assumptions needed to be done during the filter creation process.

The constructed device had undergone the experimentation process to check its applicability, main features, characteristics. The maximum force possible to be transferred by the friction in a given case has been defined to check the volume of force feedback felt by the operator. The accuracy and repeatability of the measurement unit have been determined during two experiments. Also, the correctness of created data filter was tested and verified. All the experiments, test stations, and results were presented and discussed. The overall experimentation process showed that created device is working according to the assumptions and can be used in professional applications.

References

[1] V. Ciobanu, N. Popescu, A. Petrescu, Robot telemanipulation system, IEEE conference on System Theory Control and Computing (ICSTCC), Vol. 17, pp. 681-686, October 2013

[2] A. Barrow, W. Harwin, Design and Analysis of a Haptic Device Design for Large and Fast Movements, Machines, 4(1), 2016

[3] M. Boroujeni and M. Misagh. Daly: Haptic Device Application in Persian Calligraphy, Proceedings of the 2009 International Conference on Computer and Automation Engineering, 2009

[4] J. Kutuvan: Remote Handling Technology in Nuclear Industry, Modern Manufactuing India, Vol. 1, Issue 3, 2013

[5] S. Sivčev, J. Coleman, E. Omerdić, G. Dooly, D. Toal: Underwater manipulators: A review, Ocean Engineering, Vol. 163, pp. 431-450, 2018

[6] E. Amareswar, G. S. S. K. Goud, K. R. Maheshwari, E. Akhil, S. Aashraya and T. Naveen: Multi purpose military service robot, 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 684-686, doi: 10.1109/ICECA.2017.8212752

[7] J. Leven et al.: DaVinci Canvas: A Telerobotic Surgical System with Integrated, Robot-Assisted, Laparoscopic Ultrasound Capability, Lecture Notes in Computer Science (Vol. 3749), pp. 811-818, 2005

[8] L. Podsędkowski, P. Żak: Tests on cardiosurgical robot robin heart 3, Robot Motion and Control (Vol. 396), pp. 433-442, 2009

[9] M. Moradi Dalvand, S. Nahavandi, M. Fielding, J. Mullins, Z. Najdovski and R. D. Howe: Modular Instrument for a Haptically-Enabled Robotic Surgical System (HeroSurg), IEEE Access, Vol. 6, pp. 31974-31982, 2018, doi: 10.1109/ACCESS.2018.2844563

[10] Kent, H Nelson: Hands across the ocean for world's first trans-Atlantic surgery, CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 165 10 (2001): 1374

[11] M. Fontana, E. Ruffaldi, F. Salasedo, M. Bergamasco: On the Integration of Tactile and Force Feedback, Haptics Rendering and Applications, 2012

[12] Batteau L. M., Liu A., Maintz J. B. A., Bhasin Y., Bowyer M. W: A Study on the Perception of Haptics in Surgical Simulation, Medical Simulation, Vol. 3078, 2004

[13] L. Podsedkowski, J. Moll, M. Moll, L. Fracczak: Are the surgeon's movements repeatable? An analysis of the feasibility and expediency of implementing support procedures guiding the surgical tools and increasing motion accuracy during the performance of stereotypical movements by the

surgeon, Kardiochirurgia i Torakochirurgia Polska, Volume 11, Issue 1, 2014, pp. 90-101

[14] Fracczak L, Szaniewski M, Podsedkowski L.: Share control of surgery robot master manipulator guiding tool along the standard path. Int J Med Robot Jun;15(3):e1984, 2019, doi: 10.1002/rcs.1984

[15] Mugge W, Kuling IA, Brenner E, Smeets JBJ.: Haptic guidance needs to be intuitive not just informative to improve human motor accuracy, PLoS ONE, 2016, doi: 10.1371/journal.pone.0150912

[16] P. Garrec, J. P. Friconneau, Y. Measson, and Y. Perrot: Able, an innovative transparent exoskeleton for the upper-limb, Intelligent Robots and Systems, 2008

[17] P. N. Kooren, J. Lobo-Prat, M. M. H.P. Janssen, A. Q. L. Keemink, A. H. A. Stienen, I. J. M. de Groot, M. I. Paalman, R. Verdaasdonk, B. F. J. M. Koopman: Design and control of the active a-gear: a wearable 5 dof arm exoskeleton for people with duchenne muscular dystrophy, IEEE International Conference on Biomedical Robotics and Biomechatronics, 2016

[18] G. Tholey, J. P. Desai: A General-Purpose 7 DOF Haptic Device: Applications Toward Robot-Assisted Surgery, Transactions on Mechatronics (Vol. 12, Issue 6), pp. 662-669, 2007

[19] Q. Parent, J. Perret: Usability of a large-scale force-feedback device in different immersive environments, EuroVR Conference 2016, Athens, 2016

[20] Z. Chen, Y. Zhang, D. Wang, C. Li and Y. Zhang: iFeel6-BH1500: A large-scale 6-DOF haptic device, 2012 IEEE International Conference on Virtual Environments Human-Computer Interfaces and Measurement Systems (VECIMS) Proceedings, Tianjin, 2012, pp. 121-125, doi: 10.1109/VECIMS.2012.6273226

[21] J. Perret, L. Dominjon: The INCA 6D: a Commercial Stringed Haptic System Suitable for Industrial Applications, Joint Virtual Reality Conference, 2009

[22] C. Yu, M. Jin, H. Liu: An Analythical Solution for Inverse Kinematic of 7-DOF Redundant Manipulators with Offset-Wrist, IEEE International Conference on Mechatronic and Automation Proceedings, pp. 92-97, 2012

[23] J. Wang, Y. Li, X. Zhao: Inverse Kinematics and Control of a 7-DOF Redundant Manipulator Based on the Closed-Loop Algorithm, International Journal of Advanced Robotic Systems (Vol. 7, Issue 4), pp. 1-12, 2010

[24] M. S. H. Abdullah, A. Zabidi, I. M. Yassin and H. A. Hassan: Analysis of microsoft kinect depth perception for distance detection of vehicles, IEEE 6th Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, 2015, pp. 116-119, doi: 10.1109/ICSGRC.2015.7412476

[25] M. Paleari, R. Luciani and P. Ariano: Towards NIRS-based hand movement recognition, 2017 International Conference on Rehabilitation Robotics (ICORR), London, 2017, pp. 1506-1511, doi: 10.1109/ICORR.2017.800946

[26] L. Podsędkowski P. Żak: Moduł zadawania położenia kątowego zadajników położenia, P.397176, 29.01.2016

[27] P. Żak: Using 9-axis Sensor for Precise Cardiosurgical Robot Master Angular Position Determination, Solid State Phenomena Vol. 199, pp 356-361, 2013

[28] YB. Jia: Quaternions and Rotations, Com S 477/577 Notes, Sep 10, 2013

[29] P. Żak: Ballbots rolling elements shape determination, 2016 21st International Conference on Methods and Models in Automation and Robotics (MMAR), Miedzyzdroje, 2016, pp. 1143-1147, doi: 10.1109/MMAR.2016.7575299

[30] Voyer J, Klien S, Velkavrh I, Ausserer F, Diem A. Static and Dynamic Friction of Pure and Friction-Modified PA6 Polymers in Contact with Steel Surfaces: Influence of Surface Roughness and Environmental Conditions. Lubricants. 2019; 7(2):17, https://doi.org/10.3390/lubricants7020017

[31] P. Żak: Jacobian matrix determination in a novel master manipulator device, 2016 17th International Conference on Mechatronics - Mechatronika (ME), Prague, 2016, pp. 1-5

[32] L. Kleeman: Understanding and applying Kalman filtering, Proceedings of the Second Workshop on Perceptive Systems Curtin University of Technology, 1996

[33] D. Xu, D. P. Mandic: The theory of quaternion matrix derivatives, IEEE Trans. Signal Process., Vol. 63, No. 6, pp. 1543-1556, 2015

[34] P. Żak: Master Manipulator Orientation Determination Method Using Extended Kalman Filter, 2018 18th International Conference on Mechatronics - Mechatronika (ME), Brno, 2018, pp. 276-280

[35] EN ISO 9283:1998 standard

[36] M. Aiple and A. Schiele, "Pushing the limits of the CyberGrasp™ for haptic rendering," 2013 IEEE International Conference on Robotics and Automation, 2013, pp. 3541-3546, doi: 10.1109/ICRA.2013.6631073

Outage Performance of Macrodiversity Reception in the Presence Rayleigh Short-Term Fading and Co-channel Interference

Branimir Jakšić, Jelena Todorović, Đoko Banđur, Branko Gvozdić, Miloš Banđur

Faculty of Technical Sciences, University of Pristina in Kosovska Mitrovica, Knjaza Milosa 7, 38220 Kosovska Mitrovica, Serbia
branimir.jaksic@pr.ac.rs, jelena.todorovic@pr.ac.rs, djoko.bandjur@pr.ac.rs, branko.gvozdic@pr.ac.rs, milos.bandjur@pr.ac.rs,

This paper presents a wireless macrodiversity communication system consisting of a receiver for macrodiversity selection combining (SC) receiver and two microdiversity SC receivers operating over correlated Gamma shadowed Rayleigh multipath fading environment in the presence of co-channel interference subject to Rayleigh short-term fading. First, we derive expression for cumulative distribution function of output signals of the both microdiversity SC receivers, and then capitalizing on it, we evaluate outage probability of the macrodiversity reception. The obtained probability outage results, both numerical and those obtained by simulation, as well as, the influence of Gamma long-term severity parameter and the correlation coefficient, are graphically illustrated and discussed.

Keywords: selection combining (SC); Rayleigh fading; co-channel interference; Gamma shadowing; outage probability

1 Introduction

Short-term fading, long-term fading and co-channel interference degrade mobile wireless communication system performance, cause signal envelope variation and signal envelope average power variation [1]. Macrodiversity reception simultaneously reduces both long-term and short-term, fading effects on the system performance [2-5]. Microdiversity receivers mitigate short-term fading effects while macrodiversity receiver reduces long-term fading effects on wireless communications system outage performance [6-7]. Macrodiversity system is composed of a macrodiversity SC (Selection Combining) receiver and two or more microdiversity SC receivers [8-9]. Each microdiversity SC receiver selects branch with the highest signal-to-interference ratio while macrodiversity SC receiver selects microdiversity receiver with the highest output signal power [10-11].

Various system models are presented in the publicly available literature. Some of them, proven to be successful in modeling different system operating conditions, are presented in [12-17]. These models can also be applied in the modeling of wireless communication systems that use diversity technology.

The performance analysis of a SC receiver in the presence of multipath fading and co-channel interference is studied in [18-23]. A macrodiversity system composed of one macrodiversity SC receiver and two microdiversity MRC (Maximum Ratio Combining) receivers operating over Gamma shadowed Rician multipath fading channel is considered in [18]. Closed form bit error probability expression of the microdiversity SC receivers is derived, and then that expression is used for evaluation of the entire macrodiversity system bit error probability. Macrodiversity system with a macrodiversity SC receiver with L branches and MRC microdiversity receivers with N branches operating over Gamma shadowed Nakagami-m multipath fading channel is analyzed in [19], where the second order performance metrics such as level crossing rate and average fade duration are calculated as closed form expressions.

Performance analyzes of wireless communication systems with diversity technology and the impact of different types of fading are discussed in [6-7, 24-27]. In [6], LCR (Level Crossing Rate) of the signal at the SC macrodiversity system output in the presence of α-μ short-term fading and gamma long-term fading was determined. The first order statistical characteristics of an output signal of the system consisting of three MRC microdiversity receivers and one SC macrodiversity receiver were studied in [7]. Input signals of the microdiversity MRC receivers are subject of independent k-μ short-term fading and correlated Gamma long-term fading. In [24], statistical moments of a signal at the output of system consisting of two EGC (Equal Gain Combining) microdiversity receivers and one SC macrodiversity receiver were determined in the presence of Nakagami-m fading. In [25], outage probability of a macrodiversity system consisting of two SC microdiversity receivers and one SC macrodiversity receiver in the presence of Fading and Weibull co-channel Interference was calculated. A study of wireless propagation in non-homogenous environment under non-deterministic LOS (Line-of-Sight) conditions, when the random nature of dominant/scattering components ratio has been considered and modeled as Gamma distribution is given in [26]. It includes closed-form expressions for PDF (Probability Density Function), CDF (Cumulative Distribution Function), outage probability and ABER (Average Bit Error Probability) of the observed fading process. In [27], expression for the chip error probability, the symbol error probability and the packet error probability are evaluated for IEEE 802.15.4 wireless channel in the presence of κ–μ fading, interference and additive white Gaussian noise.

A performance analysis of SC macrodiversity reception in the presence of Rayleigh short-term fading and co-channel interference is a novelty in the publicly available literature, according to authors best knowledge. Therefore, the contribution of this paper, is as follows:

- Modeling of a wireless communication system consisting of a macrodiversity SC receiver and two microdiversity SC receivers operating over correlated Gamma shadowed Rayleigh multipath fading environment in the presence of co-channel interference subject to Rayleigh short-term fading.
- Derivation of the analytical expression for outage probability at the output of the macrodiversity system.
- Analysis of the obtained numerical results, determining the impact of system parameters on the received signal quality.

2 System Model

Macrodiversity system studied in this paper is composed of a macrodiversity SC receiver and two microdiversity SC receivers operating over correlated Gamma shadowed Rayleigh multipath fading channel in the presence of the co-channel interference subject to Rayleigh short-term fading. The system model is shown in Fig. 1.

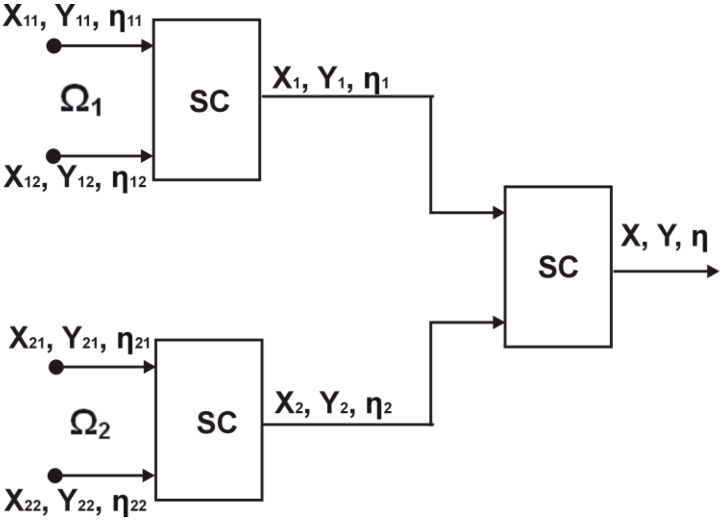


Figure 1

System model of the studied macrodiversity reception

Received signal envelopes at the branches of the first microdiversity receiver are denoted with x_{11} and x_{12}, and at the branches of the second microdiversity receiver with x_{21} and x_{22}. The co-channel interference envelopes at the branches of the first microdiversity SC receiver are denoted with y_{11} and y_{12}, and at the branches of the second microdiversity receiver with y_{21} and y_{22}. The respective signal to interference ratios are denoted with η_{11}, η_{12}, η_{21} and η_{22}. Desired signal envelopes at the output of the microdiversity SC receivers are denoted with x_1 and x_2, co-channel interference envelopes with y_1 and y_2, and signal to interference ratios with η_1 and η_2. Desired signal at the output of the macrodiversity SC receiver is denoted with x, interference with y, while η is signal to interference ratio. Average power of the desired signal at the branches of the microdiversity SC receivers is denoted with Ω_1 and Ω_2, and the average power of the interference signal with S_1 and S_2. PDF of x_{ij}, i=1,2; j=1,2 is [2, 28]:

$$p_{x_{ij}} = \left(x_{ij}\right) = \frac{2x_{ij}}{\Omega_i} e^{-\frac{x_{ij}^2}{\Omega_i}}, \quad x_{ij} \geq 0, \quad i=1,2; \quad j=1,2 \tag{1}$$

and probability density function of y_{ij} is [2, 28]:

$$p_{y_{ij}} = \left(y_{ij}\right) = \frac{2y_{ij}}{S_i} e^{-\frac{y_{ij}^2}{S_i}}, \quad y_{ij} \geq 0, \quad i=1,2; \quad j=1,2 \tag{2}$$

The signal to interference ratio at the branches of the microdiversity SC receivers is

$$\eta_{ij} = \frac{x_{ij}}{y_{ij}}, \quad i=1,2; \quad j=1,2 \tag{3}$$

Probability density function of η_{ij} is

$$p_{\eta_{ij}}\left(\eta_{ij}\right) = \int_0^\infty dy_{ij}\, y_{ij}\, p_{x_{ij}}\left(\eta_{ij} y_{ij}\right) p_{y_{ij}}\left(y_{ij}\right) = 2\Omega_i S_i \frac{\eta_{ij}}{\left(\eta_{ij}^2 S_i + \Omega_i\right)^2}, \quad i=1,2; \quad j=1,2 \tag{4}$$

CDF of η_{ij} is

$$F_{\eta_{ij}}\left(\eta_{ij}\right) = \int_0^{\eta_{ij}} dt\, p_{\eta_{ij}}\left(t\right) = 2\Omega_i S_i \int_0^{\eta_{ij}} dt \frac{t}{\left(t^2 S_i + \Omega_i\right)^2} = \frac{\eta_{ij}^2}{\eta_{ij}^2 + \frac{\Omega_i}{S_i}} =$$

$$= 1 - \frac{\frac{\Omega_i}{S_i}}{\eta_{ij}^2 + \frac{\Omega_i}{S_i}}, \quad i=1,2; \quad j=1,2 \tag{5}$$

Cumulative distribution function of at outputs of microdiversity SC receivers is

$$F_{\eta_i}\left(\eta_i\right)=F_{\eta_{i1}}\left(\eta_i\right)F_{\eta_{i2}}\left(\eta_i\right)=\frac{\eta_i^4}{\left(\eta_i^2+\dfrac{\Omega_i}{S_i}\right)^2},\quad i=1,2 \tag{6}$$

Joint probability density function of **Ω**$_1$ and **Ω**$_2$ is [29]:

$$p_{\Omega_1\Omega_2}\left(\Omega_1\Omega_2\right)=\frac{1}{\Gamma(c)\left(1-\rho^2\right)\rho^{\frac{c-1}{2}}\Omega_0^{c+1}}\sum_{i_1=0}^{\infty}\left(\frac{\rho}{\Omega_0\left(1-\rho^2\right)}\right)^{2i_1+c-1}\times$$

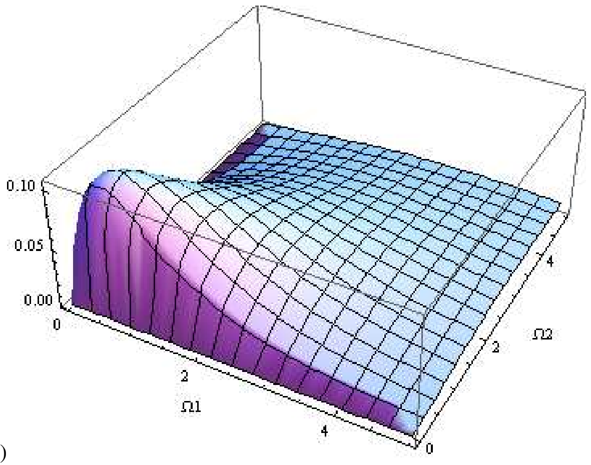
$$\times\frac{1}{i_1!\Gamma\left(i_1+c\right)}\Omega_1^{i_1+c-1}\Omega_2^{i_1+c-1}e^{-\frac{\Omega_1+\Omega_2}{\Omega_0\left(1-\rho^2\right)}} \tag{7}$$

where c is Gamma fading severity parameter, **Ω**$_0$ is the average power of **Ω**$_1$ and **Ω**$_2$ and ρ is correlation coefficient. Joint probability density function of **Ω**$_1$ and **Ω**$_2$ based on (8) is plotted in Fig. 2.

The joint probability density function of S_1 and S_2 follows independent Gamma distribution:

$$p_{S_1S_2}\left(S_1S_2\right)=\frac{1}{\Gamma(c)S_0^{c_1-1}}S_1^{c_1-1}e^{-\frac{S_1}{S_0}}\cdot\frac{1}{\Gamma(c)S_0^{c_1-1}}S_2^{c_1-1}e^{-\frac{S_2}{S_0}} \tag{8}$$

where c is Gamma severity parameter and S_0 is the average square value of S_1 and S_2.



a)

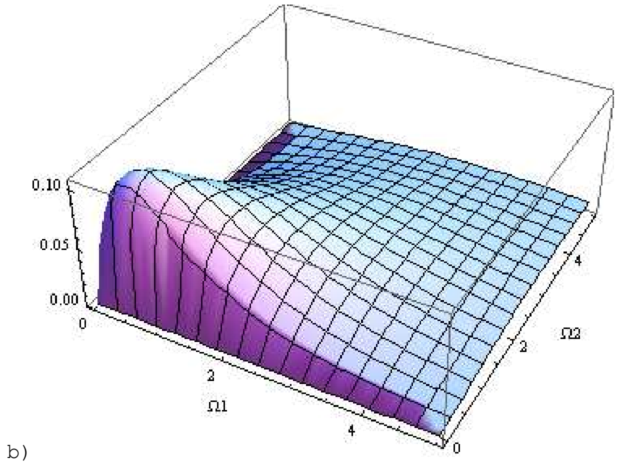


Figure 2
Joint probability density function of **Ω**$_1$ and **Ω**$_2$ for c=1.5 and: a) ρ=0.2, b) ρ=0.8

3 Analytical Results

The algorithm for calculating the analytical expression for outage probability at the output of the macrodiversity system is presented in Fig. 3.

