A Simple and Effective Heuristic Control System for the Heliostat Field of Solar Power Tower Plants

Nicolás C. Cruz[1], José Domingo Álvarez[1], Juana L. Redondo[1], Manuel Berenguel[1], Ryszard Klempous[2], Pilar M. Ortigosa[1]

[1]University of Almería, ceiA3 Excellence Agrifood Campus, Department of Informatics, Sacramento Road, n/a, 04120, La Cañada de San Urbano, Almería, Spain, ncalvocruz@ual.es, jhervas@ual.es, jlredondo@ual.es, beren@ual.es, ortigosa@ual.es

[2]Wrocław University of Science and Technology, Department of Electronics, wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Polonia, ryszard.klempous@pwr.edu.pl

Abstract: Solar power tower plants use large arrays of mirrors, known as heliostats, to concentrate solar radiation on their receiver and heat the working fluid inside them. However, receivers must not be under thermal stress. Otherwise, their life expectancy is reduced, which affects the cost and viability of production plant. Controlling the flux distributions on receivers requires selecting the active heliostats and their target points. It is a challenging task that should not be under the responsibility and expertise of human operators only. This work defines a closed-loop controller to keep the setpoint or desired flux map under certain conditions. It combines real measurements and an ad-hoc analytical model of the target field with a set of heuristic rules that covers how to activate, deactivate, and re-aim heliostats. The proposed system has been applied to a model of the CESA-I field at the Solar Platform of Almería. The open-source ray-tracer Tonatiuh represents the reality. The initial operation point has been determined with a theoretical flux distribution optimizer. According to the experimentation, the controller improves the initial and model-based flux distribution by raising its power from 708.4 to 736.4 kW (with a setpoint of 739.6 kW).

Keywords: solar power tower plant; heliostat field; flux distribution; automatic control; closed-loop control; heuristics

1 Introduction

The pollution and depletion problems associated with electricity generation through traditional fuel-based technologies have increased the interest in renewable energy [1-3]. Concentrated Solar Power systems (CSP) are especially attractive because of their hybridization capabilities, as well as, their production stability through thermal storage [4]. Among the different CSP technologies [1, 5], Solar Power Tower plants (SPT) are probably the most promising ones because the high temperatures reached result in high thermodynamic efficiency [2, 4] and less thermal storage requirements [2]. SPT plants have good development prospects linked to improving their commercial competitiveness [2, 6] and also attracts the interest of this work.

Conceptually, a SPT plant consists of a solar radiation receiver linked to a power block and a set of solar tracking mirrors known as heliostats. The heliostats follow the apparent movement of the Sun to concentrate the incident solar radiation on the receiver. The receiver, which is generally (but not necessarily [7]) on top of a tower for better focusing, contains a circuit for a Heat-Transfer Fluid (HTF). The goal is to heat the HTF with the power that the field concentrates on the receiver. Once the temperature of the HTF is appropriate, it serves to generate electricity in a power cycle (either combined, gas or steam turbine cycle). The HTF can also be stored for delivery under demand. For instance, the Gemasolar plant can generate electricity approximately for approximately 15 hours without solar radiation [2]. Figure 1 shows a simplified representation of a SPT plant with a steam turbine cycle. There exist different variations over the basis described starting from the receiver design (e.g., flat, cylindrical) and the HTF (e.g., molten salt, air, water) [4]. The interested reader can find more information about SPT plants in [8, 9].

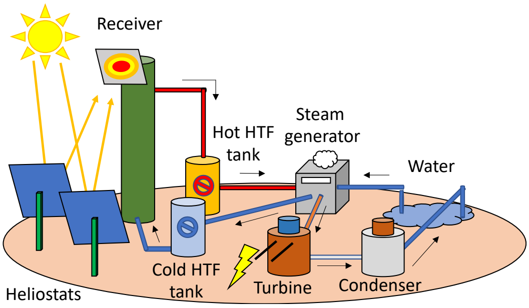


Figure 1
Simple depiction of a solar power tower plant with a steam turbine cycle

Despite its conceptual simplicity, the radiation receiver of a SPT plant is a sophisticated, expensive and fragile component, that needs special care to reduce the production costs [10] and maximize throughput [11]. The solar flux distribution that the heliostat field reflects on its receiver may cause thermal stress, premature aging, and deformation [10-12]. It is also directly related to the production efficiency of the plant as the interception factor, i.e., the ratio of non-profited radiation [10, 13, 14]. Therefore, it is necessary to control the flux distribution reflected by the field to keep the receiver safe while concentrating as much power as possible [15, 16]. For this reason, the development of optimal aiming strategies attracts the attention of many researchers [16].

The flux distribution achieved on the receiver depends on the active heliostats and their aim points [17-19]. Besides, it is affected by the apparent solar movement, the evolution of direct normal irradiance (DNI) throughout the day (including its sudden variations caused by clouds [15]), heliostat errors, the wind and other atmospheric phenomena [14, 15, 17]. Consequently, there are far too many variables to delegate the aiming control tasks to human operators [17], especially considering the development of current multi-aiming strategies [18, 20]. Many proposals have been made to control heliostat fields with different objectives and scopes [21].

This work presents a feedback control system connected to a flux optimization method. The optimizer theoretically configures the field to obtain any desired flux distribution. The controller minimizes the error between the flux distribution achieved after translating the theoretical result to reality and the desired one. The control logic is based on heuristic rules. It tries to reduce the effect of disturbances, modelling and optimization errors. The results presented in this paper show how the feedback controller improves the initial flux distribution computed with model-based optimization.

The paper is structured as follows: Section 2 contains a literature review of recent aiming methods. Section 3 summarizes the technical background of the control logic described in this work. Section 4 describes the heuristic closed-loop controller designed. Section 5 explains the experimentation carried out. Finally, the last section contains the conclusions and future work lines.

2 Literature Review

Salomé et al. [10] implement a TABU local search algorithm [22] to aim heliostats by following a combinatorial formulation of heliostats with a finite set of aim points. They design an open-loop controller that tries to obtain homogeneous flux distributions with acceptable spillage. Distant heliostats are forced to aim at central zones to reduce spillage. Besarati et al. [13] design a genetic algorithm [23] with a similar combinatorial approach to achieve flat flux forms by minimizing their standard deviation. The authors also add a dedicated component to reduce extreme spillage situations. Grobler [14] combines the two previous strategies by using the TABU search for generating initial solutions for the genetic optimizer. The overall goal is the same, but the descriptiveness of the objective function of the optimization problem is improved. Yu et al. [24] use a TABU search to minimize flux peaks and spillage by following a combinatorial formulation. Heliostats are grouped to reduce the search space, but the shape of the receiver is considered in depth to distribute the aim points.

Belhomme et al. [25] apply ant colony optimization [26] to assign the best aim point to each heliostat in a combinatorial context. The method avoids dangerous radiation peaks, but its focus is on maximizing the performance of the receiver. Maldonado et al. [27] compare the previous method to a new proposal of local scope that studies small variations to solutions. The ant colony optimizer is more robust in general, while the local one, can achieve high-quality solutions. The authors finally suggest a hybrid method that uses the ant colony to get promising initial solutions for the local method.

Ashley et al. [15] maximize the power on the receiver while keeping it in a valid range and looking for uniformity. The approach is combinatorial, and their integer programming method can find valid solutions in almost real-time, which would allow handling the effect of clouds. This aspect is covered in [28].

Astolfi et al. [11] focus on avoiding flux peaks, and the field is divided into circular sectors to adjust the aim point height of each zone. The continuous optimization problem faced focuses on the vertical aim point and grouping results. The authors test several variations considering overlapping and not doing so, which is less effective. Sánchez-González et al. [16] also consider dividing the field into sectors to achieve flat flux distributions by adjusting aim point heights. The approach relies on exploiting the symmetry of the desired flux distribution and balancing the up and bottom zones. The formulation is based on the concept of aiming factor, which allows estimating the size of the beam reflected by each heliostat. That concept was previously introduced by the authors in [20], where they design a method of two stages to maximize the thermal power output of the receiver while preserving its integrity.

Kribus [29] focuses on avoiding tracking and aiming errors, which is not addressed with open-loop approaches [10] and reduces the dependency on models. The system uses CCD cameras to detect the heliostats not correctly aimed and to calculate the correction signal. Convery [30] also opts for a closed-loop controller, but the design is especially innovative and cheap. It uses piezoelectric oscillators and photodiodes instead of cameras to identify misaiming heliostats. The oscillators serve to make each heliostat vibrate at a unique frequency detected by the photodiodes. Freeman et al. [31] try to improve the standalone capabilities of their closed-loop system. They aim to reduce the necessity of feedback from the receiver of the two previous methods. To this end, they add accelerometers and gyroscopes to the context proposed in [30]. Obtaining specific flux profiles is not covered.