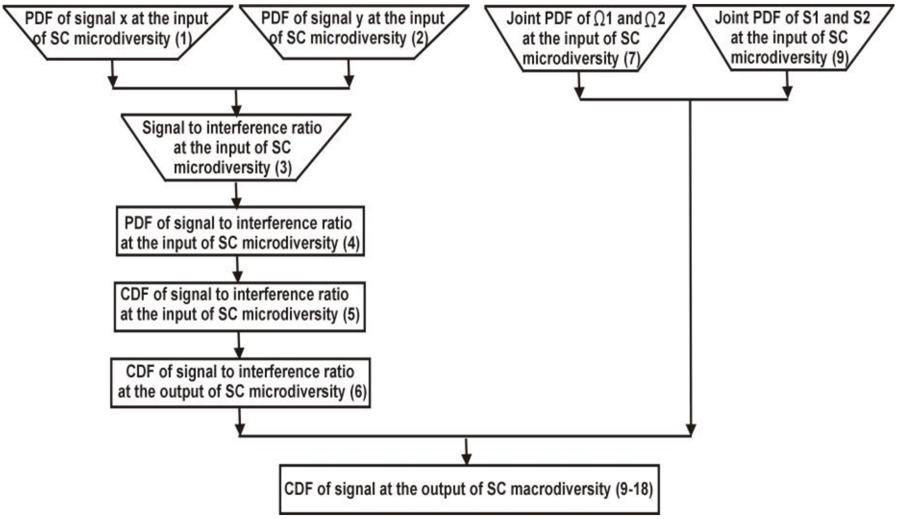

PDF of signal x at the input of SC microdiversity (1)
PDF of signal y at the input of SC microdiversity (2)
Joint PDF of Ω1 and Ω2 at the input of SC microdiversity (7)
Joint PDF of S1 and S2 at the input of SC microdiversity (9)
Signal to interference ratio at the input of SC microdiversity (3)
PDF of signal to interference ratio at the input of SC microdiversity (4)
CDF of signal to interference ratio at the input of SC microdiversity (5)
CDF of signal to interference ratio at the output of SC microdiversity (6)
CDF of signal at the output of SC macrodiversity (9-18)

Figure 3
Algorithm for outage probability calculation

The figure presents the necessary steps for obtaining the required expression, and based on the proposed system model given in Fig. 1. Necessary steps for obtaining the required expression are related to equations in the paper.

Macrodiversity SC receiver selects microdiversity SC receiver with the highest average power of the output signal. Therefore, cumulative distribution function of the macrodiversity SC receiver output signal is [25]:

$$
\begin{aligned}
&F_{\eta}\left(\eta / S_1 S_2\right)=\int_0^{\infty} d\Omega_1 \int_0^{\Omega_1} d\Omega_2 F_{\eta_1}\left(\eta/\Omega_1, S_1\right) p_{\Omega_1\Omega_2}\left(\Omega_1\Omega_2\right)+\\
&+\int_0^{\infty} d\Omega_2 \int_0^{\Omega_2} d\Omega_1 F_{\eta_2}\left(\eta/\Omega_2, S_2\right) p_{\Omega_1\Omega_2}\left(\Omega_1\Omega_2\right)=\\
&=F_{\eta}\left(\eta/S_1\right)+F_{\eta}\left(\eta/S_2\right)
\end{aligned}
\tag{9}
$$

By substituting (7) and (8) into (9) and by using (5), $F_{\eta}(\eta/S_1)$ can be expressed as:

$$
\begin{aligned}
&F_{\eta}\left(\eta/S_1\right)=\int_0^{\infty} d\Omega_1 \int_0^{\Omega_1} d\Omega_2 F_{\eta_1}\left(\eta/\Omega_1, S_1\right) p_{\Omega_1\Omega_2}\left(\Omega_1\Omega_2\right)=\\
&=\frac{1}{\Gamma(c)\left(1-\rho^2\right)\rho^{\frac{c-1}{2}}\Omega_0^{c+1}} \sum_{i_1=0}^{\infty}\left(\frac{\rho}{\Omega_0\left(1-\rho^2\right)}\right)^{2i_1+c-1} \frac{1}{i_1!\Gamma\left(i_1+c\right)}\left(\Omega_0\left(1-\rho^2\right)\right)^{i_1+c} \times\\
&\times \frac{1}{i_1+c}\left(\frac{1}{\Omega_0\left(1-\rho^2\right)}\right)^{i_1+c} \sum_{j_1=0}^{\infty} \frac{1}{\left(i_1+c+1\right)_{\left(j_1\right)}}\left(\frac{1}{\Omega_0\left(1-\rho^2\right)}\right)^{j_1} \cdot I_1
\end{aligned}
\tag{20}
$$

where $(a)_n$ denoting the Pochhammer symbol [30] and:

$$
\begin{aligned}
I_1&=\int_0^{\infty} d\Omega_1 \Omega_1^{2i_1+2c-1+j_1} e^{-\frac{2\Omega_1}{\Omega_0\left(1-\rho^2\right)}} \frac{1}{\left(1+\frac{\Omega_1}{S_1 y^2}\right)^2}=\\
&=\left(S_1\eta^2\right)^{2i_1+2c-1+j_1} \int_0^{\infty} dy\, y^{2i_1+2c-1+j_1} e^{-\frac{2S_1\eta^2 y}{\Omega_0\left(1-\rho^2\right)}} \frac{1}{(1+y)^2}
\end{aligned}
\tag{31}
$$

By using formula for confluent hypergeometric function $U(a,b,z)$ [30]:

$$
U(a,b,z)=\frac{1}{\Gamma(a)}\int_0^{\infty} e^{-zt} t^{a-1} \frac{1}{(1+t)^{a+1-b}} dt \tag{42}
$$

the integral I_1 can be written in the form:

$$I_1 = \left(S_1\eta^2\right)^{2i_1+2c-1+j_1}\Gamma\left(2i_1+2c-1+j_1\right)\times$$
$$\times U\left(2i_1+2c+j_1,\ 2a_1+2c+j_1-1,\ \frac{2S\eta^2}{\Omega_0\left(1-\rho^2\right)}\right) \tag{53}$$

After substitution (13) in (10), $F_\eta(\eta/S_1)$ becomes:

$$F_\eta\left(\eta/S_1\right) = \frac{1}{\Gamma(c)\left(1-\rho^2\right)\rho^{\frac{c-1}{2}}\Omega_0^{c+1}}\sum_{i_1=0}^{\infty}\left(\frac{\rho}{\Omega_0\left(1-\rho^2\right)}\right)^{2i_1+c-1}\frac{1}{i_1!\Gamma\left(i_1+c\right)}\frac{1}{i_1+c}\times$$
$$\times\sum_{j_1=0}^{\infty}\frac{1}{\left(i_1+c\right)_{(j_1)}}\frac{1}{\left(\Omega_0\left(1-\rho^2\right)\right)^{j_1}}\eta^{2(2i_1+2c+j_1)}\Gamma\left(2i_1+2c+j_1-1\right)S_1^{2i_1+2c+j_1}\times \tag{64}$$
$$\times U\left(2i_1+2c+j_1,\ 2a_1+2c+j_1-1,\ \frac{2S_1\eta^2}{\Omega_0\left(1-\rho^2\right)}\right)$$

The cumulative distribution function of η can be calculated by averaging (14), *F*η(η) becomes:

$$F_\eta\left(\eta\right) = \int_0^\infty dS_1F_\eta\left(\eta/S_1\right)p_{S_1}\left(S_1\right) =$$
$$= \frac{1}{\Gamma(c)\left(1-\rho^2\right)\rho^{\frac{c-1}{2}}\Omega_0^{c+1}\Gamma(c)S_0^{c_1}}\sum_{i_1=0}^{\infty}\left(\frac{\rho}{\Omega_0\left(1-\rho^2\right)}\right)^{2i_1+c-1}\frac{1}{i_1!\Gamma\left(i_1+c\right)}\times \tag{75}$$
$$\times\frac{1}{i_1+c}\sum_{j_1=0}^{\infty}\frac{1}{\left(i_1+c+1\right)_{(j_1)}}\frac{1}{\left(\Omega_0\left(1-\rho^2\right)\right)^{j_1}}\eta^{2(2i_1+2c+j_1)}\Gamma\left(2i_1+2c+j_1-1\right)\cdot I_2$$

where $F_\eta(\eta/S_1)$ is given by (14), $p_{S1}(S_1)$ is defined over (4) and I_2 is

$$I_2 = \int_0^\infty dS_2S_2^{2i_1+2c+j_1+c_1-1}e^{-\frac{1}{S_0}S_1}U\left(2i_1+2c+j_1,\ 2i_1+2c+j_1-1,\ \frac{2S_1\eta^2}{\Omega_0\left(1-\rho^2\right)}\right) \tag{86}$$

By using the formula [30]:

$$\int_0^\infty dtt^{b-1}e^{-St}U\left(a,c,t\right) = \frac{\Gamma(b)\Gamma(b-c+1)}{\Gamma(a+b-c+1)}\,{}_2F_1\left(b,\ b-c+1;\ a+b-c+1;\ 1-S\right) \tag{97}$$

the expression for $F_\eta(\eta)$ becomes:

$$\begin{aligned}
F_\eta(\eta) &= \frac{1}{\Gamma(c)\left(1-\rho^2\right)\rho^{\frac{c-1}{2}}\Omega_0^{c+1}\Gamma(c)S_0^{c_1}}\sum_{i_1=0}^{\infty}\left(\frac{\rho}{\Omega_0\left(1-\rho^2\right)}\right)^{2i_1+c-1}\frac{1}{i_1!\Gamma\left(i_1+c\right)}\times\\
&\times\frac{1}{i_1+c}\sum_{j_1=0}^{\infty}\frac{1}{\left(i_1+c\right)_{(j_1)}}\frac{1}{\left(\Omega_0\left(1-\rho^2\right)\right)^{j_1}}\eta^{2(2i_1+2c+j_1)}\Gamma\left(2i_1+2c+j_1-1\right)\times\\
&\times\left(\frac{\Omega_0\left(1-\rho^2\right)}{2\eta^2}\right)^{2i_1+2c+j_1+c_1}\frac{\Gamma\left(2i_1+2c+j_1+c_1\right)\Gamma\left(c+2\right)}{\Gamma\left(2i_1+3c+j_1+2\right)}\times\\
&\times{}_2F_1\left(\begin{matrix}2i_1+2c+j_1+c_1,\ c+2;\\ 2i_1+3c+j_1+2;\ 1-\dfrac{1}{S_0}\dfrac{\Omega_0\left(1-\rho^2\right)}{2\eta^2}\end{matrix}\right)
\end{aligned} \tag{108}$$

where ${}_2F_1$ is hypergeometric function.

In Table 1, the number of terms to be summed in order to achieve accuracy at the desired significant digit is depicted [31] in derived expression for outage probability. As we can see from the table, how increases parameter c increases the number of terms to be summed in order to achieve accuracy at the 5th significant digit. For higher values of parameter ρ, higher number of terms to achieve accuracy at the 5th significant digit is required.

Table 1

Terms need to be summed in the expression for cumulative distribution function to achieve accuracy at the 5th significant digit presented

Ω_0=1, S_0=1	c=1	c=2	c=3
ρ=0.2	58	51	44
ρ=0.4	60	51	44
ρ=0.6	66	56	48
ρ=0.8	78	62	56

4 Numerical Results

Outage probability of the studied macrodiversity system, is defined as the probability that the output SIR (Signal-Interference-Ratio) falls below a given threshold γ_{th}, and it can be expressed as $P_{out}(\gamma_{th})$ = F(γ_{th}). CDF $F(\gamma_{th})$ was defined by Eq. 18. Numerical results were obtained by implementing expression Eq. 18, in the Wolfram Mathematica software package [32]. The algorithm for calculating the CDF based on which the program code with parameter values is defined is given in Fig. 4.

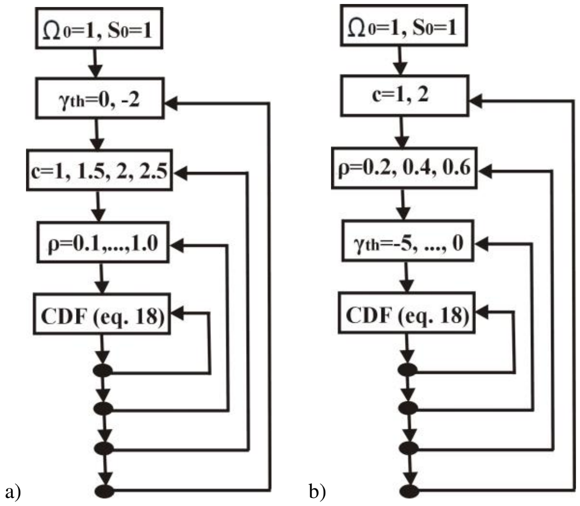


Figure 4

Algorithm for calculating the CDF of the analyzed system: a) for Figure 5, b) for Figure 6

The pseudo code for calculating the results in Fig. 5 is:

```
Ω0=1;
S0=1;
for (c=1, c<=2, c=c+1):
    for (ρ=0.2, ρ<=0.6, ρ=ρ+0.2):
        for (γth=5, γth>=0, γth=γth-1):
            calculating Eq. (18);
```

The pseudo code for calculating the results in Fig. 6 is:

```
Ω0=1;
S0=1;
for (γth=0, γth>=-2, γth=γth-2):
    for (c=1, ρ<=2.5, c=c+0.5):
        for (ρ=0.1, ρ<=1.0, ρ=ρ+0.1):
            calculating Eq. (18);
```

Outage probability of the studied macrodiversity system, and for several values of Gamma long-term fading severity parameter *c* and Gamma long-term correlation coefficient ρ is plotted in Fig. 5 and Fig. 6.

The outage probability increases as the threshold γ_{th} increases, and goes up to 1 for high values of γ_{th}. When severity parameter *c* increases, the outage probability decreases which improves system performance. The influence of the severity parameter *c* on outage probability is higher for lower values of γ_{th}. Values of the correlation coefficient ρ can range from 0 to 1. When the value of the correlation coefficient ρ is lower, especially when it is in the range of values close to 0, the received signals at the branches of the studied system are less correlated. It results in a higher probability of the correct signal detection, which consequently improves the system performance and makes probability outage lower. In the opposite case, for higher values of the correlation coefficient ρ the outage probability is higher for the same threshold values, as it could be seen in Fig. 5. In general, influence of the correlation coefficient ρ on the outage probability is higher for lower values of γ_{th} and moderately depends on the values of the parameter *c*, which can also be seen in Fig. 5.

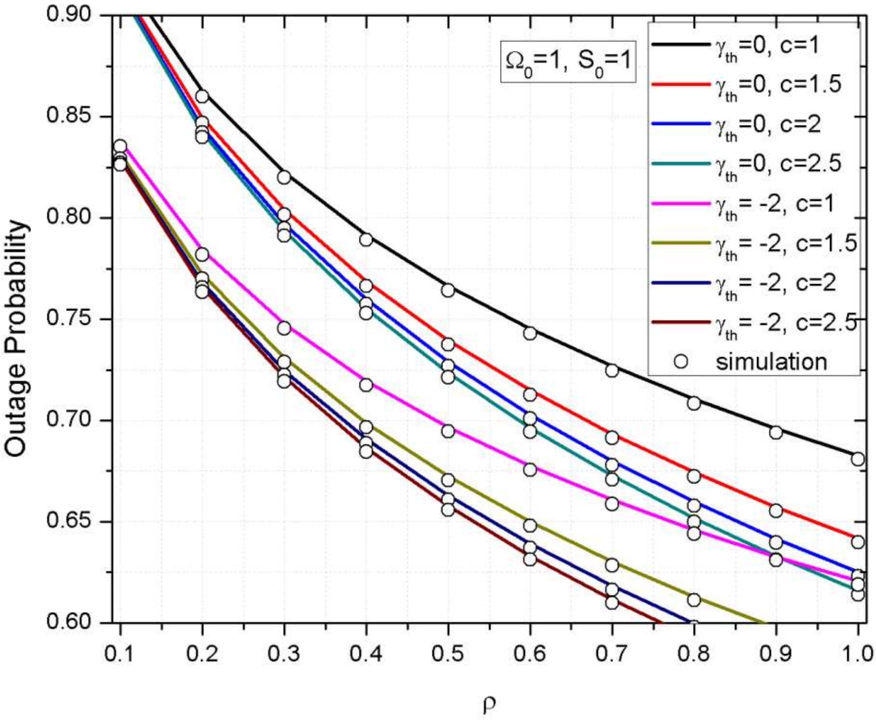


Figure 5
Outage probability versus the outage threshold for several values of *c* and ρ

The outage probability change with correlation coefficient ρ is depicted in Fig. 6. It shows that outage probability decreases when parameter ρ increases. Moreover, it is much more pronounced at the lower values of the parameter ρ. Also, it can be seen that for the higher values of parameter *c* the lower outage probability values are obtained. If the values of parameter *c* are compared, it can be seen that the difference between outage probability values is greater at lower values of parameter *c*. The behavior of the outage probability depending on the parameter ρ for the various values od the parameter *c* is identical for the various threshold values γ_{th}, with the an exception that the lower outage probability values are being obtained for the lower threshold values γ_{th}.

Numerical results were confirmed by Monte Carlo simulations using Matlab software package. Simulation results follow the results obtained by numerical calculation.

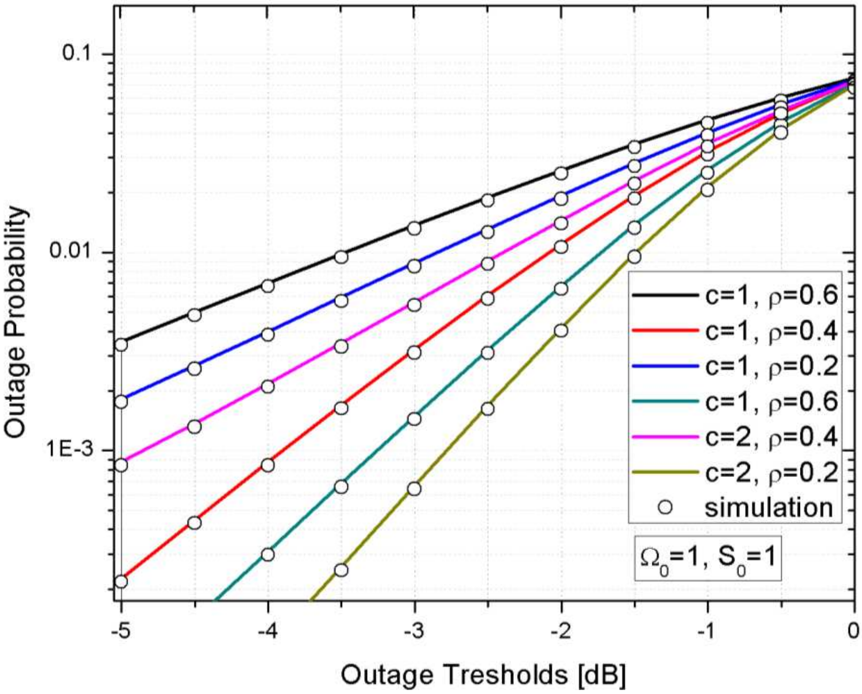


Figure 6

Outage probability versus the correlation coefficient ρ

Conclusions

In this paper, a macrodiversity system, composed of a macrodiversity SC receiver and two microdiversity SC receivers operating over Gamma shadowed Rayleigh multipath fading channel in the presence of Rayleigh co-channel interference, is considered. Macrodiversity receiver mitigates Gamma long-term fading effects, while microdiversity receivers mitigate short-term fading effects resulting in outage probability decreasing. When long-term fading severity parameter goes to infinity Gamma shadowed Rayleigh multipath channel becomes Rayleigh multipath channel, while, when Gamma long-term fading correlation coefficient goes to 1, macrodiversity system becomes microdiversity system. When correlation coefficient goes to 1, the weakest signal occurs at the both base stations, simultaneously. The closed form expression for cumulative distribution function of the microdiversity SC receiver output is derived, and then capitalizing on it, the outage probability expression of the macrodiversity SC receiver output signal is also derived. When long-term fading severity parameter goes to 1, the expression for outage probability of the studied macrodiversity system in the presence of Gamma shadowed Rayleigh multipath fading becomes expression for outage probability of macrodiversity system in the presence of Rayleigh fading. Obtained numerical results are presented graphically to show long-term fading severity parameter and correlation coefficient influences on the considered macrodiversity system outage probability. As the Gamma severity parameter increases, outage probability decreases and as the correlation coefficient decreases, the outage probability also decreases. Outage probability increases when the threshold value increases. The influence of the Gamma severity parameter and correlation coefficient, are higher for the lower threshold values.

References

[1] S. N. Moiseev, S. A. Filin, and M. S. Kondakov: Prediction of the Standard Deviation of the Signal-to-Noise Ratio in a Data-Transmission System with Orthogonal Subcarrier Frequencies, Journal of Communications Technology and Electronics, 2008, Vol. 53, pp. 794-798, https://doi.org/10.1134/S1064226908070097

[2] S. Panic, M. Stefanovic, J. Anastasov, and P. Spalevic, Fading and Interference Mitigation in Wireless Communications, 2013, CRC Press, USA

[3] A. K. Gupta, J. G. Andrews, and R. W. Heath: Macrodiversity in Cellular Networks with Random Blockages, IEEE Transactions on Wireless Communications, 2018, Vol. 17, No. 2, pp. 996-1010, https://doi.org/10.1109/TWC.2017.2773058

[4] S. K. Yoo, S. L. Cotton, W. G. Scanlon, and G. A. Conwa: An Experimental Evaluation of Switched Combining Based Macro-Diversity for Wearable Communications Operating in an Outdoor Environment, IEEE Transactions on Wireless Communications, 2017, Vol. 16, No. 8, pp. 5338-5352, https://doi.org/10.1109/TWC.2017.2709298

[5] V. B. Kreindelin, D. Y. Pankratov: Analysis of the Radio Channel Capacity of a MIMO System under the Conditions of Spatially Correlated Fadings, Journal of Communications Technology and Electronics, 2019, Vol. 64, pp. 863-869, https://doi.org/10.1134/S1064226919080242

[6] D. Krstic, B. Jaksic, M. Gligolirijevic, D. Stefanovic, and M. Stefanovic: Performance of Diversity System Output Signal in Mobile Cellular System in the Presence of α-μ Short Term Fading and Gamma Long Term Fading, Radioengineering, 2016, Vol. 25, No. 4, pp. 757-762, https://doi.org/10.13164/re.2016.0757

[7] B. Jaksic, M. Stefanovic, D. Aleksic, D. Radenkovic, and S. Minic: First-Order Statistical Characteristics of Macrodiversity System with Three Microdiversity MRC Receivers in the Presence of k-μ Short-Term Fading and Gamma Long-Term Fading, Journal of Electrical and Computer Engineering, 2016, Vol. 2016, Article ID 9689586, http://dx.doi.org/10.1155/2016/9689586

[8] G. L. Stuber: Mobile communication, 2003, Kluwer Academic Publisher, Dordrecht

[9] M. D. Yacoub: The η-μ distribution and the κ−μ distribution, IEEE Antennas and Propagation Magazine, 2007, Vol. 49, No. 1, pp. 68-81, https://doi.org/10.1109/MAP.2007.370983

[10] G. Malmgre: On the Performance of Single Frequency Networks in Correlated Shadow Fading, IEEE Transactions on Broadcasting, 1997, Vol. 43, No. 2, pp. 155-165, https://doi.org/10.1109/11.598364

[11] H. Surawereea, R. Luie, Y. Karagiannidis, G. Li, and B. Vucetic: Two Hop Amplify-and-Forward Transmission in Mixed Rayleigh and Rician Fading Channels, IEEE Communications Letters, 2009, Vol. 13, No. 4, pp. 227-229, https://doi.org/10.1109/LCOMM.2009.081943

[12] C.-F. Juang, Y.-Y. Lin, and R.-B. Huang: Dynamic system modeling using a recurrent interval-valued fuzzy neural network and its hardware implementation, Fuzzy Sets and Systems, 2011, Vol. 179, Iss. 1, pp. 83-99, https://doi.org/10.1016/j.fss.2011.05.015

[13] R. Precup, T. Teban, A. Albu, A. Borlea, I. A. Zamfirache, and E. M. Petriu: Evolving Fuzzy Models for Prosthetic Hand Myoelectric-Based Control, IEEE Transactions on Instrumentation and Measurement, 2020, Vol. 69, No. 7, pp. 4625-4636, https://doi.org/10.1109/tim.2020.2983531

[14] C. Pozna, and R.-E. Precup: Applications of Signatures to Expert Systems Modelling, Acta Polytechnica Hungarica, 2014, Vol. 11, No. 2, pp. 21-39, https://doi.org/10.12700/APH.11.02.2014.02.2

[15] R. Zall, and M. R. Kangavari: On the Construction of Multi-Relational Classifier Based on Canonical Correlation Analysis, International Journal of Artificial Intelligence, 2019, Vol. 17, No. 2, pp. 23-43

[16] E.-L. Hedrea, R.-E. Precup, R.-C. Roman, and E. M. Petriu: Tensor product-based model transformation approach to tower crane systems modeling, Asian Journal of Control, 2021, pp. 1-11, https://doi.org/10.1002/asjc.2494

[17] W. Cheng, and C. Juang: A Fuzzy Model With Online Incremental SVM and Margin-Selective Gradient Descent Learning for Classification Problems, IEEE Transactions on Fuzzy Systems, 2014, Vol. 22, No. 2, pp. 324-337, https://doi.org/10.1109/tfuzz.2013.2254492

[18] V. Milenkovic, N. Sekulovic, M. Stefanovic, and M. Petrovic: Effect of Microdiversity and Macrodiversity on Average Bit Error Probability in Gamma Shafowed Rician Fading Channels, ETRI Journal, 2010, Vol. 32, No. 3, pp. 463-467, https://doi.org/10.4218/etrij.10.0209.0448

[19] N. Sekulovic, M. Stefanovic, D. Milovic, and S. Stanojcic: Second-Order Statistics of System with N-branch Microdiversity and L-branch Macrodiversity Operating over Gamma Shadowed Nakagami-m Fading Channels, International Journal of Communication Systems, 2014, Vol. 27, No. 2, pp. 390-400, https://doi.org/10.1002/dac.2369

[20] M. C. Stefanovic, D. L. Draca, A. S. Panajotovic, and N. M Sekulovic: Performance Analysis of System with L-branch Selection Combining over Correlated Wibull Fading Channels in the Presence of Co-channel Interference, International Journal of Communication Systems, 2010, Vol. 23, No. 2, pp. 139-150, https://doi.org/10.1002/dac.1050

[21] M. C. Ju, and K. S. Hwang: Outage Equivalence of Opportunistic Relaying and Selection Cooperation in Presence of Co-Channel Interference, IEEE Transactions on Wireless Communications, 2015, Vol. 14, No. 6, 2981-2991, https://doi.org/10.1109/TWC.2015.2398871

[22] B. Bhargav, C. R. N. da Silva, Y. J. Chun, S. L. Cotton, and M. D. Yacoub: Co-Channel Interference and Background Noise in κ - μ Fading Channels, IEEE Communications Letters, 2017, Vol. 21, No. 5, pp. 1215-1218, https://doi.org/10.1109/LCOMM.2017.2664806

[23] M. D. Yacoub: The α - η - κ - μ Fading Model, IEEE Transactions on Antennas and Propagation, Vol. 64, No. 8, pp. 3597-3610, https://doi.org/10.1109/TAP.2016.2570235

[24] N. Djordjević, B. Jakšić, A. Matović, M. Matović, and M. Smilić: Moments of Microdiversity EGC receivers and Macrodiversity SC Receiver Output Signal over Gamma Shadowed Nakagami-m Multipath Fading Channel, Journal of Electrical Engineering - Elektrotechnický časopis, 2015, Vol. 66, No. 6, pp. 348-351, https://doi.org/10.2478/jee-2015-0058

[25] M. Perić, B. Jakšić, D. Aleksić, D. Randjelović, and M. Stefanović: Outage Probability of Macrodiversity Reception in the Presence Fading and Weibull Co-Channel Interference, Tehnički vjesnik - Technical Gazette, 2018, Vol. 25, No. 2, pp. 376-381, https://doi.org/10.17559/TV-20161227102847

[26] D. Bandjur, B. Jaksic, S. Panic, M. Bandjur, A. Matovic, and E. Mekic: Transmission over kappa-mu Fading channels with Gamma distributed

random Line-of-sight components, Revue Roumaine des Sciences Techniques - Série Électrotechniqueet Énergétique, 2017, Vol. 62, No. 2, pp. 179-184

[27] Đ. Banđur, B. Jakšić, A. Raičević, B. Popović, and M. Banđur: Performance Analysis of an IEEE 802.15.4 Network Operating Under κ–μ Fading, Interference and AWGN, Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 2020, https://doi.org/10.1007/s40998-020-00329-1

[28] Y. Chen and C. Tellambura: Performance Analysis of Three-Branch Selection Combining over Arbitrarily Correlated Rayleigh-Fading Channels, IEEE Transactions on Wireless Communications, 2005, Vol. 4, No. 3, pp. 861-865, https://doi.org/10.1109/TWC.2005.847109

[29] B.Jaksic, D. Stefanovic, M. Stefanovic, P. Spalevic, and V. Milenkovic: Level Crossing Rate of Macrodiversity System in the Presence of Multipath Fading and Shadowing, Radioengineering, 2015, Vol. 24, No. 1, pp. 185-191, https://doi.org/10.13164/re.2015.0185

[30] I. S. Gradshteyn, and I. M. Ryzhik: Table of Integrals, Series and Products, 2000, Academic Press, San Diego, USA

[31] B. S. Jakšić: Level Crossing Rate of Macrodiversity SC Receiver with two Microdiversity SC Receivers over Gamma Shadowed Multipath Fading Channel, Facta Universitatis, Ser. Autom. Control Robot., 2015, Vol. 14, No. 2, pp. 87-98

[32] Wolfram Mathematica: https://www.wolfram.com/

Uncertainty Analysis in the Notch Impact Test, for Materials with Different Energy Levels

Bulent Aydemir

Tubitak National Metrology Institute (TUBITAK UME) Force Laboratory,
Gebze-Kocaeli, Turkey, bulent.aydemir@tubitak.gov.tr

Abstract: The notch impact test, that is used to determine the amount of absorbed energy consumed, for breaking materials, under dynamic forces, is among the various important material mechanical tests. The quality and reliability of the results obtained, as a result of the test, are of great value for laboratories. In this study, the parameters affecting the measurement quality in the notch impact test, are expressed as uncertainty values. Charpy notch impact test was performed for materials with different absorbed energy levels in the laboratory. The uncertainty values were calculated, according to the ISO 148-1 standard, from the data obtained. Then, the uncertainty values were analyzed and their effects on the uncertainty calculation were determined. Herein, it was concluded that the parameters should be determined and analyzed correctly, in order to minimize measurement uncertainties.

Keywords: Charpy test; uncertainty calculation; notch impact test; ISO 148-1

1 Introduction

Among the mechanical characterization tests, the Charpy notch impact test is used to determine the mechanical properties of materials that work under conditions that can cause brittle fracture. The aim of the impact test in general is to determine the amount of energy and characterizing the fracture behavior of the materials under dynamic forces. This test is a simple, low-cost and reliable test that is mandatory for pipeline manufacturers, bridge construction industry and pressure vessel manufacturers. The impact test can be applied to most materials, but this paper only deals with metallic materials. Charpy notch impact test has been used extensively in mechanical testing of steel products, in research, and in purchase specifications for over four decades [1-3]. Today, there are two standards for general use of Charpy test for metallic materials. One of these standards is "Metallic Materials - Charpy Pendulum Impact Test" ISO 148-1:2016, the other is "Standard Test Methods for Notched Bar Impact Testing of Metallic Materials" ASTM E23:2018. The Charpy test is used to determine the amount of energy absorbed during fracture. The Charpy test is most commonly used to evaluate the properties of steels, based on characterizing fracture behavior or impact toughness and is therefore, often used in quality control applications, where it is a fast and economical test. A Charpy impact test measurement setup is given in Figure 1 [4]. It is used more as a comparative test rather than a definitive test. Some of the advantages of Charpy test include;

- This test is relatively easy to perform
- It is used to evaluate the toughness properties of materials
- It is useful for evaluating new products
- It provides fast measurement results

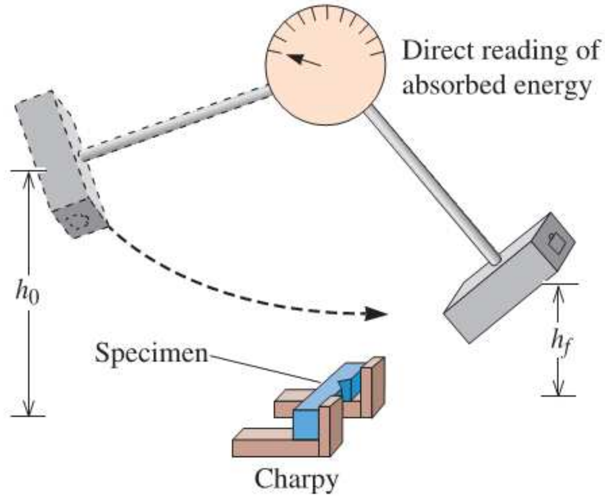


Figure 1
Principle of Charpy notch impact test machine [4]

When reporting the result of a test and analysis measurement, some quantitative indicators of the quality of the result should be given. This way, those who use it can evaluate the reliability of the test. Without such an indicator, measurement results are very difficult to compare among themselves or with reference values given in a specification or standard. Measurement uncertainty, which indicates the quality of a measurement result, is an easily applied, easily understood and generally accepted approach [5-8].

Laboratories operating under ISO/IEC 17025 accreditation and related systems are accordingly required to evaluate measurement uncertainty for test results. Taking into account all uncertainty components, which are of importance in the given situation, is obligatory, when reporting the measurement uncertainty.

Requirements for laboratories that want to be accredited in the notch impact test "General requirements for the competence of testing laboratories" ISO / IEC 17025:2017 standard has been defined. It includes the evaluation, approval and subsequent inspection of the technical competence of the laboratory according to the necessary criteria by an internationally recognized and authorized organization in order to ensure that the tests and analyzes carried out with ISO / IEC 17025 accreditation can provide confidence. Measurement uncertainty calculation is a desirable requirement.

As it is known, the indicator of measurement quality and the reliability of measurement results is the uncertainty value. For each device calibrated, an uncertainty value for that device is defined in its certificate. According to the accredited test report, it is necessary an uncertainty value must be calculated within the values obtained as a result of the test.

In this study, the uncertainty calculation, definitions, calculation formulas and sample numerical calculations for the notch impact test results are given in detail in sections. Then, the Charpy notch impact test was performed for materials with different absorbed energy levels in the laboratory. Uncertainty values were calculated according to the relevant standard for measurement data. Then, the uncertainty values were analyzed and their effects on the uncertainty calculation were determined. The aim of the study is to reveal what needs to be done in order to obtain a lower uncertainty value for notch impact testing.

2 Uncertainty Calculation

Global comparability is based on international comparisons of Charpy reference machines and approved values of certified reference test materials produced by national or international bodies using reference machines. The traceability chain begins internationally with the definition of the absorbed energy measured in the procedures described in the ISO 148-1 standard. Calibration laboratories use certified reference test material to validate reference machines. Besides, at the user level, Charpy test laboratories can verify their pendulum with reference test material to obtain reliable absorbed energy values. When research is conducted within the scope of the measurement uncertainty of the energy value that is absorbed by the specimen during testing, it is seen that the uncertainty analysis and effective factors are detailed in the ISO 148-1 standard. The 2018 version of the ASTM E23 standard does not include an uncertainty approach for the Charpy impact test. In this section, uncertainty calculations and effective factors specified in ISO 148-1, are explained in detail.

2.1 Factors Contributing to Uncertainty

Effective factors that contribute to measurement uncertainty in Charpy impact test are listed below:

a) Machine bias obtained from indirect verification

b) Homogeneity of test material and machine repeatability

c) Test temperature

Measurement equation for average absorbed energy (*KV*) is defined as formula 1:

$$KV = \bar{x} - Bv - Tx \tag{1}$$

Descriptions of expressions used in formula 1 are as follows:

$\bar{x}$ is observed averaged absorbed energy of n test specimens

Bv is machine bias based on indirect verification

Tx is bias due to temperature

2.2 Machine Bias

Machine bias is one of the effective factors that contribute to measurement uncertainty in Charpy impact test, so it is determined by indirect verification which is defined in ISO 148-2:2016 standard, as given in formula 2 below:

$$Bv = KV_v - KV_R \tag{2}$$

Descriptions of expressions in formula 2 are as follows:

KV_V is mean value of reference test specimens fractured during indirect verification

KV_R is certified value of reference test material

The uncertainty value of device deviation is calculated using the formulas given in ISO 148-2.

$$u(Bv) = \sqrt{u_v^2(x) + u_{RM}^2} \tag{3}$$

u (BV) is standard uncertainty of machine bias

u_v *(x)* is standard uncertainty of indirect validation results

u_{RM} is standard certificate uncertainty of reference test samples

$$u_v = \sqrt{u^2(Bv) + Bv^2} \tag{4}$$

Bv is machine bias based on indirect verification,

u_V is standard uncertainty of indirect verification results

2.3 Machine Repeatability and Material Homogeneity

Uncertainty of machine repeatability $u(\overline{x})$ is determined by using formula 5.

$$u(\overline{x}) = \frac{s_X}{\sqrt{n}} \tag{5}$$

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{6}$$

n is number of tested samples, x_i is absorbed energy value of the tested sample, $\overline{x}$ is the average absorbed energy value of n samples, s_X is the standard deviation of absorbed energy values obtained on n test samples. The value s_X is caused by two factors, one is machine repeatability and the other is specimen material homogeneity.

2.4 Temperature Bias

Effect of temperature bias *Tx* on absorbed energy is extremely material dependent. If steel is tested in brittle-to-ductile transition region, small changes in temperature can correspond to large differences in absorbed energy.

2.5 Machine Resolution

Effect of machine resolution is in most cases negligible in comparison with other uncertainty contributions. An exception is the case where machine resolution is large and measured energy is low. In that case, corresponding uncertainty contribution of machine resolution $u(r)$ is calculated using formula 7:

$$u(r) = \frac{r}{\sqrt{3}} \tag{7}$$

r is machine resolution in formula (7)

2.6 Combined and Expanded Uncertainty

In order to calculate combined uncertainty value of Charpy impact test measurement *u(KV)*, factors contributing to test uncertainty should be combined using formula (8) below.

$$u(KV) = \sqrt{u^2(\overline{x}) + u_V^2 + u^2(r)} \tag{8}$$

Finally, expanded uncertainty value of Charpy impact test measurement *U(KV)* is calculated using formula (9) below.

$$U(KV) = k \cdot u(KV) = t_{95}(v_{\overline{KV}}) \cdot u(KV) \tag{9}$$

$t_{95}(v_{\overline{KV}})$ is fraction value having a 95% confidence level corresponding to measurement degrees of freedom.

For expanded uncertainty calculation, first of all, measurement degree of freedom value $v_{\overline{KV}}$ should be calculated using formula (10) below.

$$v_{\overline{KV}} = \frac{u^4(\overline{KV})}{\frac{u^4(\overline{x})}{v_{\overline{x}}} + \frac{u_v^4}{v_v}} \tag{10}$$

v_v is degrees of freedom corresponding with u_v, $v_{\overline{x}}$ is degrees of freedom corresponding with $u(\overline{x})$.

Then the fraction value corresponding to the degrees of freedom calculated by the formula (10) and having a 95% confidence level is determined. Table 1 has given *tp(v)* fraction values corresponding to measurement degrees of freedom.

Table 1
Determination of *tp (v)* value according to degree of freedom

Degrees of freedom, *v*	*tp (v)* for fraction *P = 95%*
1	12.71
2	4.30
3	3.18
…	…
7	2.36
8	2.31
9	2.26
10	2.23
…	…
…	…

100	1.98
∞	1.96

3 Test Methods and Analysis of Uncertainty

The Charpy notch impact tests were performed by using a machine 300 J capacity. The impact testing machine has a certification, which was verified both with the direct and indirect verification according to ISO 148-2.

Charpy impact test specimens with different energy levels prepared concerning to ISO 148-1 standard. The dimensions of the all samples were checked against the criteria specified in ISO 148-1: length (55.00 ± 0.06) mm, height (10.00 ± 0.11) mm, width (10.00 ± 0.075) mm, notch angle (45 ± 2)°, height remaining at notch root (8.00 ± 0.075) mm, radius at notch root (0.25 ± 0.025) mm, distance between the plane of symmetry of the notch and the longitudinal axis of the test piece (27.50 ± 0.42) mm. The surface roughness of the test samples is better than 5 µm and meets all other requirements.

The measurement results of Charpy impact test specimens with different energy levels prepared according to ISO 148-1 standard are presented in Table 2 below. In table, measurement results of five Charpy impact test specimens with six group (E1, E2, E3, E4, E5, E6) different energy levels, their mean value, their standard deviation value and uncertainty value have been presented as Joule unit respectively.