Gallego et al. [32] try to obtain flat flux distributions while maximizing the incident power on the receiver. Instead of defining a combinatorial problem, as usual, the authors opt for a continuous one with two variables per heliostat. The problem is divided into simpler instances by working with different groups of heliostats, called agents, to overcome the computational expenses. The method iteratively considers the effect of groups on each other.

Finally, in [17], the authors of this paper design a general method that works in a continuous search space like the previous one, but it aims to replicate any desired flux distribution on the receiver. Therefore, it allows avoiding peaks or achieving any other feature by adjusting the flux map reference. It has two layers. The first layer serves to select the active heliostats through a genetic algorithm. The second layer adjusts their aim points by using a gradient descent method. This paper combines that basis with the modelling approach described in [33] to create an operational framework. In this context, a closed-loop controller based on heuristic rules is designed and tested. It aims to make it possible to translate and polish the instantaneous result of the method in [17]. The open-source ray-tracer Tonatiuh [34] is taken as the reality.

3 Overview of the Operational Framework

Since the designed method extends the technical context proposed in [17, 33], this section summarizes both for the sake of completeness. Section 3.1 explains how the target field is modelled, and Section 3.2 describes the optimizer that computes off-line field configurations. The interested reader is referred to the complete works for further information.

3.1 On Predicting the Behavior of the Target Field

Aim point optimization requires predicting the flux profile over the receiver surface, i.e., optical modelling [3, 10, 14]. It serves to guide the search by estimating the performance of different solutions (linked to off-line optimization [17]) and to predict the effect of different actions (related to on-line control tasks).

As summarized in [13], it is possible to compute flux maps either numerically or analytically. The numerical approach consists in simulating multiple rays through several optical stages to study their interaction with the environment. It is known as ray-tracing and can be done with software packages such as STRAL, SolTrace and Tonatiuh [3, 13, 21]. The analytical methods model the flux maps with mathematical functions such as circular Gaussian distributions as HFLCAL [13, 21].

Ray-tracing offers higher accuracy and flexibility than analytical methods at the expense of higher computational requirements. Considering potential time constraints and the fact that analytical errors attenuate according to the central limit theorem [14], the analytical approach is generally preferred [11, 13, 16]. The work in [25] is an exception example because the authors use ray-tracing and reduce its time impact by storing partial results. Regardless, creating databases with pre-computed information is also an option with analytical strategies [10, 11, 14, 24].

This work relies on an analytical model of the target field. It has been developed by following the methodology proposed by the authors of this paper in [33]. The method defines how to build an ad-hoc model based on accurate data either from ray-tracing or reality. It consists of the following steps:

1. It is necessary to register the position of a subset of heliostats covering all the zones of the target field. Next, the paths of the sun at the location of the field are sampled in a similar way. Finally, the flux map of each heliostat is gathered for each solar position. The maps can come from real measurement or ray-tracing. Notice that since the model under construction will be based on this information, its acquisition conditions will bound the prediction capabilities. For this reason, as long as the flux maps exhibit atmospheric attenuation and shading and blocking losses, the resulting model will implicitly try to consider them.

2. An analytical expression that can describe the flux map that any heliostat projects on its receiver must be selected. As said, Gaussian functions are popular for this purpose. The one chosen is explained later. After that, each of the flux maps previously gathered must be fitted to the expression. The parameters that define the overall shape must be recorded and linked to the particular heliostat and solar position.

3. The set of records is randomly split into a modelling set and a validation one. After that, the goal is to build a model that can predict the aforementioned shape parameters depending on the coordinates of the heliostats and the solar position. Any modelling technique, such as Neural Networks [35] or Random Forests [36], which is the choice in [33], can be used. The validation subset is ignored at this step.

4. It is necessary to confirm the effectiveness of the model or parameter predictor built at the previous step. To this end, it is used to estimate the shape parameters of the records left for validation. The predictions should be similar to the values gathered during the second step, from the original flux maps.

5. Finally, the model can be applied for predicting the behavior of the field, which includes the effect of control actions, as in this case. Concisely speaking, the field model consists of an expression to represent flux maps and an overlying model (known as 'meta-model' in [33]) to adjust its shape parameters on demand.

The target field for which the proposed controller has been designed is the one used to test the methodology introduced in [33], and it is described in Section 5. The analytical expression used to describe the flux map that a certain heliostat *h* projects on the receiver, f_h, is the bi-variant Gaussian distribution shown in Equation 1. It has been selected because of its flexibility to directly model non-circular maps, which is coherent with the observations made in [14, 32].

$$f_h = \frac{P}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right)} \quad (1)$$

P is the estimated total power contribution of the heliostat. It is measured in kW and can be scaled to reflect changes in DNI and unpredictable situations, such as, soiling. The variable ρ is the correlation between dimensions X and Y of the receiver. The variables σ_x and σ_y correspond to the standard deviation along X and Y, respectively. These four parameters define the overall flux shape on the receiver, and the model can compute them for every heliostat of the field at any solar position. This information is generated depending on the positions of the Sun and the heliostat and stored in a preliminary step. Regarding μ_x and μ_y, which are the means from a mathematical point of view, they are linked to the coordinates of the aim point of the heliostat. They are under the control of the method outlined in Section 3.2 first and under that of the controller described in Section 4 at the end.

Although the analytical expression which models flux distributions (Equation (1)) is the same as the one used in [17], the procedure described herein for estimating its parameters also considers the solar position, i.e., time. This aspect was not required for the instantaneous scope of that method, but it is mandatory for designing a complete controller.

3.2 On Accurately Configuring the Heliostat Field

The method designed in [17] by the authors of this paper for replicating any flux distribution on the receiver, at a given time (solar position), will be known as the model-based flux optimizer. It benefits from the analytical formulation of the flux maps projected by the heliostats of the field to compute off-line a configuration that replicates any given reference. This approach aims to be general because it permits looking for any flux feature (e.g., homogeneity or low spillage) by defining the appropriate reference.

Its design is based on solving a large-scale and continuous optimization problem. The objective function to minimize measures the difference between the reference and the result achieved according to the field optical model. The resolution method combines two separate layers to select and aim the heliostats.

The first layer is responsible for finding the subset of heliostats to activate. It uses a genetic algorithm that works with binary strings in which value 0 at position *i* means that heliostat *i* will not be activated. Analogously, value 1 has the opposite meaning. Regarding its workflow, the procedure starts by generating a random initial population. The number of active heliostats is bounded after scanning the total power contained in the reference and the offer of heliostats according to the model. The minimum number of active heliostats for any individual results from accumulating the estimated power contribution of the most powerful ones until reaching that contained in the reference. The maximum number of active heliostats for any individual is computed by repeating the previous accumulation but with the least powerful heliostats according to the model and considering 25% of their estimated power contribution as if they aimed at a corner. The interested reader can find further details about the bounds in [17]. After that, it follows a classic evolutionary loop. Each loop iteration involves the following stages: i) parent selection, ii) reproduction, iii) spring mutation and iv) replacement. When the loop finishes after a given limit of iterations, the optimizer returns the best individual found as the solution.

Every progenitor is selected as the best individual out of a random sample, which is called tournament-based selection. The reproduction is through uniform crossover by generating a random binary reproduction mask and two descendants from every pair. The first descendant takes its binary values from the first progenitor at the positions in which the binary mask is 1 and from the second one elsewhere. The second one is built by inverting this rule. Mutation consists in randomly activating and deactivating heliostats on the offspring and re-evaluating. Finally, the population for the next iteration is formed by selecting the best individuals from the current one and the descendants, which is called elitism.

Since the genetic optimizer works with binary strings, it also needs a way to automatically generate full solutions that can be evaluated in terms of the objective function during the search. Thus, the optimizer also includes a dedicated module that completes any binary string by assigning a valid aim point to each active heliostat. It is called the auto-aiming module and serves to produce final flux maps that can be compared to the reference. This module works by iteratively aiming the most powerful active heliostat to the highest difference peak between the reference flux map and the predicted one.

The second layer works with the best solution found by the previous one. It tries to improve the aim point assignment heuristically made by the auto-aiming module. To do so, it applies a gradient descent optimizer that exploits the analytical formulation of the problem. The selection of active heliostats is not further changed at this point. This part focuses on improving the initial and heuristic aim point selection to minimize the objective function. The previous layer does not work in a continuous search space, but in a discrete one: the auto-aiming module is limited to the discretization of the input reference. In contrast to it, the second layer is not limited in that way.

4 Heuristic Control Strategy

The method described in the previous section fulfils its requirements at the expense of a high computational cost, which is incompatible with on-line operation [15]. Moreover, it only works in a theoretical framework, and perfect modelling does not exist. Therefore, although the previous tools serve to configure the target field, different discrepancies may appear between the theoretical result and the flux map achieved when translating it to the plant. The control logic described in this section tries to overcome this problem.

The feedback control system developed in this work starts after applying the configuration computed for the plant by the model-based flux optimizer described in Section 3.2. It takes from the referred optimizer both the heliostats subset and their aim points as its initial state. Nevertheless, the field model is still taken into account: it serves to predict the effect of the changes proposed by the control logic between iterations before real feedback is available. Figure 2 summarizes the design approach. As shown, the system is intended for operating in closed loop: The configuration is applied, compared to the setpoint and iteratively corrected according to the feedback.

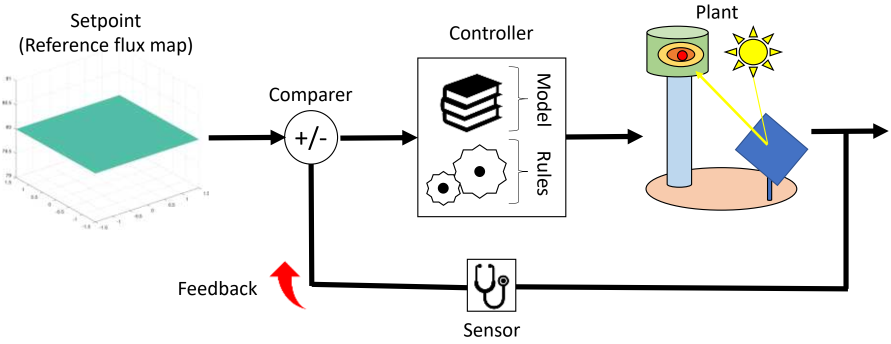


Figure 2
Design of a closed-loop controller

The control strategy considers the following three actions to improve the achieved flux distributions: i) Activating new heliostats, ii) Deactivating heliostats and iii) Changing the aim point of active heliostats.

The activation of heliostats is carried out when the difference between the total power contained in the setpoint and that reflected on the receiver is greater than a given threshold. In that situation, a new heliostat will be activated and aimed at the receiver. The process is repeated while the triggering condition is met and using the field model to update the flux map between iterations. After that, the system applies the changes and waits for new feedback to decide what to do again. The procedure is described in Algorithm 1. It takes as input the setpoint, the achieved flux map, the activation threshold ThrA (from Threshold of Activation), and a parameter Max_Scope, which is explained below.

Algorithm 1

Steps for activating new heliostats by the controller

```
1: scope = 0
2: While Power(Set_Point – Achieved) > ThrA && scope < Max_Scope {
3:     hel = Find the heliostat in Field.Inactives with the highest flux peak <= Max(Set_Point - Achieved)
4:     If hel != {} {
5:         [coord_x, coord_y] = Find_Max(Set_Point - Achieved)
6:         hel.Aim_At(coord_x, coord_y)
7:         scope = scope + 1
8:         Achieved = Achieved + Model.Inflate(hel)
9:     } else{ break }
10: }
11: Field.ApplyChanges()
```

As shown, the heliostat selection criterion requires that its flux peak is as close as possible, without exceeding, to the greatest difference between the setpoint and the achieved flux map. Every new heliostat is aimed at the point with the highest difference in flux density, which tries to compensate for the potential lack of power. However, the activation logic might decide not to add any heliostat. This happens when no heliostat with an appropriate flux peak is found. Adding that type of heliostat could generate further discrepancies between the setpoint and the achieved maps. It is also necessary to highlight that the achieved or measured map is filtered to minimize noise. Otherwise, noisy singular points can misguide the control action. Furthermore, since the heliostat selection and the estimation of the effect of changes are based on the field model, no more than "Max_*Scope*", a limited and user-given number of heliostats, can be added. This approach avoids making too many changes without feedback, which would introduce back the limitations of modelling. The appropriate number of virtual or model-based changes allowed must be adjusted after a preliminary tuning stage and depends on the quality of the model. Finally, notice how the achieved flux map must be updated according to the model for the next iteration as done at line 8.

Analogously, the system might consider that there are too many active heliostats, and it is necessary to deactivate some of them to reach the setpoint. It does not matter if they were part of the initial state or loaded in a previous activation stage. This action is triggered when there is more power on the receiver than in the setpoint, and the difference is greater than a given threshold as before. In that situation, the active heliostat with the nearest aim point to the greatest difference, and the lowest flux peak that does not exceed it, will be deactivated. The process is repeated while the triggering condition is met and limiting the maximum number of changes allowed again. After that, the system applies the changes and waits for new feedback to make a new decision. The procedure is detailed in Algorithm 2. It takes the same input as the previous one with the only exception of the threshold parameter, which is now labelled as *ThrD* (from Threshold for Deactivation) to differentiate it from the previous one. The achieved flux map is synthetically updated at line 7 in this case.

Algorithm 2

Steps for deactivating heliostats by the controller

```
1: scope = 0
2: While Power(Achieved - Set_Point) > ThrD && scope < Max_Scope {
3:     hel = Find the heliostat in Field.Actives with the closest
       aim point to Find_Max(Achieved - Set_Point) and the lowest
       flux peak <= Max(Achieved - Set_Point)
4:     If hel != {} {
5:             hel.Deactivate()
6:             scope = scope + 1
7:             Achieved = Achieved - Model.Inflate(hel)
8:     } else { break }
9: }
10: Field.ApplyChanges()
```

Finally, if no heliostat has been either activated or deactivated, the power balance is considered acceptable. Then, the controller studies the necessity of changing the aim points of the active heliostats. In this case, the focus is on the total power instead of on the differences in flux density.

The process starts by comparing the achieved flux map and the setpoint. The maximum and minimum differences are identified. The maximum difference indicates a region of the receiver in which there is more flux density than desired. On the contrary, the minimum difference means that there is less power there than expected. After identifying both points, the system seeks the heliostat that is aiming at the closest point to where the maximum difference is, and with a flux peak lower or equal to the minimum difference. If any, it aims that heliostat from its current point to where the minimum difference is. The process is repeated while the triggering condition is met and relying on the model-based updates to estimate the effect of changes. Finally, the system commits the changes to the plant and waits for new feedback to decide what to do again. Algorithm 3 contains a detailed description of the method.

Algorithm 3

Steps for re-aiming heliostats by the controller

```
1: scope = 0
2: While DefinedCondition && scope < Max_Scope {
3:     peak = Max(Achieved - Setpoint)
4:     hole = Min(Achieved - Setpoint)
5:     helioCandidates = Sort Field.Actives ascending by the
       distance of their aim point to peak
6:     hel = Starting from the beginning of helioCandidates, find
       the heliostat (not previously selected in this stage) with
       the highest flux peak <= |Min(Achieved - Setpoint)|
7:     If hel != {} {
8:             Achieved = Achieved - Model.Inflate(hel) # Old state
9:             hel.Aim_At(hole.coord_x , hole.coord_y)
10:            scope = scope + 1
11:            Achieved = Achieved + Model.Inflate(hel) # New state
12:    } else { break }
13: }
14: Field.ApplyChanges()
```

Some ideas from the previous algorithm must be further explained. Firstly, aside from not moving too many heliostats without feedback, which must be tuned as introduced, the triggering condition of this process also considers a configurable rule. This can be defined in general terms, e.g., the maximum root-mean-square error value between the setpoint and the achieved map, or with a more specific condition (such as the maximum standard deviation when looking for homogeneity). Secondly, as can be seen at line 6, re-aiming the same heliostat several times is not permitted to avoid unproductive changes and loops. Thirdly, as in the previous cases, the model serves to estimate the effect of changes. However, it is used twice this time. First, at line 8, the effect of not aiming the selected heliostat is estimated by subtracting its predicted flux map to the achieved one. After that, at line 11, the effect of re-aiming that heliostat is simulated by adding its flux map centered on its new aim point. Finally, the method might not change anything at a certain point. This does not mean that the achieved flux map is a perfect replica of the setpoint, but that the controller considers that it cannot improve the current result.

After having described the instantaneous control actions, there are also some dynamics concerns to consider when the method runs continuously. Chattering may appear, i.e., some heliostats are activated and deactivated several times from a time step to the other. This issue is addressed by defining some dwell time that restricts the minimum time allowed between two switches in the state of every heliostat.

5 Experimentation and Results

This section describes the experimentation carried out to assess the performance of the developed controller. The target field is the CESA-I, which belongs to *Plataforma Solar de Almería (PSA)* (Spanish for Solar Platform of Almería).

The field is in the south-east of Spain, at location 37°5′30″ N, 2°21′30″ W. It has 300 heliostats which are north of the receiver. Each one of them has 12 facets (3.05 x 1.1m^2 each) and a total surface close to 40 m^2. The heliostats are deployed in a nearly flat surface with a slope of 0.9% rising from the tower towards the north. The receiver, which will be treated as a flat 3 x 3 m^2 square sampled at 0.04 x 0.04 m^2 steps for defining the setpoint, is at 86.6 m above the ground. It is due north and tilted 33° towards the field. The receiver is on a cylindrical tower 5 m in diameter and approximately 80 m in height. Figure 3 shows the field distribution from the tower base (left) and a real picture taken from the tower (right).