Table 2

Measurement results of test specimens

Code	Measurement Results, J					Mean Value, x	Standard Deviation, s_x	Uncertainty, $u(x)$
	1	2	3	4	5	J	J	J
E1	21.2	21.7	22.7	21.8	22.3	21.94	0.57	**0.26**
E2	30.4	37.2	29.5	31.6	31.8	32.10	3.00	**1.34**
E3	67.2	65.6	68.4	63.5	68.8	66.70	2.20	**0.98**
E4	105.5	110.5	111.9	105.6	112.1	109.12	3.32	**1.44**
E5	124.3	119.8	122.8	123.9	114.5	121.06	4.07	**1.82**
E6	171.5	177.1	168.5	167.7	177.5	172.46	4.64	**2.08**

Calculations of statistical expressions located in Table 2 are explained in detail below. First, the arithmetic average of measurement results should be calculated for every code. The standard deviation of measurement results in Table 2 are calculated using following formula (6). Then the standard uncertainty of absorbed energy of measurement results should be calculated using formula 5.

When measurement results in Table 2 are examined, resolution of Charpy impact test machine is determined to be 0.1 Joule. Thus, uncertainty component of machine resolution is calculated using following formula (7).

$$u(r) = \frac{r}{\sqrt{3}} = \frac{0.1}{\sqrt{3}}$$

$u(r)$ = 0.06 J

In order to determine standard uncertainty of indirect verification measurement, reference samples with known certificate values should be used, which are given in Table 3 below.

Table 3

Certificate value of reference test specimen

Code	Certificate Number	Traceability	Certificate value, KV_R	Degree of Freedom	Certificate Uncertainty (k=2), U_{RM}
-	-	-	J	$V_{V\ RM}$	J
R1	XXX	XXX	123.8	7	3.4

Certified reference material has been fractured using Charpy impact testing machine with the method specified in ISO 148-2 standard. Measurement results are also given in Table 4 below.

Table 4

Measurement results of reference test sample

Measurement Results, J					Mean Value, KV_V	Standard Deviation, s_x	Uncertainty, $u_v(x)$
1	2	3	4	5	J	J	J
123.1	116.1	112.8	123.6	121.3	119.4	4.7	2.1

Calculations of statistical expressions located in Table 4 are explained in detail below. Arithmetic average, standard deviation and uncertainty values of measurement results should be calculated as similar Table 2.

Using to formula (2), the estimate of Charpy impact test machine bias is calculated as follows,

$$Bv = KV_v - KV_R = 119.4 - 123.8$$

Bv = - 4.4 J

Also, standard uncertainty of pendulum impact testing machine bias is determined by using formula (3) specified in ISO 148-2 standard below.

$$u(Bv) = \sqrt{u_v^2(x) + u_{RM}^2} = \sqrt{2.1^2 + (3.4/2)^2}$$

u(Bv) = 2.71 J

As a general rule, bias should be corrected, however due to wear of anvils and hammer parts; it is difficult to obtain a perfectly stable bias value throughout period between two indirect verifications. This is why measured bias value is often considered an uncertainty contribution, to be combined with its own uncertainty to obtain uncertainty of indirect verification result.

$$u_v = \sqrt{u^2(Bv) + B_V^2} = \sqrt{2.71^2 + (-4.4)^2}$$

u_v = 5.19 J

In order to calculate combined uncertainty value of Charpy impact test measurement, factors contributing to uncertainty should be combined as defined in ISO 148-1 standard below,

$$u_{E1,E2,E3,E4,E5,E6}(KV) = \sqrt{u^2(\bar{x}) + u_V^2 + u^2(r)} = \sqrt{u^2_{E1,E2,E3,E4,E5,E6}(\bar{x}) + 5.19^2 + 0.06^2}$$

Measurement degree of freedom value is calculated using formula 10 below.

$$v_{\overline{KV}} = \frac{u^4(\overline{KV})}{\dfrac{u^4(\bar{x})}{v_{\bar{x}}} + \dfrac{u_v^4}{v_v}} = \frac{u^4_{E1,E2,E3,E4,E5,E6}(\overline{KV})}{\dfrac{u^4_{E1,E2,E3,E4,E5,E6}(\bar{x})}{5-1} + \dfrac{5.19^4}{7}}$$

Concerning to Table 1, t_p values are determined as Table 5. Finally, expanded uncertainty of *U(KV)* is determined using formula 9 as follows.

$$U(KV) = k \cdot u(KV) = t_{95}(v_{\overline{KV}}) \cdot u(KV)$$

All uncertainty parameters calculated in numerical are summarized in Table 5 below.

Table 5

For different energy levels, calculated measurement uncertainties and values of related parameters

Code	KV$_R$	U_{RM}	*Bv*	*u(Bv)*	u_v	*u(r)*	*u(x)*	*u(KV)*	V_{KV}	*U(KV)*
	J	J	J	J	J	J	J	J	-	J
E1	123.8	3.4	-4.4	2.71	**5.19**	**0.06**	**0.26**	5.19	7.0	**12.00**
E2	123.8	3.4	-4.4	2.71	**5.19**	**0.06**	**1.34**	5.36	7.9	**12.64**
E3	123.8	3.4	-4.4	2.71	**5.19**	**0.06**	**0.98**	5.28	7.5	**12.33**
E4	123.8	3.4	-4.4	2.71	**5.19**	**0.06**	**1.44**	5.38	8.0	**12.44**
E5	123.8	3.4	-4.4	2.71	**5.19**	**0.06**	**1.82**	5.50	8.6	**12.53**
E6	123.8	3.4	-4.4	2.71	**5.19**	**0.06**	**2.08**	5.59	9.0	**12.63**

4 Discussion

The following equation was used to calculate the expanded measurement uncertainty value based on the Charpy notch impact test measurements.

$$U(KV) = k.\sqrt{u^2(\overline{x}) + u_V^2 + u^2(r)} \tag{11}$$

When the measurement uncertainty values given in Table 5 are analyzed, since all samples are tested on the same calibrated impact device, u_v and *u(r)* values are taken into account as the same value. It is only *u(x)* whose value changes in the measurement uncertainty calculation. In order to better understand the parameters that affect this uncertainty calculation, the effect percentages are given in Figure (2). It is the parameter that has the greatest impact in uncertainty calculation was u_v with an average value of 80% in Figure (2). The average effect of the *u (x)* value is 19% and the average effect of the *u (r)* value is 1%.

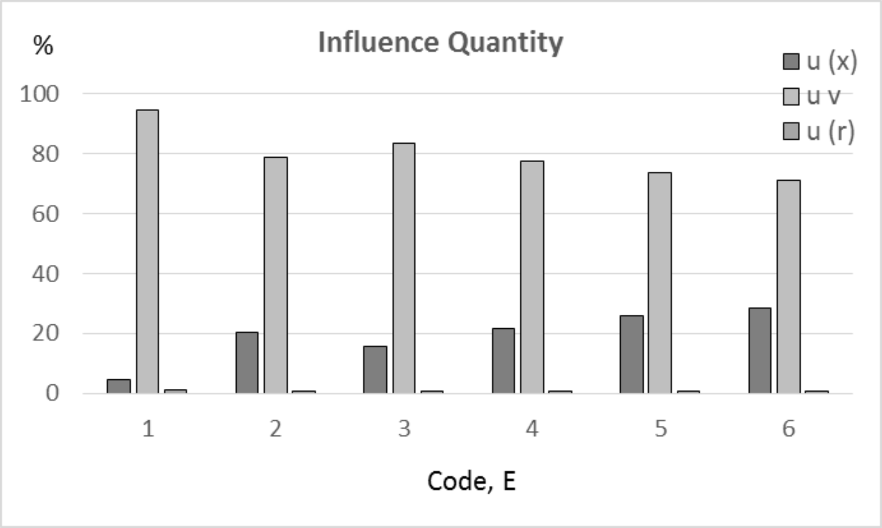


Figure 2

Influence quantity in the uncertainty calculation for code E1, E2, E3, E4, E5, E6 samples

For a more detailed analysis of the u_v parameter, which is the greatest impact in the uncertainty calculation, the calculation of the u_v parameter should be examined. The following equation is used to calculate the measurement uncertainty of indirect verification from the Charpy notch impact test machine.

$$u_v = \sqrt{u^2(Bv) + Bv^2} = \sqrt{u_v^2(x) + u_{RM}^2 + Bv^2}$$

As can be seen from the equation, three parameters are included in the uncertainty calculation. To see the changes in these parameters, the Charpy notch impact tester used in the tests was calibrated using reference test pieces with different energy levels. The values and results of the reference test pieces belonging to different energy levels used in the calibration of the notched notch impact test machine are given in Table 6. Reference test pieces (CRM) are shown with the codes R2, R3, R4. Using these reference pieces, the measurements were taken on the same Charpy impact testing machine and the $u_v(x)$ and *Bv* values were determined.

The effect percentages of the parameters that affect the standard uncertainty of the indirect validation results in Table 6 are given in Figure (3). U_{RM} was the greatest impact parameter in the uncertainty between 35% and 55% in the calculation of indirect verification results. Similar results were obtained in the study by NIST [9] [10]. In figure (3), the average effect of the *Bv* value is 34% and the average effect of the $u_v(x)$ value is 20%. As a result, the parameters that have a major impact on the uncertainty calculation are determined as the uncertainty value of the U_{RM} reference test pieces and the deviation error in the *Bv* reference test sample values.

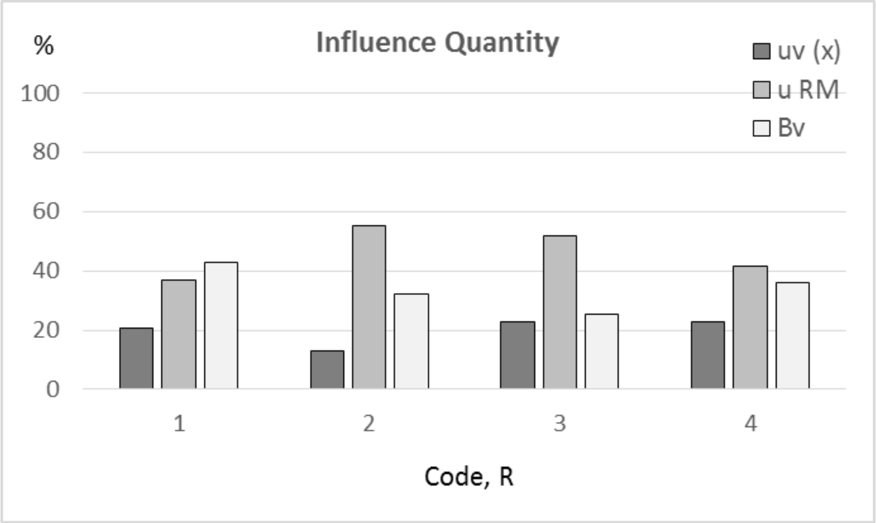


Figure 3

Influence quantity in the uncertainty calculation for code R1, R2, R3, R4 of CRMs

As a result, the parameters that are the impact in the uncertainty calculation in the notch impact testing machine and the percentages of their effects are given. There are different references for uncertainty calculations for notch impact testing [11-14]. However, there is no analysis study on uncertainty calculation in this area. The laboratory that wants to reduce the uncertainty value of the notch impact test can solve its problem by examining the calculations given in this study. In this study, the uncertainty calculation in the notch impact test was analyzed and the greatest impact parameters for the uncertainty calculation related to the subject were presented to accredited laboratories and institutions using the Charpy notch impact test.

Table 6
For different energy levels of reference Charpy test pieces' materials, certificated and calculated uncertainties and values of related parameters

Code	Certificate value, KV_R	Degree of Freedom	k	Certificate Uncertainty, U_{RM}	Uncertainty, $u_v(x)$	Bv
-	J	$V_{V\,RM}$	t_p	J	J	J
R2	216.3	35	2.03	**8.6**	**2.0**	**5.0**
R3	195.6	54	2.01	**7.3**	**3.2**	**3.6**
R4	21.9	35	2.03	**2.2**	**1.2**	**1.9**

Conclusions

In this study, the general uncertainty calculation of the Charpy impact test is explained in detail, in accordance with the ISO 148-1 standard. For the absorbed energy value obtained, as a result of the Charpy impact test, the effective factors contributing to the measurement uncertainty calculation are specified. In order to provide a better understanding of the methods and formulas in the uncertainty approach, the uncertainty calculation for different energy level test samples, is presented in this work as well. Two groups of three-parameter uncertainty parameters, that are the impact in uncertainty calculating of the Charpy test are defined. Its effects on uncertainty calculation have been analyzed. In the first group, the parameters included in the uncertainty calculation are $u(x)$, u_v and $u(r)$ values. When all samples were tested on the same notch impact test machine, the u_v and $u(r)$ values were taken into account as the same value. In the measurement uncertainty calculation, only the $u(x)$ value determined from the scattering of the test results, was taken into account, as the variable value. For the parameters included in the uncertainty calculation in the first group, the u_v value was observed to be the greatest impact parameter in the uncertainty calculation, with an average of 80%. In the second group, the parameters of $u_v(x)$, u_{RM} and Bv values were examined in the uncertainty calculation of the u_v value. By using different reference test specimens, measurements were taken using the same Charpy notch impact test machine and $u_v(x)$ and Bv values were determined. The u_{RM} value was calculated as the greatest impact parameter in the second group uncertainty parameters, with an average of 46%. Then the average effect of the Bv value was calculated as 34%.

To develop the uncertainty value, which is an indicator of measurement quality, it is necessary to analyze the uncertainty calculation well. According to the analysis given in this study, in order to improve the notch impact test uncertainty value, it is necessary to reduce the greatest impact parameters. In this study, the uncertainty calculation, in the notch impact test, was examined and effective parameters, to reduce the uncertainty value were presented. To reduce the uncertainty value of notch impact testers, u_v the greatest impact parameter should be reduced. Examining the calculation of u_v the most influential parameter comes from the uncertainty of the reference materials, u_{RM}. To reduce this parameter, choose a reference material with a lower uncertainty value. In this way, a reduction in the uncertainty calculation is achieved. Low uncertainty value can be attained by analyzing other effective parameters, similarly, in the uncertainty calculation.

Practically, the measurement uncertainty value, is calculated from the effective parameters that can cause a change in a measurement result. Measurement uncertainty, defines an uncertainty band at a 95% confidence level according to the uncertainty approach. By lowering the uncertainty value and narrowing the uncertainty band, it allows for results with higher quality and much closer to the average values.

References

[1] Aydemir, B. "Application of Measurement Uncertainty in Notch Impact Tests", 3nd International Conference on Material Science and Technology in Cappadocia (IMSTEC'18), Nevsehir (17-19/09/2018): 5 p.

[2] Aydemir, B., 2011, Training of calibration of notch impact test machines, G2KV-050, Oct. 2011, TUBITAK UME

[3] Aydemir, B., 2017, Training of uncertainty calculation for material testing-G2KV-120, May 2017, ISDEMIR, Iskenderun

[4] Askeland,_ D. R., Wright, W. J.,_2014, The Science and Engineering of Materials, 7th edition, Cengage Learning

[5] Evaluation of measurement data - Guide to the expression of uncertainty in measurement, JCGM 100:2008

[6] EA guidelines on the expression of uncertainty in quantitative testing, EA-4/16, G:2003

[7] T. M. Adams, A2LA Guide for Estimation of Measurement Uncertainty in Testing, July 2002, Guidance, G104

[8] Stephanie Bell, A Beginner's Guide to Uncertainty of Measurement, 1999

[9] Lucon, E., McCowan, C., "Impact Testing Yesterday and Today" ASTM Workshop, 13/11/2011, NIST

[10] J. D. Splett H. K. Iyer C.-M. Wang, C. N. McCowan, Special Publication 960-18 NIST Recommended Practice Guide: Computing Uncertainty for Charpy Impact Machine Test Results, 2008

[11] Lont, M. A., "The Determination of Uncertainties in Charpy Impact Testing," Manual of Codes of Practice for the Determination of Uncertainties in Mechanical Tests on Metallic Materials, UNCERT COP 06:2000, 2000

[12] Takagi, S. and Yamaguchi, Y., "Uncertainty Analyses of Reference Specimens for the Verification of Charpy Impact Test Machines," J. of Material Testing Research Association of Japan, Vol. 48, No. 4, 2003

[13] Splett, J. D. and Wang, C. M., "Uncertainty in Reference Values for the Charpy V-Notch Verification Program," J. Testing and Evaluation, 2002, pp. 362-369

[14] Roebben, G., Lamberty, A. and Pauwels J. “Certification of Charpy V-Notch Reference Test Pieces at IRMM” Journal of ASTM International, July/August 2005, Vol. 2, No. 7

Lumped Element Method – A Discrete Calculus Approach for Solving Elliptic and Parabolic PDEs

Zoltán Vizvári[1], Mihály Klincsik[1], Zoltan Sári[1], Péter Odry[2]

[1]University of Pécs, Szentágothai Research Centre
Ifjúság útja 20, H-7624 Pécs Hungary
vizvari.zoltan@mik.pte.hu, klincsik@mik.pte.hu, sari.zoltan@mik.pte.hu

[2]University of Dunaújváros, Institute of Information Technology
Táncsics M. u. 1/A, H-2401 Dunaújváros, Hungary, podry@uniduna.hu

Abstract: In this report, we introduce a novel discrete calculus method for obtaining the numerical solution of parabolic and elliptic type partial differential equations. The discrete operators applied during the process of obtaining the numeric solution have the same advantageous properties as those of their continuous counterparts: orthogonality, conservation laws, and minimum-maximum principle. The results of our proposed solution method are interpreted in terms of the underlying physics and the material properties on the same graph. This can significantly simplify the solution of the discretised system that originates from the advantageous properties of discrete operators defined on weighted graphs. We demonstrate the applicability of the presented approach by using it to calculate the numeric solutions of an elliptic and a parabolic model problem and compare these results to the solutions of the same problems calculated using a well-known FEM solver.

Keywords: Discrete Calculus; Elliptic and Parabolic Problems; Differential Operators on Graphs; Lumped Elements

1 Introduction

Elliptic and parabolic partial differential equations (PDEs) are fundamental in the mathematics of physical laws. The practical impact of the solution of these kinds of equations is profound since many important physical phenomena used in various industrial applications are described by them. These problems are in disparate fields which include thermal energy transport, diffusion, electrostatics, electrodynamics etc. The general formulation of the aforementioned equations mentioned is as follows:

$$\frac{\partial u}{\partial t} + \nabla \cdot (-\kappa \nabla \mathrm{u}) = s \qquad (1)$$

where $u = u(\boldsymbol{x}, t)$ is the unknown scalar field (potential), $s = s(\boldsymbol{x}, t)$ is the source term, $\kappa = \kappa(\boldsymbol{x}, t)$ is the isotropic transport coefficient, $\underline{\boldsymbol{x} = (x, y, z)}$ and t are the space and time coordinates respectively, ∇u denotes the gradient of the scalar field u, and $\nabla \cdot \boldsymbol{v}$ is the divergence of the vector field $\boldsymbol{v} = -\kappa \nabla u$. The application of the weighted residual method (WRM) is a popular approach for solving these kinds of physical problems [1]. The application of WRM leads to several different well-known numerical methods such as the Finite Difference Method (FDM), Finite Volume Method (FVM), Finite Element Method (FEM), etc. All of these methods lead to an approximate solution, which satisfies the PDE and the boundary conditions as well [2]. Concerning the method of approximation, there are different approaches such as the variation principle, Galerkin-method, or the integral equations using the Green-function [2].

Another kind of distinction can be made based on the specific basis functions utilised for the various methods. Taking these into consideration, the following categories can be derived: discrete, semi-discrete and continuous [3]. In the discrete approach, the resulting approximate solution is defined only at discrete locations (points) in the investigated domain, which results, for example, in FDM. In this case, the solution satisfies the discretised equation in the interior nodes of the domain and the boundary conditions at all the boundary nodes as well. Thus, the solution is referred to as the strong form solution [2]. Semi-discrete methods use continuous basis functions that are defined on each of the discrete sub-domains. The approximated solution consists of the linear combination of basis functions. The most popular method in this category is the FEM, which was first proposed by Ritz and Galerkin [2]. The finite element method determines the weak form solution to the governing PDE. This method was introduced first for problems in structural mechanics. Since then, researchers have exploited this technique for application to problems in other physical disciplines (for example fluid dynamics, heat transfer, etc.) [2]. In the continuous case (for example Analytic Element Method (AEM)), the approximated solution is defined on the entire domain (without discretisation). Currently, the AEM is the most applicable technique for groundwater simulations. In this case, AEM has some advantages over FDM/FEM which include precise estimation of the hydraulic head, generation of continuous flow solutions throughout the domain, and more accurate estimation of water budget, etc. [3].

Certainly, there are several other methods (for example Monte-Carlo Method [2], meshless, meshfree methods [2], Boundary Element Method (BEM) [2], etc.), but they are not relevant with regard to the purpose of this investigation.