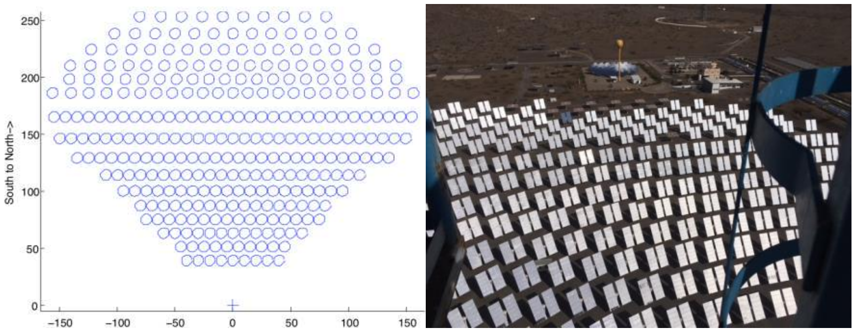


Figure 3

Design of the heliostat field CESA-I (left) and real picture of it from the tower (right)

The solar field has been modelled with the ray-tracing software Tonatiuh, which is considered as the reality, in this work, for practical reasons. Further information regarding this model can be found in [33].

The date considered is 21st June, 2018, at 12:00 (local time). The measured DNI is 745.467 W/m^2. The goal is to produce a homogeneous flux distribution of 80 kW/m^2 and 739.6 kW over the receiver, which is a popular kind of setpoint in the literature. Figure 4 shows the setpoint defined for the model-based flux optimizer (left), the result that it achieves according to the model (right), and its real shape after applying the result to reality (bottom-middle). As can be seen, the theoretical result fulfils the requirements: The model-based flux optimizer automatically activates and aims 60 heliostats achieving a flux distribution in range [78, 82] kW. It takes approximately 222 seconds, and the theoretical result contains 739.1 kW with a Standard Deviation (STD) of 1.2 kW/m^2 (that of the setpoint is 0 kW/m^2).

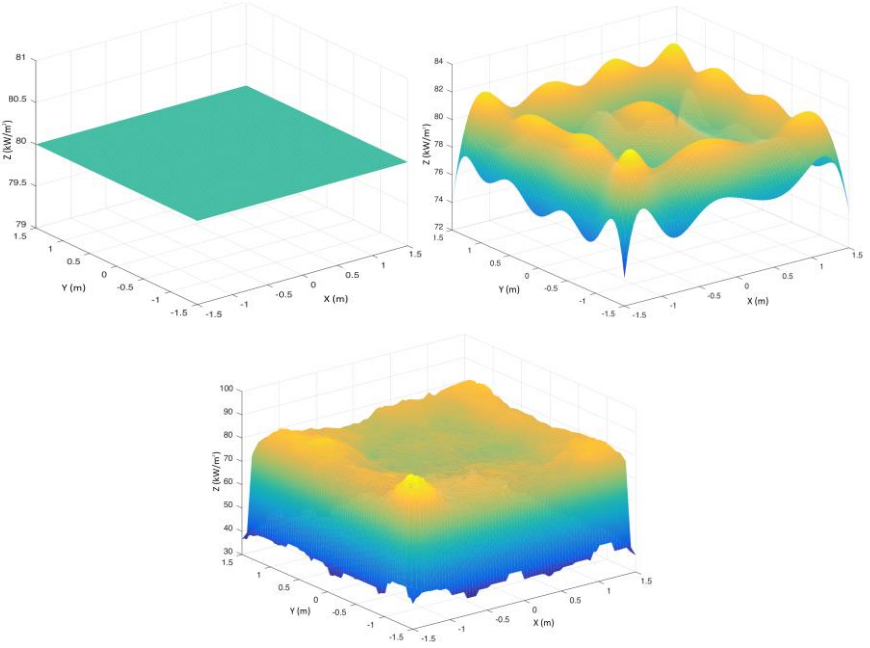


Figure 4

Setpoint to achieve on the receiver (left), theoretical result of the optimizer (right) and real output of the plant under that configuration (bottom-middle)

Logically, when the computed configuration is translated to reality (Figure 4, bottom-middle), the quality decreases. It contains 708.4 kW and its STD is 3.5 kW/m^2. The result is valid, especially considering the yearly scope of the field model and the general design of the model-based flux optimizer. It is even better, i.e., STD from 3.5 to 2.6 kW/m^2, if the metrics focus on the central zone to coincide with the real receiving area. Nevertheless, the result is expected to improve with the designed controller. For instance, the flux density in a corner is over the setpoint, while the central zone is below it.

The controller has been tuned as follows: The maximum number of heliostats that can be activated, deactivated or moved without obtaining new feedback, i.e., Max_Scope, is set to 4. The significant difference in power between the setpoint and the achieved flux map is set to 8 kW for ThrA and 2.4 kW for ThrD. These values are based on the total power contributions estimated by the field model at the considered instant and on the fact that deactivating might be more urgent, i.e., ThrD < ThrA. Finally, the ad-hoc triggering condition to launch the re-aiming action depends on the STD of the achieved flux map. It is launched with values greater than 2.0 in the central zone.

The controller starts by adding two heliostats to the initial 60. This first change increases the total power from 708.4 to 739.2 kW, which is almost equal to the target power. The overall STD is improved from 3.5 to 3.4 kW/m^2 (2.6 to 2.1 kW/m^2 in the central area). Note that the power predicted with the model after updating was 740.4 kW, and the overall and central STD values were 3.3 and 2.2 kW/m^2, respectively. Thus, the model was useful to estimate the effect of the changes applied.

At the second step, the controller opts for changing the aim points of some of the active heliostats. Specifically, it executes four iterations of its internal loop that moves heliostats from the hottest zones to those with less power. However, it is only able to move three heliostats in the end. The last iteration fails to find an active heliostat with flux peak as low as required and different from those chosen in the three previous iterations. After applying these changes, there are 742.2 kW on the receiver, i.e., only 0.35% higher than the 739.6 kW of the setpoint. The overall STD is improved from 3.4 to 3.0 kW/m^2, and from 2.1 to 1.7 kW/m^2 in the central zone. This time, the behavior predicted by the model after applying the changes, i.e., moving three heliostats, where 742.0 kW in power, and overall and central STD values of 3.0 and 1.6 kW/m^2, respectively. Therefore, the field model is not only useful for initial the model-based flux optimizer, but also for the controller to synchronize the changes proposed over gathered the flux map.

Figure 5 shows the flux distribution obtained after applying the changes of the controller over the initial field configuration up to the second step. As can be seen, the flux distribution is homogeneous in general. The effect of reducing the STD values from 3.5 and 2.6 to 3.0 and 1.7 kW/m^2, respectively, is evident. It is clearly better than the initial flux map before applying the control changes (Figure 4, bottom-middle), especially considering the depression of the central zone in that flux map. In fact, the total power is now almost identical to the target, i.e., 739.6 kW, namely, 742.2 kW after the controller versus the initial value of 708.4 kW.

Finally, the deactivation stage is triggered with the previous input. It deactivates one heliostat to decrease the highest flux peak at the bottom left corner. After doing so, the method leaves 61 active ones. The result achieved is shown in Figure 6 and, as can be seen, it is even better than the previous one: the bottom left flux peak is significantly less pronounced. In fact, the visualization axes were automatically reduced from 100 to 90 kW/ m^2. At this point, the total power is slightly lower than desired, i.e., 736.4 kW, but still almost identical to the target of 739.6 kW. Moreover, the STD values are now 2.9 kW/ m^2 for the overall shape and 1.5 kW/ m^2 for the central area. In this case, the model predicted 735.0 kW in power and 3.0 and 1.6 kW/ m^2 in overall and central STD, which are near to the values achieved in the reality.