During the numerical solution of PDEs several additional problems and difficulties can arise as follows:

- Conservation laws: in some cases, neither semi-discrete nor continuous methods guarantee global conservation or local conservation [4, 5],
- Handling of complicated spatial geometries [4, 6, 7],
- Discontinuity or abrupt changes of material properties over the physical domain [8, 9],
- Complex-valued material properties, parameters, and field quantities, which can be effectively addressed by the proposed numerical method described in the following sections [10].

1.1 Advantages of Discrete Calculus Methods

The Discrete Calculus (DC) numerical approaches are able to effectively capture the physics of the underlying PDEs. Thus, compared to other methods (FEM, BEM, etc.), the DC Methods are more favoured by researchers in the field. The key feature of the application is the exact discretisation of the underlying physics and calculus [11] (before making any approximations) to exploit all the discrete differential operators (they mimic the mathematical properties of the continuous differential operators). [12-23] In general, the implementation of DC methods consists of two main steps: [4, 24]

1. Discretisation process: the continuous PDE system is substituted by its discrete counterpart,
2. Approximation process: solving the discrete system on the discretised domain.

The nature of a PDE (orthogonality, conservation, wave propagation, etc.) never depends on the details of the material. Thus, the DC approach always represents the physics of the PDE by transferring all numerical approximation errors to the material properties [11]. All approximations of material properties are based on the constitutive equations, which are strong physical statements. Errors in material properties do not affect the physical properties of a PDE system (for example local conservation of energy).

In general, a numerical solution method actually uses two meshes [24]. These include DC methods, FEM, FVM, and FDM as well as staggered mesh schemes. The solution is approximated on one mesh, and the equations are approximated on the other [24].

In this report, we propose a novel DC method to determine the solution of PDEs of the elliptic and parabolic type in particular. In order to carefully explain the novel approach, our report focuses only on the unsteady diffusion equation (1), which has the complexity to present the fundamental ideas of our DC approach. In the following, we establish the foundations of a novel discretisation method, which can be applied to the numerical solutions of various physical problems in the form (1).

2 The Proposed Method

The novel method proposed by the authors incorporates the advantageous properties of the finite element and finite difference methods, thereby maintaining the simplicity of the discretisation approach of the finite difference method, while at the same time enabling arbitrary triangular discretisation of the domain.

The underlying principle of our method is to construct a topological operation, which forms a circular graph directly from a given physical domain (continuum). This way, not only the continuous domain is transformed into a graph, but the PDE of the problem will also be represented by discrete operators. The physical parameters become weights of the edges of the graph, and finally, the boundary conditions are represented by virtual edges connected to the boundary vertices of the graph (Fig. 1). By the application of discrete versions of the differential operators, the original problem can be directly discretised on a graph corresponding to an arbitrary triangular mesh. In this way, we do not attempt to obtain an approximate solution of the continuous problem, but instead, we solve the fully discretised version of the original problem directly and interpret the solution on the graph.

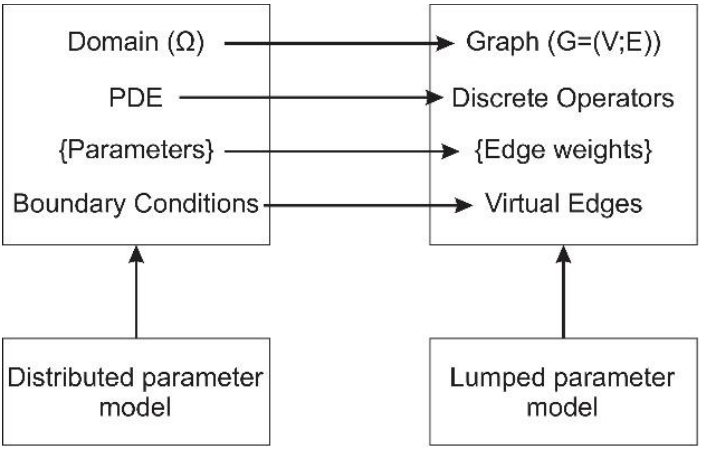


Figure 1

The concept of transformation

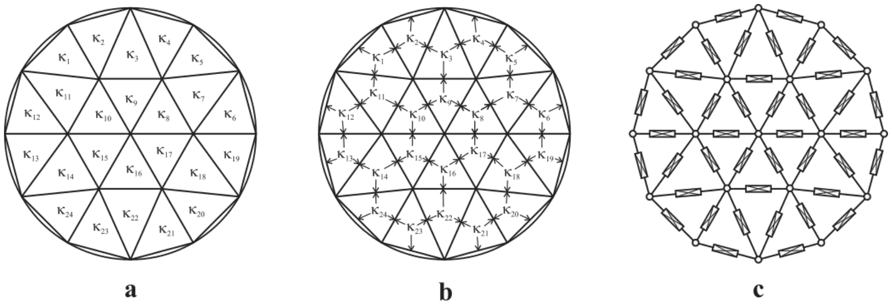


Figure 2
Transformation of the physical parameters κ_i from the triangular domain to the edges of the triangles $(a) \rightarrow (b)$, and representation by network (c)

As a result of this transformation, the material property values are lumped to the edges of the graph, and the original distributed parameter model becomes a discretised, lumped parameter model of the problem. As observed in Fig. 2, the continuous physical domain is subdivided into triangular domains, where the material properties (κ) are assumed to be constant. The interface conditions are handled during the discretization of κ, where similarly to the nature of Fourier-approximation of functions having discontinouinity the κ value on the interface will be equal to the average of the neighbouring domain values resulting in a kind of smoothing effect. Based on this approach, it is possible to interpret the physics (PDE) and the material properties on the same graph, which significantly simplifies the solution of the discretised system. The system of linear equations contains the properties of the Kirchhoff current law, so for interfaces, we assume continuity everywhere. This means that equal potential values are observed in the common nodes of triangles and the same fluxes in the case of their common sides.

2.1 The Process of 'Lumping'

In a three-dimensional case, Fig. 3(a) shows a tetrahedral element T_e on which the set of lumped parameters is constructed from material property κ on the domain.

The steps of the process are the follows:

1. Connect the centre of the circumsphere of the tetrahedral element with the middle-points of the edges. (Fig. 3(b))
2. The resulting sub-elements generally have irregular shapes, thus the cross-sectional area (A) perpendicular to the direction of flux is a function of the length along the direction of the flux. (Fig. 3(c))
3. The sub-elements can be handled as regular prisms with the same length and same volume (V) as the original sub-element, while also having a constant cross-sectional area $(\bar{A})$, which is the average of A. (Fig. 3(d))

4. This way the one-dimensional model of conductivity can be applied. (Fig. 3(e))

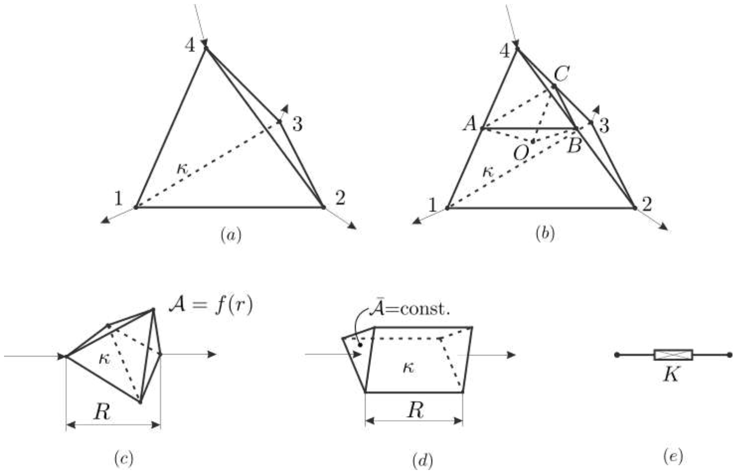


Figure 3
Process of constructing a lumped parameter K corresponding to a three-dimensional tetrahedral element with material property κ

The previously described process can be represented formally as follows:

$$K = \kappa \cdot \frac{\bar{A}}{R} = \kappa \cdot \frac{\frac{1}{R}\int_{T_e} A(r)dr}{R} = \kappa \cdot \frac{V}{R} \cdot \frac{1}{R} = \kappa \cdot \frac{V}{R^2} \tag{2}$$

where R is the radius of the circumsphere. In the two-dimensional case, the process is similar to the three-dimensional case as shown in Fig. 4, but we start with a triangular element T_e for this approach with a height h. The formula corresponding to the two-dimensional case is as follows:

$$K = \kappa \cdot \frac{\bar{m}\cdot h}{R} = \kappa \cdot \frac{\frac{1}{R}\int_{T_e} m(r)dr}{R} \cdot h = \kappa \cdot \frac{A}{R} \cdot \frac{1}{R} \cdot h = \kappa \cdot \frac{A}{R^2} \cdot h = \kappa \cdot \frac{V}{R^2} \tag{3}$$

where R is now the radius of the circumcircle. The calculation of the lumped parameter values must be performed for each sub-element in both the three-dimensional (3) and the two-dimensional (2) cases.

2.2 Graph Representation

The main advantage of constructing a graph from a continuum compared to a traditional discretisation method (like FEM, FDM) is that since the graph is independent of the underlying physical space, the abrupt changes of material properties and boundary structures are handled in a straightforward manner, regardless of the dimension of the underlying space. Moreover, all the relevant results from discrete calculus and graph theory can be applied. Fig. 5 shows a few possible surface geometries represented by the same graph after the topological transformation. The original domain shapes are rectangular (a), semi-cylindrical (b), disk (c), and hemispherical (d) respectively. Regardless of the shape of the original geometry, the matrices that represent the topology of the graph will be the same. The only difference is in the actual weights of the graph edges.

The definition and most important properties of the graph used are the followings

$$G = (V, E) \tag{4}$$

where V is the set of vertices and E is the set of edges, with weights $w: E \rightarrow \mathrm{C}$. The real and imaginary parts of the weights are physical properties corresponding to the material in the physical domain. The discretisation operator is defined as:

$$D: \Omega \times K \rightarrow G \tag{5}$$

where $\Omega \subset \mathrm{R}^n$ is the physical domain, n is the number of space-dimensions, K is the set of physical parameters, and G is the set of graphs with properties defined in (4). The operator defined in (5) can be applied to an ample set of physical problems (processes which can be represented by elliptic PDEs), and the discretisation always leads to a linear system of equations corresponding to a lumped parameter model.

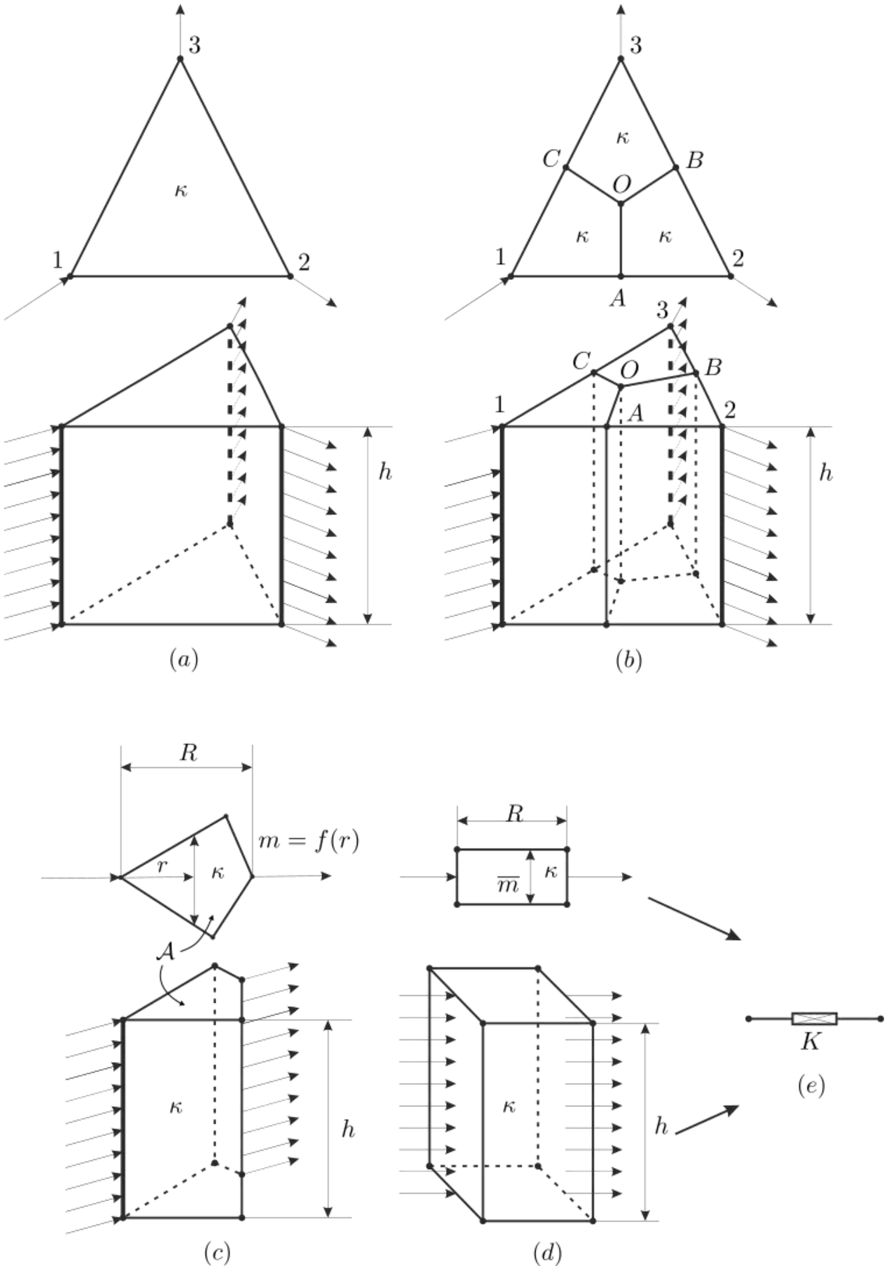


Figure 4
Process of constructing a lumped parameter K corresponding to a two-dimensional triangular element with material property κ

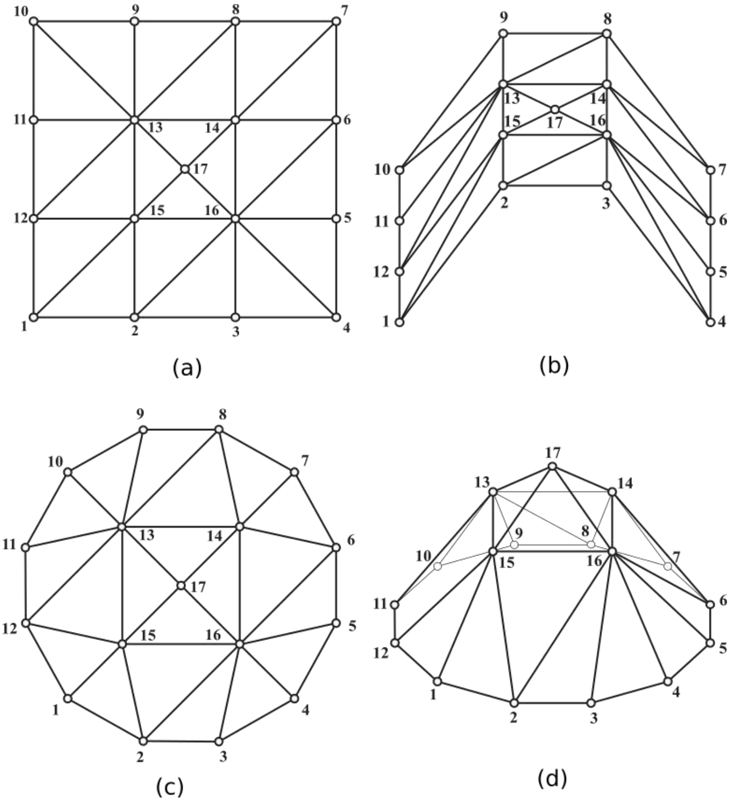


Figure 5

Different surface geometries represented by the same graph

Table 1 represents a comparison between the continuous and discrete versions of a few important differential operators.

Table 1

Analogy between continuous and discrete operators [25]

Name	Continuous	Discrete
Kirchhoff Current Law	$\nabla \cdot J = s$	$\mathbf{A}\cdot \mathbf{j}_b$=$\mathbf{j}_s$
Kirchhoff Voltage Law	$\nabla \times \nabla u = 0$	$\mathbf{B}^T\cdot \mathbf{u}_b$=**0**
Linear transport equation	$J = -\kappa \cdot \nabla u$	$\mathbf{j}_b$=-$\mathbf{K}\cdot \mathbf{A}^T\mathbf{u}_n$
Elliptic transport equation	$-\nabla \cdot (\kappa \cdot \nabla u) = s$	-($\mathbf{A}\mathbf{K}\mathbf{A}^T\cdot\mathbf{u}_n$)=$\mathbf{j}_s$

In Table 1 **J** is the extensive current density, **A** is the node-branch incidence matrix, **B** is the branch-loop incidence matrix, **K** is a diagonal matrix of weights on the edges of the graph, $\mathbf{j}_b$ is the extensive current on the branches, $\mathbf{j}_s$ is the extensive source current, $\mathbf{u}_n$ is the potential in the nodes, and $\mathbf{u}_b = \mathbf{A}^T\cdot\mathbf{u}_n$ is the potential difference (intensive quantity) on the branches. All of the discrete operators enlisted in Table 1 have the same advantageous properties (orthogonality, conservation, minimum-maximum principle) as those of their continuous counterparts [25]. Due to the strict approach, our proposed method guarantees that the resulting discrete operator always leads to a solvable system of linear equations (for example in the case of elliptic transport, the resulting discrete operator will be a Laplacian matrix interpreted on a graph).

To determine the solution of a PDE, we generally need a boundary condition. In the case of solving the discrete version of the problem on a graph (which is loosely connected to the geometry of the physical domain), we need to extend the graph using virtual branches (Fig. 6). This set of branches form the boundary, and all generally used boundary conditions (Dirichlet, Neumann, Outward/Inward Flux, etc.) can be defined on these branches.

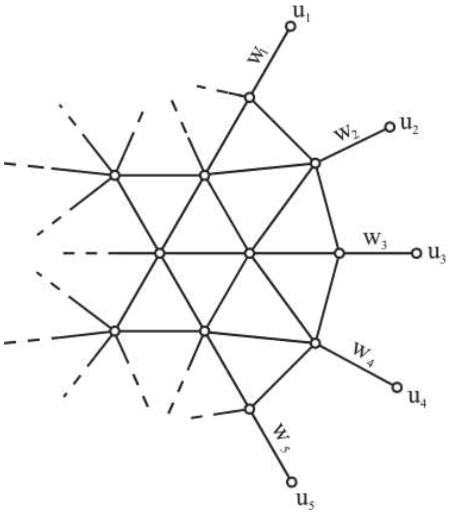


Figure 6
Boundary conditions represented by virtual branches

Currently, we have investigated the Neumann and Dirichlet boundary conditions only since these are the most frequently occuring ones. Even in case of heatflow problems, we can handle certain boundaries as Dirichlet types while others as Neumann type boundaries, but not the mixed version in our current implementation. As a future work, the inclusion of a third type boundary condition is also planned to make our method more flexible. During the definition of boundary conditions, the values of the weights w_i can be zero, or infinity as well. For example, a particular Dirichlet boundary condition can be handled by a virtual edge with $w_i \to \infty$ rendering the potential of the inside node equal to the prescribed potential of the outside node (i.e. ground node, $u_i = 0$ in the electric case). In the case of the graph representation, the handling of the sources can be performed by adding potential and current sources for any of the nodes of the graph (see Table 2).

Table 2
Fundamental boundary conditions defined by virtual branchesImplementation of the method

Boundary condition	**Weight**	**Potential**	**Flux**
Dirichlet	w	u	-
Neumann	w	-	i
Insulation	0	-	-

Concerning the special approach and the properties of the proposed method, we decided to describe the technique as the Lumped Element Method (LumEM). As an introductory model example, let us assume we have a sufficiently small domain of a 2D space, where the material properties can be assumed to be homogeneous. In the case of such a simple domain, the procedure of discretisation is given in the following section.

As the first step of discretisation the domain has to be partitioned into a triangular mesh (Delaunay triangulation), where the definitions are as follows:

- M: mesh
- M^0: nodes (vertices, points)
- M^1: branches (edges, lines)
- M^2: triangles (faces, loops)

In the following subsections, we examine the steps of discretisation for two simple geometries containing one and two triangles.

2.3 Simple Disk with One Triangle

In the case of a simple disk domain assuming homogeneous material properties κ containing one triangle, Fig. 7 shows the process of transforming the material property κ of the homogeneous material to the edges of a triangle. The methods shown below can be used during discretization for any triangle created by Delunay triangulization. The boundary of the domain has three nodes where extensive flux can flow into or out of the domain. The triangular domain is subdivided into three smaller domains (see A_1, A_2, A_3 in Fig. 7) by connecting the centre of the circumcircle to the midpoints of the sides.

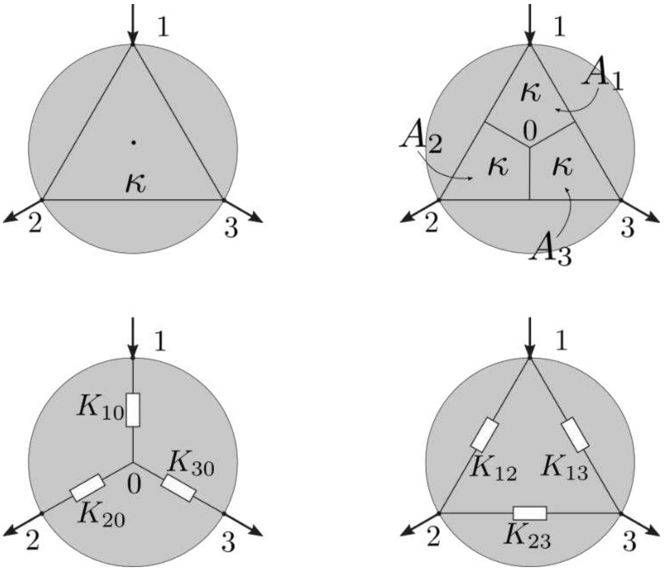


Figure 7
Development of a transport coefficient network for one triangle

After dividing the domain into three parts, each area can be replaced by an edge with a weight that concentrates the κ of the domain to the edge. The edges are constructed by connecting the vertices of the triangle to the centre of the circumcircle, hence the lengths of the resulting edges are equal. Mapping the material property κ to the edges can be done by the following set of rules

$$K_{i,0} = \kappa \frac{A_i}{R^2}\, h, \;\; i = 1,2,3 \tag{6}$$

where the resulting $K_{(i,0)}$ can be interpreted as the concentrated material properties which will form graph edges after a star-delta conversion as follows:

$$K_{i,j} = K_{j,i} = \frac{K_{i,0}K_{j,0}}{\sum_{i=1}^{n} K_{i,0}} = \kappa \frac{h}{R^2} \frac{A_i A_j}{\sum_{\ell=1}^{n} A_\ell} = \kappa\, t_{i,j}, \;\; (i,j = 1,2,3, i \neq j) \tag{7}$$

The weights $K_{(i,j)}$ on the resulting edges are equal because of the assumption of a homogeneous distribution of material properties on the disk. As such, the disk domain is transformed into a triangular subgraph.

2.4 Simple Disk with Two Triangles

In the next case, the disk-shaped domain is decomposed into two disjunct triangles (Fig. 8). The boundary of the domain has four nodes where extensive flux can flow into or out of the domain. The transport coefficients κ_1 and κ_2 on the triangular domains are assumed to be homogeneous inside each triangle. From this point onwards, the triangles are treated in a similar way to the case of the one-triangle example in the preceding section as seen in (8) and (9):

$$K_{i,P} = \kappa^{(1)} \frac{A_i^{(1)}}{\left(R^{(1)}\right)^2} h, \;\; (i = 1,2,3) \tag{8}$$

$$K_{i,Q} = \kappa^{(2)} \frac{A_i^{(2)}}{\left(R^{(2)}\right)^2} h, \;\; (i = 1,2,3) \tag{9}$$

where $R^{(1)}$, $R^{(2)}$ are the radii of the circumcircles of triangles 1 and 2 respectively.

After this, we perform a Y-Δ transformation on each triangle, described by the following formulas:

$$K_{i,j}^{(1)} = K_{j,i}^{(1)} = \kappa^{(1)}\, t_{i,j}^{(1)}, \;\; (i,j = 1,2,3, \;\; i \neq j) \tag{10}$$

$$K_{i,j}^{(2)} = K_{j,i}^{(2)} = \kappa^{(2)}\, t_{i,j}^{(2)}, \;\; (i,j = 1,2,3, \;\; i \neq j) \tag{11}$$

where $t_{i,j}^{(1)}$, $t_{i,j}^{(2)}$ are the factors corresponding to the subdomains 1 and 2 respectively, with the same form as introduced in (7). As a result of the Y-Δ transformation, there will be a pair of edges ($K_{1,3}^{(1)}$, $K_{1,3}^{(2)}$) which connect the same two nodes. In this case, the resulting transport coefficient is the sum of those of the branches according to the characteristics of the flow of the extensive flux.

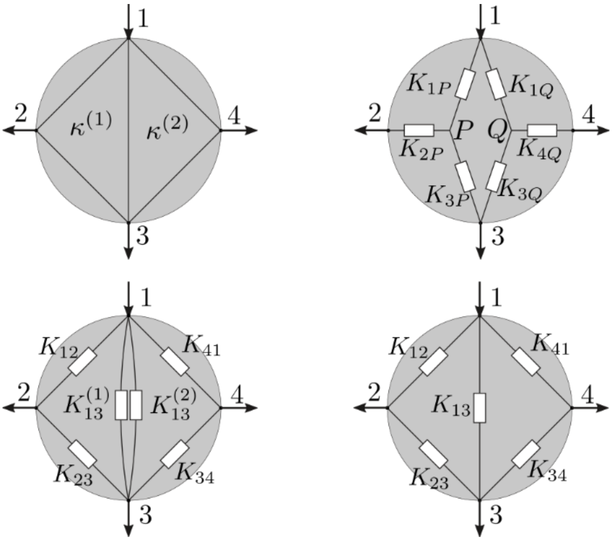


Figure 8

Development of a transport coefficient network for two adjacent triangles

Based on this outlined concept, the weights of the resulted graph are as follows:

$$\begin{cases} K_{1,2} = \kappa^{(1)} t_{1,2}^{(1)} \\ K_{2,3} = \kappa^{(1)} t_{2,3}^{(1)} \\ K_{3,4} = \kappa^{(2)} t_{3,4}^{(2)} \\ K_{4,1} = \kappa^{(2)} t_{4,1}^{(2)} \\ K_{1,3} = \kappa^{(1)} t_{1,3}^{(1)} + \kappa^{(2)} t_{1,3}^{(2)} \end{cases} \tag{12}$$

which can be written in matrix form as:

$$\begin{bmatrix} t_{1,2}^{(1)} & 0 \\ t_{2,3}^{(1)} & 0 \\ 0 & t_{3,4}^{(2)} \\ 0 & t_{4,1}^{(2)} \\ t_{1,3}^{(1)} & t_{1,3}^{(2)} \end{bmatrix} \begin{bmatrix} \kappa^{(1)} \\ \kappa^{(2)} \end{bmatrix} = \begin{bmatrix} k_{1,2} \\ k_{2,3} \\ k_{3,4} \\ k_{4,1} \\ k_{1,3} \end{bmatrix} \tag{13}$$

from this, a compact matrix form can be

$$\mathbf{T} \cdot \boldsymbol{\kappa} = \mathbf{k} \tag{14}$$

where κ is the vector of transport coefficients corresponding to triangular domains, **k** is the vector of weights of the resulting graph, and **T** is a transformation matrix. Equations (12) and (13) illustrate the essence of the material discretisation process described in Section 2, according to which the discontinuity of material properties taken as constant on the elements are smoothed at the connections of the elements.

The transformation matrix $\mathbf{T} \in \mathrm{R}^{|M^1| \times |M^2|}$ has only three non-zero values t_j in the j-th column, corresponding to the incident branches of the j-th triangle[1]. The representation of the boundary branches in the matrix ensures full column rank for any **T**. Because of this property, there exists a pseudo-inverse $\mathbf{T}^{+}$ of **T**, and (14) can also be solved for κ in view of **T** and **k**. Thus, the subdivision procedure described can be easily generalised to an arbitrary number of triangles.

3 Case Studies

To demonstrate the effectiveness and robustness of the method, case studies have been created for which the results obtained by the LumEM method are illustrated and compared with a well-known FEM solving procedure realized in COMSOL

[1] Matrix **T** is essentially a branch-loop incidence matrix.

Multiphysics environment. In the case studies, we will discuss in detail the implementation of the LumEM method. During discretization, using the same mesh for both methods, the LumEM and FEM potential values assigned at the nodes become easy and comparable. In addition, relative errors between the two methods are calculated and graphically interpreted. In the first case study, we construct a fundamental electrical problem and then modify it to generate a numerical example that demonstrates the robustness of the LumEM method, even with unrealistic material property variation. In the third case, we solve a fundamental heat transfer problem using the LumEM method in order to show that the method is also consistent with time-dependent parabolic PDEs. Of course, a comparison with FEM is carried out for each case study.

3.1 Basic Electric Problem

In this section, an application of the previously described discretisation method is presented for a simple electric problem, assuming time-harmonic functions. Let's assume a circular domain (with a diameter of 1 m) shown in Fig. 9 with given physical parameters.

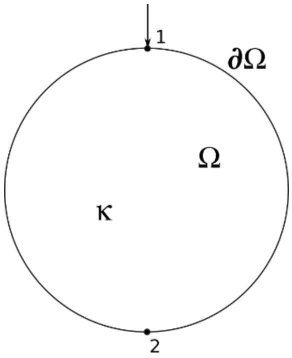


Figure 9

Simple domain

$$\kappa_i = \sigma_i + j\omega\varepsilon_i \tag{15}$$

where σ_i is the conductivity, ε_i is the relative permittivity, $i = 1 \ldots M^2$. After discretisation, the parameters corresponding to the branches of the resulting graph, are given as:

$$K_k = G_k + j\omega C_k \tag{16}$$

where G_k is the conductance, C_k is the capacitance, $k = 1 \ldots M^1$. The fundamental equation describing the problem examined is given as:

$$\nabla \cdot (\kappa \nabla u) = 0 \tag{17}$$

Since κ is a constant over the whole domain Ω, the following form can be written:

$$\Delta u = 0 \tag{18}$$

The mathematical model can be used to mimic a physical problem in which a voltage generator is connected to a single point (N_1) on the boundary. The ground point is also placed at a highlighted point on the boundary (N_2). The remaining part of the boundary (Γ) is formed from electrical insulating material. Another purpose of this case study is to demonstrate that the method can be easily and efficiently applied to point sources.

To define the boundary conditions, we denote the potentials on the boundary by $\phi(s)$, where the argument s is the normalised boundary length ($s \in [0,1]$). The applied boundary conditions are a Neumann-boundary $i_g\delta(s = 0)$ at node N_1 and a Dirichlet-boundary $\phi(s = 1/2) = \phi_0 = 0$ (ground) at node N_2. For all of the other boundary nodes (in Γ), the boundary condition is given as:

$$\mathbf{n} \cdot \nabla\phi(s) = 0 \quad s \in \Gamma \tag{19}$$

where $\partial\Omega = \Gamma + N_1 + N_2$.

The solution steps are as follows:

1. Construct a triangulation of the investigated domain (for example using the Distmesh algorithm [26]).
2. Transform the domain into a graph by applying (14).
3. Identify solution of the equation system

$$\mathbf{AKA^T\ u = i_g} \tag{20}$$

where $\mathbf{A} \in \{0,1,-1\}^{|M^0|\times|M^1|}$ is the reduced incidence matrix[2], $\mathbf{K} \in \mathrm{C}^{|M^1|\times|M^1|}$ is a diagonal matrix built from the elements of $\mathbf{k}$ (see 3.2 Section), $\mathbf{u} \in \mathrm{C}^{|M^0|}$ is the vector of potentials, and $\mathbf{i_g} \in \mathrm{C}^{|M^0|}$ is the vector of generator currents. The matrix $\mathbf{AKA^T}$ in (20) can be interpreted as a Laplace-operator (weighted Laplacian matrix) [25] for the nodes.

3.2 Basic Parabolic Problem

In this section, the discretisation method is applied to a basic time-dependent heat conduction problem. Similarly to the electric problem, we assume a circular domain (with a diameter of 1 m) shown in Fig. 10, with given physical parameters

$$\kappa_i = k_i \tag{21}$$

[2] The reduced incidence matrix is created by deleting the row corresponding to the ground node from the matrix **A** [70].

where k_i is the thermal conductivity.

The fundamental equation defining the problem is

$$\rho\, c_p \frac{\partial u}{\partial t} + \nabla \cdot (-\kappa \nabla u) \;=\; 0 \tag{22}$$

where u is the temperature, ρ is the density, c_p is the specific heat capacity, and κ is the thermal conductivity. Since κ is constant over the entire domain (22) can be written as:

$$\rho c_p \frac{\partial u}{\partial t} - \kappa\, \Delta\, u = 0 \tag{23}$$

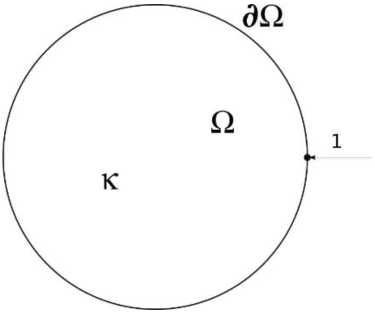


Figure 10
Simple domain for a parabolic problem

The boundary conditions applied are a Dirichlet-boundary ϕ_0 at node N_1. For all of the other boundary nodes, the boundary condition is given as:

$$\mathbf{n} \cdot \nabla\phi(s) = 0 \quad s \in \Gamma \tag{24}$$

where $\partial\Omega = \Gamma + N_1$. Of course, for non-homogeneous Dirichlet boundary conditions, the function is evaluated in the nodes on the boundary and these values are used to implement the solver.

The solution steps are as follows:

1. Construct a triangulation of the investigated domain (for example using the Distmesh algorithm [26]).
2. Transform the domain into a graph by applying (14).
3. Identify a solution of the following equation system by applying the backward-Euler method for discretisation of the time derivative.

$$\rho c_p \frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{\Delta\, t} - \mathbf{AKA}^{\mathbf{T}}(\mathbf{u}_{i+1}) = 0 \tag{25}$$

denoting the relevant time steps by indexes i and i+1. From this form, we can rearrange the terms to get:

$$-\frac{\rho c_p}{\Delta t}\mathbf{u}_i = \left(\mathbf{L}_\kappa - \frac{\rho c_p}{\Delta t}\mathbf{I}\right)\mathbf{u}_{i+1} = \mathbf{M}\ \mathbf{u}_{i+1} \tag{26}$$

where $\mathbf{L}_\kappa = \mathbf{AKA}^\mathrm{T}$ is the discrete Laplace operator (weighted Laplacian matrix) [70] containing the material property κ. Based on this result, with the introduction of the matrix $\mathbf{M}$, we have the following form:

$$\mathbf{u}_{i+1} = -\frac{\rho c_p}{\Delta t}\ \mathbf{M}^{-1}\ \mathbf{u}_i \tag{27}$$

The solution of (27) can be acquired by iterating over the time steps starting from an initial condition u_0 and taking into account the boundary conditions.

4 Results and Discussion

4.1 Solution of the Electric Problem

In the first example (in 4.1 Section) we have assumed a homogeneous material distribution over the whole domain, and the value of the transport coefficient κ defined in (15) is $\kappa = 1[S/m] + j\omega 1[F/m]$ (assuming a fixed $\omega = 1[1/s]$) and according to the two-dimensional nature of the problem, the domain is assumed to have unit height $h = 1[m]$. In the formula we have substituted the numerical value into epsilon, the construction of epsilon based on relative permittivity is also applicable without any complications.

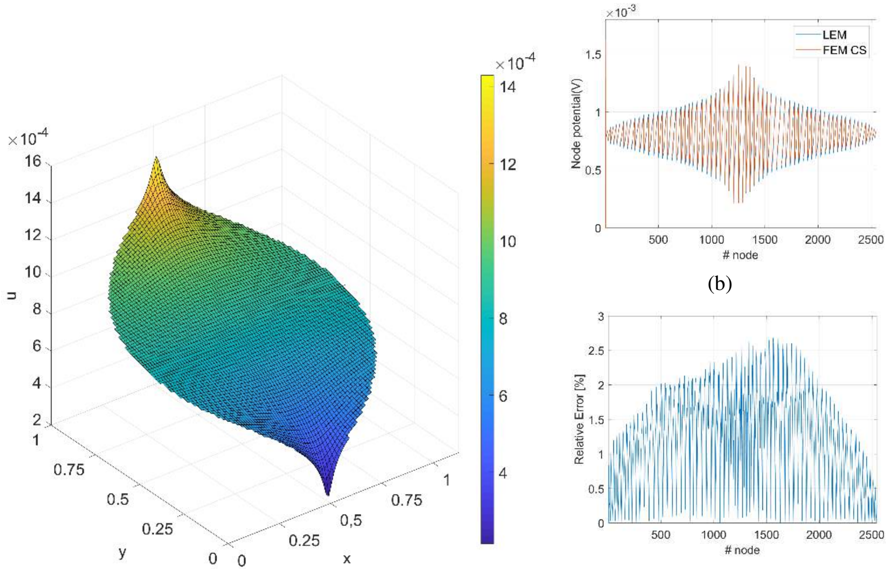
Solving (20) provides the potentials $\mathbf{u}$. The solution can be seen in Fig. 11 compared to a solution calculated using the COMSOL FEM environment. The figure shows the nodal values of the potential and the relative error calculated as

$$\epsilon = \frac{|u_n - u_{ref}|}{max(u_{ref})}. \tag{28}$$

It can be seen from the figure that there is a very good agreement (the maximum relative error is less than 3 %) between our solution and the reference solution (by COMSOL).

The next example covers a non-homogeneous material distribution, where there is a concentric disk $(x_0 = 0,\ y_0 = 0,\ r = 0.35)$ inside the domain, where the transport coefficient is $\kappa_1 = 10^{10} + 10^{10}j$, and the remaining area has a transport coefficient $\kappa = 1 + j$.

Fig. 12 shows a comparison of the calculated potential fields with the corresponding relative error. From the figure, it is clear that there is an excellent agreement for the non-homogeneous case as well since the maximum relative error is approximately 4 %.

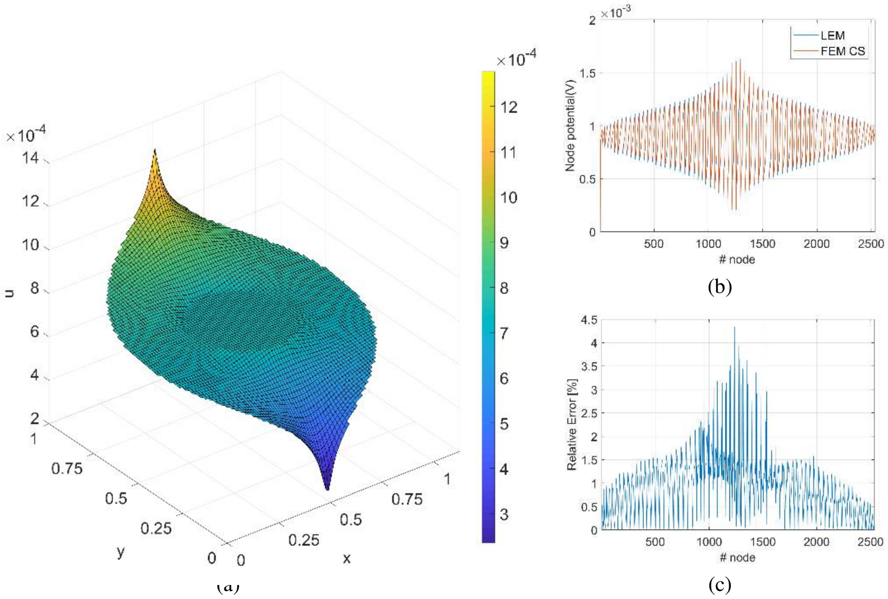


(b)

(c)

Figure 11

The potential distribution calculated by LumEM method (a), the comparison of numerically calculated potentials (b), and the corresponding relative errors (c) for the homogeneous case



(b)

(a)

(c)

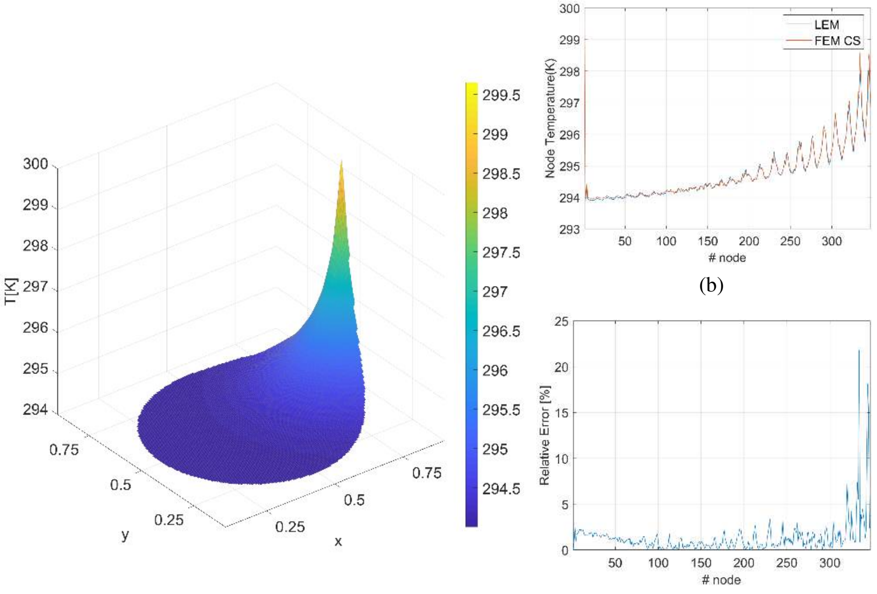
Figure 12

The potential distribution calculated by LumEM method (a), the comparison of numerically calculated potentials (b), and the corresponding relative errors (c) for the non-homogeneous case

4.2 Solution of the Heat Transfer Problem

The actual parameters used for the heat transfer problem discussed in Section 4.2 are the thermal conductivity $k = 1\ [W/(m\ K)]$, density $\rho = 1\ [kg/m^3]$, and specific heat capacity $c_p = 1\ [W/(kg\ K)]$. According to the two-dimensional nature of the problem, the domain is assumed to have a unit height $h\ = 1\ [m]$.