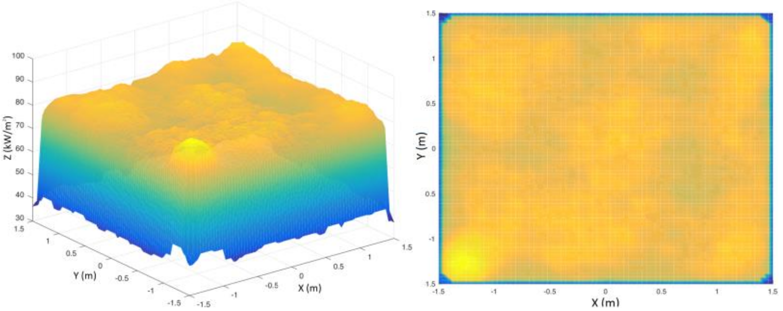


Figure 5

Flux distribution achieved on the receiver surface after applying the two changing steps proposed by the designed controller. Overall (left) and top (right) views

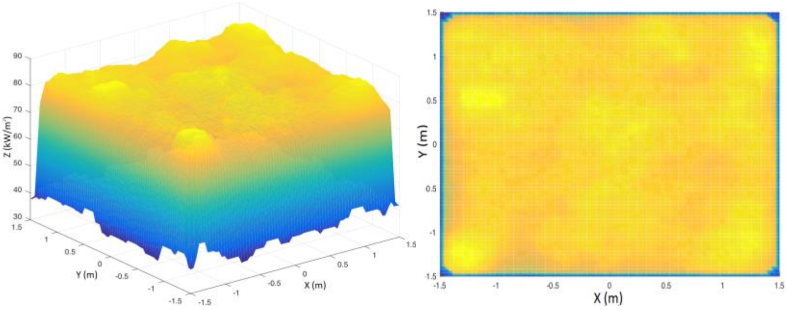


Figure 6

Flux distribution achieved on the receiver surface after applying the three changing steps proposed by the designed controller. Overall (left) and top (right) views

After the previous change, the measured feedback does not trigger any additional action of the proposed controller. The difference in power is not big enough to consider activating or deactivating, and the STD in the center is also in the desired range. Thus, the controller has only executed 3 phases, and it has improved the initial and model-based starting point. The total power has raised from 708.4 to 736.4 kW (739.6 kW was desired), and the overall STD has been reduced from 3.5 to 2.9 and from 2.6 to 1.5 kW/ m^2 in the overall shape and in the center, respectively.

Conclusions

Controlling the flux distribution that a heliostat field produces on its receiver is important. It affects the efficiency and safety of the plant, and even its production costs, considering the expenses associated with the receiver. Unfortunately, it is also a challenging problem that involves selecting the heliostats to use and their aim points. There are different methods to address this problem in the literature. While most of them focus on adjusting the aim point and obtaining homogeneous flux distributions, the authors of this work have proposed a general flux optimization method linked to a new modelling strategy. Its theoretical results fulfil the requirements, but their quality worsens when translated to the plant due to inherent modelling errors.

This paper proposes a simple feedback controller that can perform three different actions: activating, deactivating and re-aiming heliostats. The decision between the two first options depends on the difference in power between the setpoint and the obtained flux map. The activation of the third one is adapted to the goal and based on comparing flux densities between the setpoint and the achieved flux map. The process combines real feedback with an analytical field model. The control logic has been added to the workflow defined by the flux map optimizer and its associated field model. It compensates for internal modelling errors by making a few changes based on the input flux map.

In the experimentation carried out, the setpoint has been set to a homogeneous flux distribution, which is the most studied target in the literature. According to the results obtained, the control logic improves the flux distribution that results from directly applying the configuration computed by the model-based flux optimizer. Considering that the setpoint has 739.6 kW and 0 kW/m^2 in standard deviation, the controller raises the achieved power from 708.4 to 736.4 kW and reduces the central standard deviation from 2.6 to 1.5 kW/m^2.

As future work, the control strategy will be tested considering the apparent solar movement instead of a fixed point in time. The software package Tonatiuh will also be replaced with real measurements when they can be taken from the target field.

Acknowledgement

This work has been supported by the Spanish Ministry of Economy, Industry and Competitiveness under grants RTI2018-095993-B-100 and DPI2017-85007-R, and the Spanish Junta de Andalucía under grant UAL18-TIC-A020-B, co-funded by FEDER funds. N. C. Cruz has been supported by an FPU fellowship from the Spanish Ministry of Education under grant FPU14/01728.

References

[1] S. Kiwan and S. Al Hamad, "On analyzing the optical performance of solar central tower systems on hillsides using biomimetic spiral distribution", *Journal of Solar Energy Engineering*, Vol. 141, No. 1, pp. 1-12, 2019

[2] J. Wang, L. Duan and Y. Yang, “An improvement crossover operation method in genetic algorithm and spatial optimization of heliostat field,” *Energy*, Vol. 155, pp. 15-28, 2018

[3] K. Wang, Y. L. He, Y. Qiu and Y. Zhang, "A novel integrated simulation approach couples MCRT and Gebhart methods to simulate solar radiation transfer in a solar power tower system with a cavity receiver," *Renewable Energy*, Vol. 89, pp. 93-107, 2016

[4] M. Saghafifar, M. Gadalla, and K. Mohammadi, "Thermo-economic analysis and optimization of heliostat fields using AINEH code: Analysis of implementation of non-equal heliostats (AINEH)," *Renewable Energy*, Vol. 135, pp. 920-935, 2019

[5] V. S. Reddy, S. C. Kaushik, K. R. Ranjan and S. K. Tyagi, "State-of-the-art of solar thermal power plants — A review," *Renewable and Sustainable Energy Reviews*, Vol. 27, pp. 258-273, 2013

[6] F. J. Collado and J. Guallar, "Fast and reliable flux map on cylindrical receivers," *Solar Energy*, Vol. 169, pp. 556-564, 2018

[7] C. J. Noone, A. Ghobeity, A. H. Slocum, G. Tzamtzis and A. Mitsos, "Site selection for hillside central receiver solar thermal plants," *Solar Energy*, Vol. 85, No. 5, pp. 839-848, 2011

[8] S. Alexopoulos and B. Hoffschmidt, “Advances in solar tower technology,” *WIREs Energy and Environment*, Vol. 6, No. 1, pp. 1-19, 2017

[9] D. Y. Goswami, *Principles of Solar Engineering (3rd Ed).* Taylor & Francis, 2015

[10] A. Salomé, F. Chhel, G. Flamant, A. Ferrière and F. Thiery, "Control of the flux distribution on a solar tower receiver using an optimized aiming point strategy: Application to THEMIS solar tower," *Solar Energy*, Vol. 94, pp. 352-366, 2013

[11] M. Astolfi, M. Binotti, S. Mazzola, L. Zanellato and G. Manzolini, "Heliostat aiming point optimization for external tower receiver," *Solar Energy*, Vol. 157, pp. 1114-1129, 2017

[12] K. Wang, Y. L. He, X. D. Xue and B. C. Du, "Multi-objective optimization of the aiming strategy for the solar power tower with a cavity receiver by using the non-dominated sorting genetic algorithm," *Applied Energy*, Vol. 205, pp. 399-416, 2017

[13] S. M. Besarati, D. Y. Goswami and E. K. Stefanakos, "Optimal heliostat aiming strategy for uniform distribution of heat flux on the receiver of a solar power tower plant," *Energy Conversion and Management*, Vol. 84, pp. 234-243, 2014

[14] A. Grobler, “Aiming strategies for small central receiver systems,” Master’s degree dissertation, Stellenbosch University, 2015

[15] T. Ashley, E. Carrizosa and E. Fernández-Cara, "Optimisation of aiming strategies in Solar Power Tower plants," *Energy*, Vol. 137, pp. 285-291, 2017

[16] A. Sánchez-González, M. R. Rodríguez-Sánchez and D. Santana, "Aiming factor to flatten the flux distribution on cylindrical receivers," *Energy*, Vol. 153, pp. 113-125, 2018

[17] N. C. Cruz, J. D. Álvarez, J. L. Redondo, M. Berenguel and P. M. Ortigosa, "A two-layered solution for automatic heliostat aiming," *Engineering Applications of Artificial Intelligence*, Vol. 72, pp. 253-266, 2018

[18] E. F. Camacho and A. J. Gallego, “Advanced control strategies to maximize ROI and the value of the concentrating solar thermal (CST) plant to the grid,” in *Advances in Concentrating Solar Thermal Research and Technology*, Woodhead Publishing, Elsevier, 2017, Ch. 14, pp. 311-336

[19] L. Roca, R. Díaz-Franco, A. de la Calle, J. Bonilla, L. J. Yebra and A. Vidal, "A combinatorial optimization problem to control a solar reactor," *Energy Procedia*, Vol. 49, pp. 2037-2046, 2014

[20] A. Sánchez-González, M. R. Rodríguez-Sánchez, and D. Santana, "Aiming strategy model based on allowable flux densities for molten salt central receivers," *Solar Energy*, Vol. 157, pp. 1130-1144, 2017

[21] A. Grobler and P. Gauché, "A review of aiming strategies for central receivers," in Proceedings of the second Southern African Solar Energy Conference, pp. 1-8, 2014

[22] F. Glover and M. Laguna, Tabu Search. Kluwer, Norwell, MA, 1997

[23] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Machine Learning*, Vol. 3, No. 2, pp. 95-99, 1998

[24] Q. Yu, Z. Wang and E. Xu, "Analysis and improvement of solar flux distribution inside a cavity receiver based on multi-focal points of heliostat field," *Applied Energy*, Vol. 136, pp. 417-430, 2014

[25] B. Belhomme, R. Pitz-Paal and P. Schwarzbözl, "Optimization of heliostat aim point selection for central receiver systems based on the ant colony optimization metaheuristic," *Journal of Solar Energy Engineering*, Vol. 136, No. 1, pp. 011005-011012, 2014

[26] M. Dorigo and C. Blum, "Ant colony optimization theory: A survey," *Theoretical Computer Science*, Vol. 344, No, 2. pp. 243-278, 2005