During the solution of (27), we assumed a 1 sec interval during time-stepping and we have investigated the solution at the last time step corresponding to t = 1 sec. Fig. 13 shows the solution compared to those obtained from COMSOL and the corresponding relative errors taking the COMSOL solution as the basis of the comparison. The average relative error is 1.32%. The maximal values of the errors located in the close vicinity of the boundary node (large node indexes), where the temperature changes very sharply according to a Dirac-like boundary condition.



(b)

(a)

(c)

Figure 13

The temperature distribution calculated by LumEM method (a), the comparison of numerically calculated temperature values (b), and the corresponding relative errors (c) for the heat transfer problem

Conclusion and Future Work

In this report, we have introduced a new discrete calculus-based approach for the solution of elliptic and parabolic type linear partial differential equations. Based on the special properties of the proposed method, it was called the Lumped Element Method. This is because the approach relies on a special topological transformation from the continuous domain to a weighted graph containing lumped parameters that correspond to the material properties. We implemented the approach in the solution of two different model problems and investigated the results of the numerical solutions. It was determined that the proposed method is applicable to obtain the numerical solution of linear elliptic and parabolic PDEs. A future research goal is the generalisation of the method to a broader class of PDEs (hyperbolic, non-linear, etc.) and the investigation of the possible applications of the method to more complicated topologies and 3D problems. In addition, the method provides an opportunity to extend it towards analytical solving methods with one dimension or even a higher dimension [27].

Acknowledgement

This research was funded by project EFOP-3.6.1-16-2016-00004 and 2020-4.1.1-TKP2020, co-financed by the Ministry for Innovation and Technology in Hungary and the European Union.

References

[1] Hatami, M., (2017) Weighted Residual Methods. Elsevier Science Publishing Co Inc. [30] Hiptmair, R., 2001, Discrete hodge-operators: An algebraic perspective. Progress In Electromagnetics Research 32, 247-269

[2] Mazumder, S., (2015) Numerical Methods for Partial Differential Equations. Elsevier Science Publishing Co Inc.

[3] Majumder, P., Eldho, T. I., (Feb 2016) A new groundwater management model by coupling analytic element method and reverse particle tracking with cat swarm optimization. Water Resources Management 30 (6), 1953-1972

[4] Asaithambi, R., Mahesh, K., (Jul 2017) A note on a conservative finite volume approach to address numerical stiffness in polar meshes. Journal of Computational Physics 341, 377-385

[5] Shi, C., Shu, C.-W., (Jun 2018) On local conservation of numerical methods for conservation laws. Computers & Fluids 169, 3-9

[6] Petras, A., Ling, L., Ruuth, S., (Oct 2018) An RBF-FD closest point method for solving PDEs on surfaces. Journal of Computational Physics 370, 43-57

[7] Serkh, K., Rokhlin, V., (Jan 2016) On the solution of elliptic partial differential equations on regions with corners. Journal of Computational Physics 305, 150-171

[8] Coco, A., Russo, G., (May 2018) Second order finite-difference ghost-point multigrid methods for elliptic problems with discontinuous coefficients on an arbitrary interface. Journal of Computational Physics 361, 299-330

[9] Hou, T. Y., Hwang, F.-N., Liu, P., Yao, C.-C., (May 2017) An iteratively adaptive multi-scale finite element method for elliptic PDEs with rough coefficients. Journal of Computational Physics 336, 375-400

[10] Toth, F., Kaltenbacher, M., (Oct 2016) Time and frequency domain finite element implementation of a fully anisotropic linear visco-elastic model. PAMM 16 (1), 397-398

[11] Lipnikov, K., Manzini, G., Shashkov, M., (Jan 2014) Mimetic finite difference method. Journal of Computational Physics 257, 1163-1227

[12] Frank, J., Moore, B. E., Reich, S., (Jan 2006) Linear PDEs and numerical methods that preserve a multisymplectic conservation law. SIAM Journal on Scientific Computing 28 (1), 260-277

[13] Nicolaides, R. A., Wu, X., (Dec 1997) Covolume solutions of threedimensional div-curl equations. SIAM Journal on Numerical Analysis 34 (6), 2195-2203

[14] Nicolaides, R., (Jun 1989) Flow discretization by complementary volume techniques. In: 9th Computational Fluid Dynamics Conference. American Institute of Aeronautics and Astronautics

[15] Perot, B., (Mar 2000) Conservation properties of unstructured staggered mesh schemes. Journal of Computational Physics 159 (1), 58-89

[16] Perot, B., Nallapati, R., (Jan 2003) A moving unstructured staggered mesh method for the simulation of incompressible free-surface flows. Journal of Computational Physics 184 (1), 192-214

[17] Perot, J., (Sep 1993) An analysis of the fractional step method. Journal of Computational Physics 108 (1), 51-58

[18] Perot, J. B., (Oct. 1995) Comments on the Fractional Step Method. Journal of Computational Physics 121, 190-191

[19] Perot, J., Subramanian, V., (May 2007) Discrete calculus methods for diffusion. Journal of Computational Physics 224 (1), 59-81

[20] Perot, J. B., Vidovic, D.,Wesseling, P., (2006) Mimetic reconstruction of vectors. In: Compatible Spatial Discretizations. Springer New York, pp. 173-188

[21] Chang, W., Giraldo, F., Perot, B., (Jul 2002) Analysis of an exact fractional step method. Journal of Computational Physics 180 (1), 183-199

[22] Desbrun, M., Hirani, A. N., Leok, M., Marsden, J. E., (2005) Discrete exterior calculus. arXiv:Math/0508341

[23] Wenneker, I., Segal, A., Wesseling, P., (Jan 2003) Conservation properties of a new unstructured staggered scheme. Computers & Fluids 32 (1), 139-147

[24] Subramanian, V., (2007) Discrete calculus methods and their implementation. PhD. thesis, Mechanical Engineering, University of Massachusetts

[25] Leo J. Grady, J. P., (2010) Discrete Calculus. Springer London

[26] Persson, P.-O., Strang, G., (Jan 2004) A simple mesh generator in MATLAB. SIAM Review 46 (2), 329-345

[27] Vizvari, Z., Sari, Z., Klincsik, M. Odry, P. (2020) Exact schemes for second-order linear differential equations in self-adjoint cases. Adv Differ Equ, 497, https://doi.org/10.1186/s13662-020-02957-7

Introducing the Concept of Internet of Digital Reality – Part I

Péter Baranyi[1], Ádám Csapó[1], Tamás Budai[2], György Wersényi[2]

[1] Dept. of Computer Science, Széchenyi István University
Egyetem tér 1, Győr, Hungary
baranyi.peter@sze.hu, csapo.adam@sze.hu

[2] Dept. of Telecommunications, Széchenyi István University
Egyetem tér 1, Győr, Hungary
budai.tamas@sze.hu, wersenyi@sze.hu

Abstract: With the growing pervasiveness of Virtual Reality, Augmented Reality, Mixed Reality and Digital Twins combined with Artificial Intelligence, 5G networks and the omni-present 2D Web, we are entering into a new era, characterized by a multi-modal entangled combination of previously disparate realms of IT with human and social cognitive systems. This process of entanglement is arguably leading to a new, qualitatively different kind of reality, in which the borders between the physical and digital world, as well as digital representations and simulations thereof are becoming increasingly fuzzy. Based on these developments, the paper re-interprets some well-known concepts – including virtual reality, augmented reality, mixed reality, as well as virtual / digital simulations and virtual / digital twins. Based on these new interpretations, the paper introduces the higher-level structures of Digital Reality (DR) and Internet of Digital Reality (IoD). It is argued that these structures can lead to a better understanding of the new possibilities afforded to humanity by pervasive digital technologies.

Keywords: Digital Reality; Future Internet; Internet of Things; Internet of Everything; Internet of Digital Reality

1 Introduction

With its rapid evolution in the past decades, modern information technology has arguably led to a fundamental re-shaping of the landscape of human cognition, for better or worse [1, 2]. However, regardless of the ways in which individual capabilities, behaviors and socio-cognitive structures are impacted, digitization is a phenomenon that is not only here to stay with us, but can also be expected to deepen in its pervasiveness within all aspects of the everyday human experience. This development has led to a variety of new conceptualizations in terms of the co-evolution of humans, machines, and computer networks – including human-machine entanglement, cognitive merging / co-evolution, singularity, and even post-humanism. All of these concepts highlight the general idea that the border between humans and technology is becoming increasingly fuzzy [3, 4, 5, 6], and all of them are deeply investigated under the scientific discipline of Cognitive Infocommunications (CogInfoCom) [6, 7].

This paper is based on this realization that humanity has reached an inflection point in its social and technological evolution, and details how the continued deepening of human-ICT entanglement can be expected to lead to a qualitatively new kind of reality, referred to as the *digital reality*. In turn, the communication, management and harmonization aspects of different segments of this reality are covered by the concept of *Internet of Digital Reality (IoD)*.

Note that the term 'digital reality' is not entirely new – Deloitte Consulting LLP and the Consumer Technology Association have introduced 'digital reality' as a trademarked term to refer to "*technologies and capabilities that inhere in AR, VR, MR, 360° video, and the immersive experience, enabling simulation of reality in various ways*" as described in [8, 9, 10]. In this paper, our goal is to add more context and a wider perspective to this term – focusing not only on the technology (how information is visualized) but also highlighting artificial intelligence related, capability-oriented and social cognition aspects. We also consider the ramifications of the concept of digital reality in the emergence of a new kind of interconnectedness, referred to as the Internet of Digital Reality (IoD). In a separate paper (Part II), we provide a more in-depth discussion on the infrastructural background of IoD [11].

2 Key Concepts

In order to motivate and clearly delineate Digital Reality as a concept, we first take a look at some related technological and social concepts. Our exploration involves casting new light on some commonly held assumptions about the fields of virtual reality, augmented reality, mixed reality and others. Based on these discussions, a new perspective emerges which helps us define Digital Reality at the end of this section. Later in the paper, the term 'Internet of Digital Reality' is further explored.

2.1 The Concept of Reality

First of all, it is important to clarify what we mean by the term 'reality'. In this paper, we use the working definition of reality as being ***a set of conceptions and perceptions that form an integrated unit of comprehension, and create an understanding of what is possible, desirable and actual***.

As we will see later, this definition covers quite many details of human experience at any given time and in any given historic moment. By 'integrated unit of comprehension', we mean that even if a reality is comprised of many different components, at a higher level (the level of the reality), those components together create a new meaning which does not exist at the individual level of the components. By highlighting what is possible, desirable and actual, we mean that a reality can shape our experience at a basic level.

The definition also highlights the fact that as long as a set of possibilities, desires or actual manifestations of the world are mutually exclusive, or alternatively, as long as a set of possiblities, desires or actual manifestations are focused on entities that cannot have a meaningful influence over one another, then we can speak of different realities. For example, the files on one person's computer and the files on another person's computer belong to different realities – at least in a technical sense – as long as the two people are strangers to each other and live their lives in complete independence of one another. At the same time, the files on one person's computer and the same person's filing cabinet could likely belong to the same reality – the given person's reality – regardless of the former existing in digital form, and the latter in physical form. These examples show that the concept of reality is orthogonal to the dimensions of physical / digital, and real / virtual – an idea that may seem counterintuitive at first, given the colloquial associations created by words such as 'virtual reality', generally used in to cover computer-generated 3D applications.

Expressed differently, the fact that virtual reality, augmented reality and physical reality are treated as different concepts today (based on the medium through which they are primarily accessed) merely shows that the technology behind them has not (yet) achieved a level of pervasiveness based on which we are able to access, manipulate and organize the same information regardless of the medium we are using. If this were the case, virtual and physical reality would significantly overlap.

2.2 Layers of virtuality

In its colloquial sense, the word '**virtual**' is often used to refer to something that is not real, but is a manifestation that somehow resembles a real counterpart. Although widespread associations behind the term 'virtual' most often have to do with digital manifestations that resemble physical objects, we consider this interpretation to be limited, especially from the perspective of the definition of reality provided in Section 2.1. For this reason, ***we consider any manifestation to be 'virtual' that has a referential aspect, regardless of whether that manifestation appears purely in someone's imagination, or in a specific physical or digital solution, and regardless of whether it points to a real (physical) or a purely imaginary concept, or to a specific, concrete object (i.e. an instantiation of a concept)***.

Going deeper into this term, we note that:

- the term 'virtual' is not an absolute term and (contrary to the commonly held viewpoint) is not even necessarily digital – e.g. both a card game on a table or a 3D digital game can be thought of as virtual, depending on what realities the cards and the games themselves refer to. A tangible user interface (a TUI, which is completely physical) can be considered as a virtual tool, provided that it refers to (controls and / or is updated in synchrony with) a 'backend system' (whether digital or physical – though mostly digital in the original formulation of TUIs [12, 13]). Similarly, open-air museums of traditional villages (e.g. Skansen) can be considered as 'virtual' (despite being physical and not even referring to anything digital), as they represent an architecture and way of life from a bygone era (to visit Skansen is to visit a location that resembles a typical village from the past). Based on the above, the term 'virtual' can be conceived of as a relative term that focuses on an aptly defined border between artificial and natural, or some constructed reality and a base reality;

- the term 'virtual' does not necessarily refer to visual (or visual only) manifestations. While most people associate the term 'virtual' (as in 'virtual reality' described below) with visual inputs, resemblance to the referent counterpart will in general depend on the context - whether that counterpart has a visual appearance (either in 2D or 3D), or if it has affordances in other modalities, or if it is purely an abstract concept. Note that in some interpretations, such as in the framework of embodied cognition, nothing exists that is purely abstract (i.e., divorced from the spatial experience / spatial interactions of the human mind-body) [14], we can at least assume that it is possible to distinguish among concepts that are closer to or more removed from physical reality [15, 16].

In summary, it is possible to characterize both physical and digital objects that are perceived by humans as being virtual to a different extent, depending on their proximity to the reality they refer to. A digital representation of a physical object that in its specific form does not exist in reality can seem more virtual (less real) than a digital representation of an object (an instantiation of the concept) that specifically really exists in physical reality. By contrast, physical objects (like paper money) can often seem even less virtual, though in a sense they are absolutely virtual in that they are an abstraction that correlates with someone's services to society as recognized by that society.

2.3 Characterizing Realities by Degree of Physicality: Virtual, Augmented and Mixed Realities

According to the commonly held view, **virtual reality (VR)** refers to a higher-level arrangement of computer-generated and visually displayed objects that resemble objects in physical reality.

However, as described above, the concepts 'reality' and 'virtual' are orthogonal to the medium or specific technology that is being used. Therefore, in much the same way as the term 'virtual', ***'virtual reality' can be conceived of as any artificially constructed environment (whether physical or digital) that contains virtual objects***. It is important to point out that VR can be similar to or completely different from the real physical world, and can be experienced using both physical or digital tools. In other words, the sign / symbol and the referent (i.e., the referred-to) 'components' of a virtual reality can be either physical or digital, in any combination.

Today, VR is typically implemented on digital devices (usually 3D displays) and most recently it has become associated with 3D glasses, which provide an immersive experience into a complete virtual environment generated by computers. In this regard, VR is generally considered as a set of 3D visualizations describing any artificial or real world (although other perceptual modalities like auditory, haptic, somatosensory and olfactory cues can also be relevant. However, the fact alone that an environment is 2D (and not 3D), not primarily visual (as opposed to visual), augmentative (in the sense that it does not 'shut out' the physical environment as opposed to being completely immersive), or physical (as opposed to digital) – indeed, any combination of the above – ***does not mean that the environment cannot be artificially constructed and that it cannot refer to other concepts or objects. Further, it has no bearing on what kind of 'mixture' it represents between the physical and the digital*** – one can be standing in an insulated room wearing a completely immersive headset, and yet one's thoughts and feelings can refer to concepts and objects in any substrate, both physical and digital – hence refer to the same 'reality'. It can even be argued that VR is a novel kind of infocommunication tool, with direct consequences to everyday life [17]. As we will see, this is an important feature of all computer-generated realities, especially when considering their capability of referring to realities at different levels of virtuality, as described later in Section 2.4.

Augmented reality (AR) most often refers to a computer-generated reality that involves virtual objects presented on a visual field inside a physical environment. Note, that as in the case of VR, an augmented reality need not necessarily be generated by a computer. However, it is clear that the key distinguishing feature of AR as opposed to VR is that the user is primarily located in a 'physical base reality', which is somehow augmented.

In general, the physical objects in the user's environment can serve as fixtures that can accommodate various data representations – e.g. a large dashboard or video display might be projected onto an empty vertical surface like a wall, whereas a more structured and single-purpose display like a clock might be projected onto a smaller, flat surface like the top of a night stand. Since the human body is also a part of physical reality, we consider any augmentation of the human body (e.g. with wearable displays or bionic devices integrated into the body) as a part of augmented reality.

Because of its supplementary nature to physical reality, AR is often also referred as **extended reality (XR)**. Also somewhat related to AR is the concept of **mixed reality (MR)**, which is defined as an environment in which "real world and virtual world objects are presented together within a single display", and that falls "anywhere between the extrema of the Virtuality Continuum" [18] – a hypothetical continuum between the two extremes of the purely physical and the purely virtual, encompassing both augmented reality and augmented virtuality. In the terminology of the virtuality continuum, VR is regarded as a singular 'limiting case' in which everything is purely virtual. All other realities are mixed by necessity. However, if the notions of reality and virtuality are considered in the sense described in Sections 2.1 and 2.2, the notion of virtuality hinges not on a specific substrate (digital or physical) and even a single reality can reach across the border between digital and physical. This is well demonstrated by the fact that the terms virtual reality and augmented reality are often used side-by-side or interchangeably [19][1].

As described in the following subsection, computer-generated realities can be characterized not only based on the levels of virtuality in which and with which they operate, but also based on the qualities of the 'referent realities' that are being referred to by the objects and processes within those computer-generated realities.

2.4 Linking Together Realities via Representation and Simulation

Computer-generated realities derive their strong potential not simply from their ability to present objects that are reminiscent of physical objects in some abstract or entertainment-driven sense, nor from their ability to overlay information about imaginary entities on physical objects. By the same token, their power isn't necessarily derived from the way in which they are merged with our 'physical base reality' – whether they appear on portable, see-through displays or on the screen of a desktop computer.

Rather, computer-generated realities are immensely powerful due to their ability to "speak" to humans not just about concepts, but also about objects that exist in different realities, using a varied language that includes both familiar objects and abstract representations. In this subsection, the referential capabilities of computer-generated realities (i.e. their 'aboutness' with respect to other realities) are described via the concepts of *simulation* and *twin representation*.

A **virtual simulation (VS)** is a simulation of the 3D motion and changes in the geometrical 3D representation of an object (whether physical or abstract) in a computer-generated or in a physical environment. Note that although in most cases, virtual simulations involve depicting the appearance and motions of a physically existing object in a computer-generated world, the definition of VS nevertheless allows other combinations between the physical and non-physical to refer to each other, as long as both the sign / symbol and the referent have temporal geometric properties. This is the reason why a purely abstract process, like the training process of a deep neural network, can also be the subject of a virtual simulation, if the goal is to show some (simplified, aggregated) aspects of the training process to e.g. a group of students or experts using graphical (geometric) concepts. In this sense, any graphical model that has a temporal dimension can be considered as a virtual simulation, when the temporal dimension is being explored. In any case, a VS by definition focuses on

[1] For a concise historical overview of both VR and AR applications, see [20].

the visualization of geometrical properties and / or motion in 3D.

In some sense more generally, and in another more specifically, a **digital simulation (DS)** is a computer-generated simulation of a temporal process (whether a real, physical, purely imagined or digital process). In the case of DS, the 3D virtual (geometric) representation of various objects is not in primary focus (or, in the extreme case, has no meaning at all), since the purpose is to represent processes, which by necessity appear as changes through time, using digital technologies. The concept of DS therefore includes components that may abstract away from the 'high-fidelity visual replica' viewpoint of VR and can be considered as a digitized version of a given physical process, either actually taking place or imagined. In some cases, even the potential future states of the system can be estimated (or simulated) and presented to users, thereby providing a 'multi-state' representation. Returning to the previous example of simulating deep neural networks, if the goal is not to present the workings of the model to humans, but rather to run it in order to obtain viable estimates for certain hyper-parameters (which are then later used as a part of the 'real' training process), then this hyper-parameter search can be considered as a digital simulation.