[27] D. Maldonado, R. Flesch, A. Reinholz and P. Schwarzbözl, "Evaluation of aim point optimization methods. In AIP Conference Proceedings," in *AIP Conference Proceedings*, Vol. 2033, No. 1, pp. 1-8, 2018

[28] T. Ashley, E. Carrizosa and E. Fernández-Cara, "Inclement weather effects on optimal aiming strategies in solar power tower plants" in Proceedings of SolarPACES, DOI: 10.1063/1.5067041, 2018

[29] A. Kribus, "Closed loop control of heliostats," *Energy*, Vol. 29, No. 5-6, pp. 905-913, 2004

[30] M. R. Convery, "Closed-loop control for power tower heliostats," in *Proceedings of SPIE*, Vol. 8108, pp. 81080M-1, 2011

[31] J. Freeman and L. R. Chandran, " Closed loop control system for a heliostat field," in *2015 IEEE International Conference on Technological Advancements in Power and Energy*, pp. 272-277, 2015

[32] A. J. Gallego and E. F. Camacho, "On the optimization of flux distribution with flat receivers: A distributed approach," *Solar Energy*, Vol. 160, pp. 117-129, 2018

[33] N. C. Cruz, R. Ferri-García, J. D. Álvarez, J. L. Redondo, J. Fernández-Reche, M. Berenguel, R. Monterreal and P. M. Ortigosa, "On building-up a yearly characterization of a heliostat field: A new methodology and an application example," *Solar Energy*, Vol. 173, pp. 578-589, 2018

[34] M. Blanco, A. Mutuberria, A. Monreal and R. Albert, "Results of the empirical validation of Tonatiuh at Mini-Pegase CNRS-PROMES facility," in *Proceedings of SolarPACES*, 2011

[35] C. H. Chen, C. C. Chung, F. Chao, C. M. Lin and I. J. Rudas, "Intelligent robust control for uncertain nonlinear multivariable systems using recurrent cerebellar model neural networks," *Acta Polytechnica Hungarica*, Vol. 12, No. 5, pp. 7-33, 2015

[36] L. Breiman, "Random forest," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001

Applying Expert Heuristic as an a Priori Knowledge for FRIQ-Learning

Tamás Tompa, Szilveszter Kovács

Department of Information Technology, University of Miskolc,
Miskolc-Egyetemváros, H-3515 Miskolc, Hungary
e-mail: tompa@iit.uni-miskolc.hu, szkovacs@iit.uni-miskolc.hu

Abstract: Many Reinforcement Learning methods start the learning phase from an empty, or randomly filled knowledge-base. Having some a priori knowledge about the way as the studied system could be controlled, e.g. in the form of some state-action control rules, the convergence speed of the learning process can be significantly improved. In this case, the learning stage could start from a sketch, from a knowledge-base formed based upon the already existing knowledge. In this paper. the a priori (expert) knowledge is considered to be given in the form state-action fuzzy control rules of a Fuzzy Rule Interpolation (FRI) reasoning model and the studied reinforcement learning method is restricted to be a Fuzzy Rule Interpolation-based Q-Learning (FRIQ-Learning) method. The main goal of this paper is the introduction of a methodology, which is suitable for merging the a priori state-action fuzzy control rule-base to the initial state-action-value function (Q-function) representation. For demonstrating the benefits of the suggested methodology, the a priori knowledge-base accelerated FRIQ-Learning solution of the "mountain car" benchmark is also discussed briefly in the paper.

Keywords: Reinforcement Learning; Heuristically Accelerated Reinforcement Learning; Fuzzy Rule Interpolation; Q-Learning; FRIQ-Learning

1 Introduction

The reinforcement learning (RL) (originally introduced in [22]), is still a popular machine learning algorithm among the devices of the computational intelligence. The RL methods are kind of trial-and-error type algorithms, solving problems without the explicit knowledge about the solution, but based on rewards leading to the targeted behaviour of the system. The rewards (reinforcements) are given by the environment, according to the observed and targeted behaviour, independently from the inner states of the RL agent. The original Q-learning [31] and the Fuzzy Q-learning (FQ-learning) [10], [4], [7] algorithms are starting with an empty knowledge-base and building their approximated Q-function during iterations based on the gained reward values. The Q-function representation in the case of

Q-learning is a Q-table. In the case of FQ-learning it is a fuzzy rule-base describing the Q-function in continuous state and action universes. These RL algorithms automatically build their knowledge-base during the learning process. Therefore, they can be applied in such situations, where there is no initial knowledge about the system to be controlled. These methods are starting from an empty knowledge-base at the beginning of the learning process, and their Q-function approximation is built through iterations. On the other hand, if there is an initial knowledge-base about the operating process may exist, this knowledge could form a draft for the initial RL model.

There are existing solutions for combining the RL methods with an initial expert knowledge-base, which is noted in this case as "heuristic". These methods can be used in such systems, where the knowledge, or a portion of the knowledge about the system operation already exists, but the full control needs to be extended and adjusted based of the feedback of the working environment. One of these methods is the Heuristically Accelerated Reinforcement Learning, HARL [5]. In the HARL the heuristic is given in a form of a heuristic function ($H_t(s_t,a_t)$). It defines for the agent, which "a_t" action should be selected in the state "s_t", at the time "t". The combination of the HARL model with the traditional Q-learning, is called Heuristically Accelerated Q-learning (HAQL) [5]. Another possible solution for describing the heuristic is the formal knowledge representation with a declarative language. The "GOAL" is an agent programming language, which is defining the action selection for the agent by a set of "if then" type conditions [8]. In fuzzy rule-based Q-learning [10] the Q-function is represented by a fuzzy rule-base. In this case, it is straightforward that the a priori information of the expert should be also represented as a fuzzy rule-base. In [21] Pourhassan et al. are proposing a way to incorporate the expert knowledge in the Q-learning by means of fuzzy rules.

The main goal of this paper is to suggest a way for extending the Fuzzy Rule Interpolation (FRI) model-based Reinforcement Learning (FRIQ-Learning) methods to be able to adopt a priori expert knowledge about the problem solution in the form of fuzzy rules.

2 The FRIQ-Learning

The Fuzzy Rule Interpolation-based Q-learning (FRIQ-learning) [26][9] is an extension of the traditional Q-learning with continuous state and action space, represented by a FRI model. For Fuzzy Q-learning (FQ-learning) techniques, usually, the classical 0-order Takagi-Sugeno Fuzzy Inference model is adopted (see, e.g. [1], [3] and [11] for more details). In case of FRIQ-learning adapting FRI methods (see e.g. [2] for a short overview of FRI methods) for the FQ-learning can reduce the size of the fuzzy rule-base, by permitting the use of sparse fuzzy rule-bases for fuzzy knowledge representation. The sparse rule-base built for the FRI can represent the same or nearly the same approximated Q-function as the complete fuzzy rule-base with the classical (e.g. CRI [18]) fuzzy reasoning.

One of the available FRI techniques is the Fuzzy Rule Interpolation based on Vague Environment (FIVE) FRI, which was originally introduced in [13], [14] and [15]. The FIVE is a multidimensional, application-oriented FRI technique, which is based on the Vague Environment (VE) [12] concept. According to the VE concept, the fuzzy partitions of the antecedent and consequent universes can be represented by scaling functions [12]. The similarities of fuzzy sets can be calculated as the scaled distances of crisp points. Therefore, the FIVE can give a crisp conclusion directly without any additional defuzzification step. The combination of the FQ-learning with the FIVE FRI is called FRIQ-learning [26]. In the FRIQ-learning the state-action-value function is described by a sparse fuzzy rule-base and the Q-function is approximated by the FIVE FRI. The form of the i^{th}, $i \in [1,r]$ fuzzy rule in the Q-function rule-base is the following:

If s_1 **is** S^i_1 **And** s_2 **is** S^i_2 **And … And** s_n **is** S^i_n **And** a **is** A^i **Then** $\tilde{Q}(\boldsymbol{s},a) = q^i$ (1)

where S^i_j $j \in [1,n]$ is a label of a fuzzy set in the j^{th} dimension of the n dimensional state space $\boldsymbol{S}$, $\boldsymbol{s} \in \boldsymbol{S}$ is the n dimensional state observation, s_j is the j^{th} dimension of the state observation $\boldsymbol{s}$, A^i is the label of a fuzzy set in the one-dimensional action space U, $a \in U$ is the selected action, $\tilde{Q}(\boldsymbol{s},a)$is the approximated Q-function, q^i is the singleton conclusion of the i^{th} fuzzy rule. Applying the FIVE FRI model for the Q-function representation, according to [16], we get the following formulas:

$$\tilde{Q}(\boldsymbol{s},a) = \begin{cases} q^i & \text{if } (\boldsymbol{s},a) = (\boldsymbol{s}^i,a^i) \text{for some } i, \\ \sum_{i=1}^{r} \left(\left(q^i \big/ \left(\delta_{\mathrm{v}}^i\right)^{\lambda} \right) \Big/ \left(\sum_{j=1}^{r} 1 \big/ \left(\delta_{\mathrm{v}}^j\right)^{\lambda} \right) \right) & \text{otherwise.} \end{cases} \tag{2}$$

where q_i is the consequent of the i^{th} rule, $(\boldsymbol{s},a)$ is the crisp observation, λ is the Shepard parameter and r is the number of the rules in the rule-base. The δ_{v}^i is the scaled distance of the actual observed state, selected action $(\boldsymbol{s},a)$ value and the i^{th} fuzzy rule antecedent according to the scaling function v of the corresponding vague environment [12]:

$$\delta_{\mathrm{v}}^i = \delta_{\mathrm{v}}\left((\boldsymbol{s},a),(\boldsymbol{s}^i,a^i)\right) = \left[\sum_{j=1}^{n} \left(\int_{s_j^i}^{s_j} \mathrm{v}_j(s_j) ds_j \right)^2 + \left(\int_{a^i}^{a} \mathrm{v}(a) da \right)^2 \right]^{1/2} \tag{3}$$

where $(\boldsymbol{s},a)$ is the actual state and action, $(\boldsymbol{s}^i,a^i)$ is the antecedent part of the i^{th} rule, s_j is the j^{th} dimension of the n dimensional state universe, $\mathrm{v}_j(s_j)$ is the scaling function of the s_j state universe, $\mathrm{v}(a)$ is the scaling function of the action universe *U*.

Substituting the formulas of the FIVE FRI (2) to the update form of the Q-learning, we get the q_i rule consequent of the *i*th fuzzy rule in the (*k*+1)th iteration in the following form:

$$q_i^{k+1} = \begin{cases} q_i^k + \Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) & \text{if } (\boldsymbol{s},a) = (\boldsymbol{s}^i, a^i) \text{ for some } i, \\ q_i^k + \Delta\tilde{Q}^{k+1}(\boldsymbol{s},a)\cdot\left(1/\delta_{\mathrm{v},i}^{\lambda}\right)/\left(\sum_{i=1}^{r} 1/\delta_{\mathrm{v},i}^{\lambda}\right) & \text{otherwise.} \end{cases} \tag{4}$$

where $\Delta\tilde{Q}^{k+1}(\boldsymbol{s},a)$ is the (*k*+1)th update value of the Q-function in $(\boldsymbol{s},a)$:

$$\tilde{Q}^{k+1}(\boldsymbol{s},a) = \tilde{Q}^{k}(\boldsymbol{s},a) + \Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) \tag{5}$$

$$\Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) = \alpha\cdot\left(g(\boldsymbol{s},a,\boldsymbol{s}') + \gamma\cdot\max_{a'\in U}\tilde{Q}^{k}(\boldsymbol{s}',a') - \tilde{Q}^{k}(\boldsymbol{s},a)\right) \tag{6}$$

In this form, as in the original Q-learning [31], γ is the discount factor and $\alpha\in[0,1]$ is the step size parameter. The q_i^{k+1} is the *k*+1 iteration of the singleton conclusion of the *i*th fuzzy rule, taking action *a* in state $\boldsymbol{s}$, $\boldsymbol{s}'$ is the new observed state, $g(\boldsymbol{s},a,\boldsymbol{s}')$ is the observed reward completing the $\boldsymbol{s}\rightarrow\boldsymbol{s}'$ state-transition. The $\tilde{Q}^{k}$ and the $\tilde{Q}^{k+1}$ are the *k*th and the (*k*+1)th iteration of the Q-function approximated by the FIVE FRI (2).

For the action selection policy, the FRIQ-learning applies the greedy policy (or optionally the ε-greedy policy) [27], which is always selecting the action having the greatest Q value (or in case of ε-greedy, the greatest with ε probability) in the corresponding state. The greedy policy can be described by the following form:

$$\pi(s) = \arg\max_{a\in U} Q^{\pi}(\boldsymbol{s},a) \tag{7}$$

The FRIQ-learning was also extended with an automatic incremental rule-base creation method [27]. In this technique, based on reinforcements, the Q-function rule-base can be built automatically through iterations. The method starts from an "empty" rule-base, in which the rules are at the corners of the (n+1)-dimensional action-state space hypercube (this is required because of the definition of the interpolation, where n is the dimension of the state). In the further iteration steps, the initial rule-base grows according to the values of the updating rule (4). A new rule is inserted to the rule-base if the updating value of the state-action-value function ($\Delta\tilde{Q}$) is greater than a predefined limit and the existing rules are farther than a given distance. The position of the newly inserted rule is the closest possible (enabled) rule position (see [27] for more details). In case if the update value is smaller than the predefined limit or the given state-action point is close to an existing rule, then only the existing fuzzy rules consequents are updated.

The method has rule-base reduction strategies too. They are based on the approach, that for approximating the Q-function there is no need for all the rules created in the incremental phase. Some of the (redundant) rules could be removed without a relevant change in the Q-function representation. Applying the reduction strategies, these rules can be omitted from the rule-base. There are four different reduction strategies defined in [24], [28] and [29]. Three of them are based on the differences in the close rule consequences (Q-values) [28] [29] and one is based on rule clustering [24].

3 Expert Heuristic as a Priori Knowledge for Defining the Initial Rule-Base of the FRIQ-Learning

In case if there are some a priori knowledge about the system to be controlled, e.g. some kind of expert rules exist, the convergence speed of the Q-learning could be improved by their adoption. In this paper, the suggested way of this adoption is the merging of the expert knowledge to the initial rule-base of the FRIQ-learning. The expert knowledge, as an a priori information, is defined by a human expert before the learning process. In this paper, the suggested way of the expert knowledge expression is in the form of fuzzy rules. This case the a priori rule-base can be directly adapted to the initial fuzzy rule-based Q-function representation of the FRIQ-learning. During the learning process this initial knowledge representation will be tuned and modified (extended or reduced), then at the end of the process the expert rules can also be fetched back e.g. for expert rule validating purposes.

For merging the expert rules to the initial rule-base of the FRIQ-learning, the problem of the different rule representations must be solved. In the case of the FRIQ-learning, the fuzzy rules are state-action-value rules according to form (1), while the expert knowledge is usually expressed in the form of state-action production rules (8). The fuzzy rule consequences of the state-action-value Q-function representation in FRIQ-learning are the Q-values. On the other hand, the fuzzy rule consequences of the state-action production rules are expert-defined actions. The suggested form of the i^{th}, $i \in [1, r]$ expert-defined production fuzzy rule is the following:

$$\textbf{If } s_1 \textbf{ is } \hat{S}_1^i \textbf{ And } s_2 \textbf{ is } \hat{S}_2^i \textbf{ And … And } s_n \textbf{ is } \hat{S}_n^i \textbf{ Then } a = \hat{A}^i \qquad (8)$$

In structure, the form of (8) is very similar to (1), except the different types of consequents and the missing action antecedent in (8). The i^{th} rule consequent $\hat{A}^i$ is the expert-defined rule action for a given n dimensional state $\hat{\boldsymbol{S}}^i = \left[\hat{S}_1^i, \hat{S}_2^i, \ldots, \hat{S}_n^i\right]$.

For adapting the expert rules (8) in the initial rule-base of the Q-function representation, the missing Q-values must be determined. Considering the initial expert rules, as "valuable" decisions about the actions, and taking into account of the planned greedy action selection policy, the Q-values of the expert rules must be set to relatively higher initial values.

Having a different action selection policy than a greedy one, the given expert rules can be also considered to be a heuristic policy modifier [5]. Considering the expert rules to be always and unquestionably true, the greedy policy of the FRIQ-learning can be turned into a heuristic policy, which obeys the expert rules in the following form:

$$\pi(\boldsymbol{s}) = \begin{cases} a = \hat{A}^i, & \text{if } \boldsymbol{s} = \hat{\boldsymbol{S}}^i, \textit{ for some } i \\ \arg\max\limits_{a \in U} Q^{\pi}(\boldsymbol{s}, a) & \textit{otherwise.} \end{cases} \tag{9}$$

where $\hat{\boldsymbol{S}}^i$ and $\hat{A}^i$ are the state and action fetched from the i^{th} expert rule, and $\boldsymbol{s}$ is the actual state observation. If the actual observation $\boldsymbol{s}$ matches the state $\hat{\boldsymbol{S}}^i$ of one of the expert state-action rules, then the selected action will be the corresponding action $\hat{A}^i$. Otherwise, the greedy action selection of the FRIQ-learning will be followed.

Considering the expert rules to be always and unquestionably true, with the greedy policy for the rest of the state-action space, the FRIQ-learning can only extend the initial rule-base of the expert by additional rules. In this case, the goal of the suggested FRIQ-learning-based methodology is the extension of the expert-defined state-action rules by additional state-action rules for the state space area uncovered by the expert rules.