Note that both VS and DS can be used to simulate purely mental or digital phenomena (i.e. objects and processes can refer to cognitive objects and cognitive representations of processes, or to processes that are already being run on computers). Due to its process-based quality, in an extreme case, DS does not necessitate the use of any specific 3D representation from the physical world – whereas in the case of VS, the 3D representations are by necessity object-based as they involve the transmission of geometrical properties.

Moving further on in our exploration of bridges between physical and mental / computer-generated realities, it is useful to consider the concepts of virtual twins and digital twins.

A **virtual twin (VT)** is an object that refers to another (referent) object in a different reality, and whose geometric appearance is continuously updated to reflect the geometric appearance of that referent object. VT primarily refers to a high fidelity geometric copy with a photo-realistic visualisation of various physical features such as color, materials / textures etc. in VR. The 3D motion of a VT can also be synchronized in real time with that of its counterpart, in cases where the VT is used to control the real object, or the real object is used to control the VT (control / monitoring). In practice, the motion of the 3D representation of the VT is based on a combination of information from simulations (VS) and the real object. In general, we can say that if there are two objects, one being a real physical object, and the other one represented in VR, and the two are purposely similar in visual appearance and in 3D motion, then the application is an instance of a VT.

A **digital twin (DT)** is an extension of VT with DS, such that not only the 3D visualisation and 3D motion of an object is in focus, but all process-based functionalities and interactions of and with the object(s) is represented, even when no direct 3D

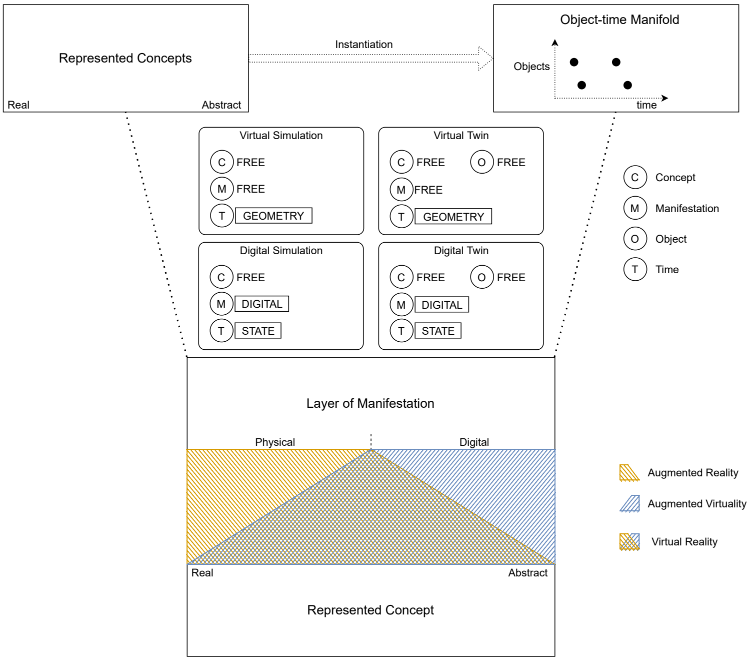


Figure 1
Key concepts introduced in Section 2. As the bottom of the figure shows, augmented reality augments a physical manifestation with concepts – possibly both real or abstract. Augmented virtuality on the other hand augments a digital manifestation with real or abstract concepts. Virtual reality is taken to integrate both of these possibilities, since the manifestation of VR can be both digital or physical (as in the case of an open-air historic replica of a bygone era). As the top of the figure shows, simulations and twins can be represented as a combination of concepts, manifestations, temporal representations and (in the case of twins) objects. The categories of virtual / digital and simulation / twin constrain some of these dimensions – i.e. virtual simulations / twins focus on geometry, while digital simulations / twins focus more on abstract state, and also, digital simulations / twins are constrained to the digital world, whereas virtual simulations / twins can be both physical and virtual.

geometrical representation for the given phenomenon exists. DT extends the concept of VT with user interactions and various further functionalities that are based on DS or obtained via a connection to the referent object. In general, if there are two objects, one being a real physical object and the other represented in VR, and the two are not only visually similar but can also be interacted with in similar ways and in synchrony, then the application is an instance of DT.

2.5 Summary of Key Concepts

To summarize, virtual, augmented and other mixed realities are arrangements of virtual objects, i.e. objects that in some way resemble a counterpart from a different

	Environment	Symbols	Referents		
	Physical / Digital	Physical / Digital	Geometric / Abstract	Static / Temporal	Real / Imagined
Virtual reality (VR)	P/D	P/D	G/A	S/T	R/I
Augmented reality (AR)	P	P/D	G/A	S/T	R/I
	include as components:				
Virtual simulation (VS)	P/D	P/D	G	T	R/I
Digital simulation (DS)	D	D	A	T	R/I
Virtual twin (VT)	P/D	P/D	G	T	R
Digital twin (DT)	D	D	A	T	R

Table 1
Summary of concepts defined in Section 2. As indicated in the subheaders of the columns, the letters P, D, G, A, T, R, and I stand for Physical, Digital, Geometric, Abstract, Static, Temporal, Real and Imagined, respectively.

reality - whether physical, digital, real or imagined. The purpose of such realities can be, but is not necessarily to convey any immediate message about these counterparts; instead, their purpose can simply be to provide an aesthetically pleasing experience, or to serve as a medium (an environment) of communication through which other, related concepts can be conveyed more easily [17]. The main distinction between virtual and augmented reality is that an augmented reality is by necessity primarily physically based, with some (physical or digital) extensions. Even the physical or digital augmentation of the human body (viewed as a physical substrate) can be considered as an augmented reality, if it has the qualities of a reality as defined in Section 2.1.

Virtual and digital simulations, as well as virtual and digital twins can be seen as components that may be integrated into a virtual or augmented reality. At a conceptual level, virtual simulations and digital simulations can be defined as temporal representations of changes in the physical appearance, or in the state of some referent objects. Note that these referent objects need not actually exist in another reality (i.e. can be purely imagined), and the main focus is on temporal changes in the states of referent objects that may possibly exist.

Finally, we have defined virtual twins and digital twins as representations whose state evolves with the states of some referent objects which actually exist in a different reality. The purpose of such representations is to support the control and monitoring of systems that exist in a base reality that has a unique importance.

A summary of these concepts is provided in Table 1 and in Figure 1.

3 Definition of the Digital Reality

3.1 Deepening Pervasiveness of 2D Digital Environments

The term **2D digital environment (2DE)** refers to all software applications available to users on their infocommunication devices, i.e. any digital feature that provides 2-dimensional visual and interaction surfaces, including all web applications, 2D collaborative solutions and cloud based solutions that we use for work, entertainment and digital social life can be considered as 2D digital environments.

2DEs have become a central tool in the lives of most people living in the developed world for at least the past decade. However, the pervasiveness and influence of these environments is now deepening, as a result of a confluence of factors enabling automation at several different conceptual levels:

- Automation at the network level is leading to increased throughput and geographically more widespread availability, thereby enabling users to be presented with richer content in increasingly remote corners of the world.
- Automation dealing with big data solutions is allowing service providers to gather more information about and better understand the habits and preferences of their customers.
- Finally, artificial intelligence (AI) is making it possible to filter through this deluge of information and enables a more efficient curation of digital content in a context-sensitive way.

The net effect of these trends is not simply that humans are able to obtain the same information faster, nor simply that humans are able to obtain more information at any given time; but rather to release a catalytic process that is transforming our everyday reality – which is to say, *our conception and perception of what is possible, desirable and achievable based on an integrated set of physical, digital and mental concepts*.

Today, the World Wide Web is not just an Internet-based set of technologies that we use to access information and to send text, voice and / or video messages. Rather, we can increasingly communicate with machines directly and they are also capable of doing so with each other. Algorithms are ‘watching over us’, the big tech companies can easily learn things about us that we ourselves do not know. Further, they are able to obtain information that we willingly (or unknowingly) provide via a growing range of modalities. As non-human entities are entering the public Internet, and are evolving to a level where they can communicate, and eventually even decide / act on their own based on sophisticated adaptive methods, the associated cognitive load placed on humans is expected to be enormous [21].

3.2 The Emergence of Cognitive Entities and Digital Reality

When a human is growing up in such a complex digital environment his or her brain, mentality and complete psychological system is highly influenced (in many aspects irreversibly). In addition, the vast information that is collected and processed by AI algorithms is also qualitatively different compared to the past decades. In short, both humans' mental worlds and the digital world are influenced in a tight iteration loop. As a result, a markedly hybrid system is emerging in which the human and digital components are inseparable. This can be conceived of as the emergence of the **Cognitive Entity**, in which artificial and real cognitive systems are co-evolved to form qualitatively new capabilities [7, 22, 23, 24].

This view of humans will be even more relevant in the case of the coming generations. And it is based on this view, influenced by the concepts of VR, AR, VS, DS, VT, DT, AI and others in the field of cognitive infocommunications (CogInfoCom) [7, 6] that the concept of Digital Reality can be defined as follows:

A **Digital Reality (DR)** ***is a high-level integration of virtual reality (including augmented reality, virtual and digital simulations and twins), artificial intelligence and 2D digital environments which creates a highly contextual reality for humans in which previously disparate realms of human experience are brought together. DR encompasses not only industrial applications but also helps increase productivity in all corners of life (both physical and digital), thereby enabling the development of new social entities and structures, such as 3D digital universities, 3D businesses, 3D governance, 3D web-based digital entertainment, 3D collaborative sites and marketplaces.***

A key question that may arise is how a discussion on Digital Reality differs from an analysis on "*all digital solutions*" or "*everything that is digital*". In our view, the term DR is more specific in that it refers to an integration of digital solutions that points to and as a result, helps create a new reality that involves human immersion (via co-evolution as discussed under Cognitive Infocommunications) and can be highly contextual.

As an example, consider an 'integration' of a physical refrigerator, microwave oven, smart stove and deep fryer along with their corresponding digital twins, together with various further virtual objects in a shareable VR space, which also includes a 2D digital environment of collaborative cookbooks, cooking notes, cooking blogs, and video chats with a connection to the most renowned experts in the culinary arts. In such an integration, with physical devices, functional replicas (for e.g. training and visualization), documentations and even access to the world's best cooks, the end result is a mixed reality that embodies a qualitatively more advanced reality than do its individual components - a digital reality for cooking. A DR like this might also include AI algorithms for making suggestions as the user is doing the cooking, or for aggregating the calories and nutritional value of the user's consumption over time, or for replenishing the supply of raw ingredients in a predictive manner.

As demonstrated by this example, a digital reality should not be taken to mean everything that is digital; much rather, it defines a level of compactness, or concentratedness of components (whether physical, digital, static, dynamic, pre-programmed or AI-driven) that enables higher-level goals and functionalities to be formulated and implemented. This ability of DR to stretch even our mental concepts (of what is possible) and of extending our human capabilities in radically new ways is what sets it apart from mere collections of digital tools. Of course, DRs involve digital tools (among others), but those tools are organized and integrated in a specific, topic-oriented way. In short, DRs are capable of forming new realities analogous to what we mean by the term ‘reality’ as a web of concepts and experiences in terms of aspirations and actualization in real life.

4 Internet of Digital Reality

Based on the above, the Internet of Digital Reality (IoD) is a set of technologies that enables digital realities to be managed, transmitted and harmonized in networked environments (both public and private), focusing on a higher level of user accessibility, immersiveness and experience with the help of virtual reality and artificial intelligence. Connections among various cognitive entities also have to be handled not only at the end user level of virtual reality displays and software, but also at the levels of network protocols and network management, physical media (wired or wireless), hardware interfaces, and other equipment. AI is a key component of both digital reality and IoD, that enables a cohesion of context-driven content and intelligent network routing to emerge. In this section, we provide a brief overview of IoD. Further discussions of IoD, as well as its ramifications and the technologies behind it can be found in [11].

4.1 Historical perspectives behind IoD

The Internet of Things (IoT) introduced the world of networked “things” – e.g. sensors and actuators, wearables, digital twins – by integrating distributed computation with intelligent connections. Digital representations of physical entities in the real world can thus be connected, interacted with, managed, and they are able to communicate with each other even without constant human supervision.

The Internet of Everything (IoE) was defined by Cisco in 2013 as “*bringing together people, process, data, and things to make networked connections more relevant and valuable than ever before, turning information into actions that create new capabilities, richer experiences, and unprecedented economic opportunity for businesses, individuals, and countries*” [25, 26]. A key difference between IoE and IoT is the intelligence of connections. IoT is mostly about physical objects and concepts communicating with each other but IoE is what brings in network intelligence to bind all these concepts into a cohesive system.

The proliferation of fields called ‘*Internet-of-X*’ (X being something different in each case, as in the case of e.g. Internet of Nano Things, Internet of Mission-Critical Things, Internet of Mobile Things [27]) somewhat resembles the proliferation of more traditional fields focusing on different kinds of interactions – e.g. human-computer interactions, human-machine interactions, human-robot interactions etc. The emergence of such fields with similar concepts and methodologies (albeit applied to different contexts) was part of the impetus behind the definition of cognitive infocommunications (CogInfoCom) [6, 7, 24]. Similarly, we propose the term ‘Internet of Digital Reality (IoD)’ to partially integrate, partially complement and more importantly augment earlier notions of the form ‘Internet-of-X’. After all, it can be argued that the real motivation behind all of these technological directions was always to merge, augment and share realities – an idea that is already present (excluding the network aspects) in Digital Reality.

Perhaps the most relevant aspect of IoD is that it connects digital realities, that is, combinations of technologies and data that create a higher-level functional integration. For example, IoD is the network through which combinations of 3D virtual spaces and their real-world counterparts can be shared, together with all relevant data and interactive support for extended capabilities.

4.2 Pillars of IoD

In one sense, IoD supersedes IoT and IoE where not only physical “things”, but also complete digital reality are connected via a (public or private) network.

Today, cognitive entities and virtual / digital twins are strongly based on Internet connections [28]. Pillars of the Internet of Digital Reality include:

- Cognitive entities (qualitatively new capabilities emerging via a combination of machines, "things", sensors, AI, digital twins, avatars, algorithms, bots, RPA, higher level organizations etc.) interacting in the digital realm
- Information (data, web content, control)
- Communication networks (intelligent connection, wired and RF physical layers, public and private networks)
- Artificial intelligence, which gains new significance as a global network capable of distributed learning and continued evolution as digital realities become connected and shared across the globe
- Access devices and interfaces (headsets, tactile devices, AR/MR/VR spaces, 360-degree immersive scenarios, mobility and navigation)
- Cognitive infocommunications (sensation and perception, human factors and human-ICT co-evolution)

- Safety and security (data, information, or even the physical safety of the user etc.)
- Digital business and legal issues
- Digital Society (education, acceptance of technology, digital work, e-government, digital arts and gaming)

All of the above pillars are a part of the science and practice of IoD; however, these pillars are also broadened in scope based on the radically new perspectives of IoD. For example, artificial intelligence is transformed into a global network (networked AI) that is capable of learning based on user activities and user data on the fly. Of course, the breaking down of barriers between various contexts and use cases also creates new challenges in terms of data security and legal issues. These are just two examples of how the pillars of IoD both help form and are transformed by IoD. A detailed discussion of IoD, as well as its ramifications and the technologies behind it can be found in Part II of this paper [11].

Conclusions

In this paper, a short introduction was given on the concept of Digital Reality, which allows users to access, manipulate and interact with integrated sets of content – whether on display screens, immersive settings or overlaid on physical objects – that represent concepts and / or objects that can be both physical or digital, as well as both real or imagined. The paper showed that this new form of reality born in today's high level entangled combination of digital technologies and humans needs a new definition for future scientific works. Further, the concept of Internet of Digital Reality (IoD) was introduced, as a set of network and related technologies for the management, transition and harmonization of Digital Realities. Further details on IoD can be found in [11].

Acknowledgement

This research was supported by the Digital Development Center in the national framework GINOP-3.1.1-VEKOP-15- 2016-00001 "Promotion and support of cooperations between educational institutions and ICT enterprises".

References

[1] K. K. Loh and R. Kanai, "How has the Internet Reshaped Human Cognition?" *The Neuroscientist*, vol. 22, no. 5, pp. 506–520, 2016.

[2] M. Prensky, "Homo Sapiens Digital: From Digital Immigrants and Digital Natives to Digital Wisdom," *Innovate: Journal of Online Education*, vol. 5, no. 3, 2009.

[3] C. Frauenberger, "Entanglement HCI the Next Wave?" *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 27, no. 1, pp. 1–27, 2019.

[4] R. Kurzweil, *The singularity is near: When humans transcend biology*. Penguin, 2005.

[5] A. Pilsch, *Transhumanism: Evolutionary Futurism and the Human Technologies of Utopia*. U of Minnesota Press, 2017.

[6] P. Baranyi and Á. Csapó, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 67–83, 2012.

[7] P. Baranyi, A. Csapo, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*. Springer, 2015.

[8] P. D. Ramani Moses, Nikita Garia, "Digital Reality – A technical primer," https://www2.deloitte.com/insights/us/en/topics/emerging-technologies/digital-reality-technical-primer.html, 2021.

[9] J. Schwartz, S. Hatfield, R. Jones, and S. Anderson, "What is the Future of Work," *Redefining work, workforces, and workplaces. Part Of A Deloitte Series On The Future Of Work*, 2019.

[10] Y. Kang and K. C. Yang, "Employing Digital Reality Technologies in Art Exhibitions and Museums: A Global Survey of Best Practices and Implications," in *Virtual and Augmented Reality in Education, Art, and Museums*. IGI Global, 2020, pp. 139–161.

[11] G. Wersényi, Á. Csapó, T. Budai, and P. Baranyi, "Internet of Digital Reality: Infrastructural Background – Part II," *Acta Polytechnica Hungarica*, vol. 18, no. 8, pp. 91–104, 2021.

[12] H. Ishii, "The tangible user interface and its evolution," *Communications of the ACM*, vol. 51, no. 6, pp. 32–36, 2008.

[13] Y. Zhao, Y. Qin, Y. Liu, S. Liu, and Y. Shi, "Qook: A new physical-virtual coupling experience for active reading," in *Proceedings of the adjunct publication of the 26th annual ACM symposium on User interface software and technology*, 2013, pp. 5–6.

[14] A. M. Glenberg, "Few believe the world is flat: How embodiment is changing the scientific understanding of cognition." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 69, no. 2, p. 165, 2015.

[15] L. W. Barsalou *et al.*, "Perceptual symbol systems," *Behavioral and brain sciences*, vol. 22, no. 4, pp. 577–660, 1999.

[16] B. Z. Mahon and A. Caramazza, "A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content," *Journal of physiology-Paris*, vol. 102, no. 1-3, pp. 59–70, 2008.

[17] Á. B. Csapó, I. Horvath, P. Galambos, and P. Baranyi, "VR as a medium of communication: from memory palaces to comprehensive memory management," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2018, pp. 000 389–000 394.

[18] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, "Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum," in *Telemanipulator and telepresence technologies*, vol. 2351. International Society for Optics and Photonics, 1995, pp. 282–292.

[19] R. J. Grossmann, "Virtual reality, Augmented Reality – I call it i-Reality," *Mhealth*, vol. 1, 2015.

[20] D. Williams, "The Internet of Everything - Cisco IoE Value Index Study," https://www.huffpost.com/entry/the-history-of-augmented-_b_9955048, 2017.

[21] F. Pasquale, "The alien intelligence of automated media," in *New Laws of Robotics*. Harvard University Press, 2020, pp. 89–118.

[22] P. Baranyi and A. B. Csapo, "Revisiting the concept of generation CE-Generation of Cognitive Entities," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2015, pp. 583–586.

[23] L. I. Komlósi and P. Waldbuesser, "The cognitive entity generation: Emergent properties in social cognition," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2015, pp. 439–442.

[24] J. Katona, "A Review of Human–Computer Interaction and Virtual Reality Research Fields in Cognitive InfoCommunications," *Applied Sciences*, vol. 11, no. 6, p. 2646, 2021.

[25] Cisco, "The Internet of Everything - Global Private Sector Economic Analysis," https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoE_Economy_FAQ.pdf, 2013.

[26] ——, "The Internet of Everything - Cisco IoE Value Index Study," https://www.cisco.com/c/dam/en_us/about/business-insights/docs/ioe-value-index-faq.pdf, 2013.

[27] C. Srinivasan, B. Rajesh, P. Saikalyan, K. Premsagar, and E. S. Yadav, "A Review on the Different Types of Internet of Things (IoT)," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, no. 1, pp. 154–158, 2019.

[28] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the Digital Twin: A systematic literature review," *CIRP Journal of Manufacturing Science and Technology*, vol. 29, pp. 36–52, 2020.