In case of supposing that the expert rules may be false, or incorrect, the goal of FRIQ-learning-based methodology can be extended by the tuning of the initial expert rule-base (moving, removing, or updating the expert rules).

It is also important to note that because of the incremental manner of the rule-base construction during the learning phase, there is no need for defining state-action expert rules for all the possible states (like an optimal policy), but it is sufficient to give the expert rules only in any states, where the expert has knowledge about the system. Therefore, the expert might define any number of state-action pairs. Thus, if the tuning of the initial expert rule-base is permitted during the learning phase, the quality of the expert-defined a priori information effects only the convergence rate of the FRIQ-learning.

4 Adaptation of the Expert Rule-Base

The suggested expert rule-base adaptation method of the FRIQ-learning is built upon two phases. In the first phase, the Q-value determination method calculates the initial approximated Q-values for the expert-defined rules. In the second phase the rule-base adaptation method combines the expert-defined state-action a priori rule-base (6) with the FRIQ-learning initial rule-base (1). Thereafter the combined rule-base will serve as the initial Q-function approximation rule-base of the FRIQ-learning process.

4.1 Determining the Initial Q-Values of the Expert Rules

The consequents of the expert rules are actions (defining a states-action function). Therefore, the rules of the expert knowledge representation have no Q-values. On the other hand, the rule representation of the FRIQ-learning describes a state-action-quality function (Q-function), where the quality of the state-action pairs must be determined. Therefore, to adopt the a priori expert rules to the Q-function rule representation, the corresponding Q-values must be determined. It must be done before the learning phase as an initialization step of the Q-function rule-base generation.

The goal of the proposed Q-function rule-base initialization method is to determine the initial, estimated Q-value ($\tilde{Q}_{\text{init}}$) for each expert-defined state-action rules before the learning phase. According to the proposed concept, the rule Q-values should be initialized with an expert-defined quality ($\hat{Q}_{\text{init}}$) value. I.e. the Q-values of the expert rules, together with the state reward value definitions are an inherent part of the expert knowledge representation. With full confidence, these values can not be determined independently from the corresponding expert rules. On the other hand, it could happen, that the expert heuristic contains only the worthy production rules, without any additional information related to the initial Q, or reward values. In this case it can be supposed, that the expert knowledge representation contains only the most important correct rules, and if the expert rule Q-values are missing, the initial Q-values of the expert rules can be approximated by a relatively “high” Q-value. The relatively “high” Q-value is an estimation. As a first straightforward estimate, if the initial Q-values definition is missing from the expert knowledge representation, this paper suggests setting the initial Q-value to be the same for all the rules, as a fraction of the estimated maximal achievable Q-value ($\tilde{Q}_{\max}$). Where $\tilde{Q}_{\max}$ can be approximated based on the maximal reward can be given by the environment. The initial Q-value $\tilde{Q}_{\text{init}}$ can estimated by the following formula:

$$\tilde{Q}_{\text{init}} = \eta \cdot \tilde{Q}_{\max} \tag{10}$$

$$\tilde{Q}_{\max} = \lim_{k\to\infty}\tilde{Q}^{k+1}\left(s^*,a^*\right) = \lim_{k\to\infty}\left(\tilde{Q}^{k}\left(s^*,a^*\right)+\alpha\cdot\left(g\left(s^*,a^*,s^*\right)+\gamma\cdot\tilde{Q}^{k}\left(s^*,a^*\right)-\tilde{Q}^{k}\left(s^*,a^*\right)\right)\right)$$

$$\tilde{Q}^{k}\left(s^*,a^*\right) = \max_{a'\in U}\tilde{Q}^{k}\left(s^*,a'\right) \text{ and } g\left(s^*,a^*,s^*\right) = \max_{s\in S, a\in U} g\left(s,a,s'\right) = g_{\max}$$

$$\begin{aligned}\tilde{Q}_{\max} &= \lim_{k\to\infty}\tilde{Q}^{k+1}\left(s^*,a^*\right) = \lim_{k\to\infty}\left(\tilde{Q}^{k}\left(s^*,a^*\right)+\alpha\cdot\left(g_{\max}+\left(\gamma-1\right)\cdot\tilde{Q}^{k}\left(s^*,a^*\right)\right)\right) = \\ &= \tilde{Q}^{k}\left(s,a\right)+\alpha\cdot g\left(s,a,s'\right)+\alpha\cdot(\gamma-1)\cdot\tilde{Q}^{k}\left(s',a'\right) = \frac{\alpha\cdot g_{\max}}{-\alpha\cdot\left(\gamma-1\right)} = \frac{g_{\max}}{1-\gamma}\end{aligned} \quad (11)$$

$$\tilde{Q}_{\text{init}} = \eta\cdot\frac{g_{\max}}{1-\gamma}\text{, in case if } \gamma<1 \quad (12)$$

where $\tilde{Q}_{\max}$ is the estimated maximal achievable Q-value having $g_{\max}$ being the maximal reinforcement could be given by the environment. $\eta\in[0,1]$ is the discount factor of the $\tilde{Q}_{\text{init}}$ estimation.

There are other initial Q-values approximation methods can be found in the literature, e.g. for discrete state Q-learning in [19] the initial Q-values are determined based on the reward value of the goal state applying a binary reward function, for fuzzy Q-learning in [21] the initial Q-values are determined based on expert knowledge related to the estimated Q values of some states.

4.2 Merging the Expert Rules with the Initial Rule-Base of the FRIQ-Learning

Being in interpolated Q function representation, the initial rules of the FRIQ-learning has to hold the corners (corner rules) of the $(n+1)$-dimensional state-action space [27]. Therefore, the number of the initial fuzzy rules are 2^{n+1}. E.g. in case of two states and one action, it is $2^{2+1}=8$. If the number of the expert rules is $\hat{r}$, the size of the initial merged rule-base has $2^{n+1}+\hat{r}$ rules. According to the FRIQ-learning initial rule definition suggested in [27], the initial rule consequent values of the corner rules are $q_i=0$. Therefore, the *i*th corner rule *r*□i has the following format:

If *s*$_1$ **is** S$_1$□i **And** *s*$_2$ **is** S$_2$□i **And…And** *s*$_n$ **is** S$_n$□i **And** *a* **is** A□i **Then** $\tilde{Q}(s,a)$=0 (13)

where S$_l$□i $\in[\min(S_l),\max(S_l)],\forall i,l$, A□i $\in[\min(A),\max(A)],\forall i$ are the corner state and action values and see Eq. (1) for the rest of the notation.

For the expert rules, the initial rule consequent values are $q_i=\tilde{Q}_{\text{init}}$, $i\in[1,\hat{r}]$ hence the ith expert rule, $\hat{r}$ i has the following format:

If *s*$_1$ **is** $\hat{S}_1^i$ **And** *s*$_2$ **is** $\hat{S}_2^i$ **And…And** *s*$_n$ **is** S^i_n **And** *a* **is** $\hat{A}^i$ **Then** $\tilde{Q}(s,a)=\tilde{Q}_{\text{init}}$ (14)

where see Eq. (8) for the notation.

In case if there is an expert rule, which hit the position of the initial corner rules, the overlapping expert rule will replace the corresponding initial corner rule.

The main steps of the suggested Q-function rule-base initialization are summarized on Figure 1.

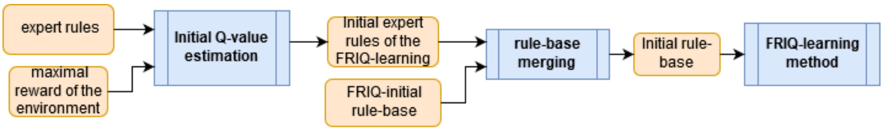

expert rules
maximal reward of the environment
Initial Q-value estimation
Initial expert rules of the FRIQ-learning
FRIQ-initial rule-base
rule-base merging
Initial rule-base
FRIQ-learning method

Figure 1
The suggested FRIQ-learning rule-base initialization

5 FRIQ-Learning and the Mountain Car Problem

The goal of this section is to give an application example for the suggested merging of the a priori state-action expert rules with the initial corner rule-base of the FRIQ-learning (i.e. for the suggested rule-base initialization) in a standard RL task.

The chosen example is the well-known “mountain car” RL benchmark example. The mountain car problem is the task of a car for getting out of a deep valley. Initially, the car is situated in the center of the valley. The goal is to get out of the valley by going to the top of the hill within a given time frame. In this example the problem is considered to be solved if the car gets out of the valley in less than 1000 iteration steps. If the car reached the goal or the 1000 iteration steps elapsed, then an episode is completed. The full learning phase will be done if the Q-update values are smaller to a predefined Q-update limit (e.g. 0.05) through some episodes and if the size of the rule-base is not changed (not adding a new rule).

The problem has two states and one action variable. The states descriptors are the velocity and the position of the car and the action variable is the left, right, or neutral movement of the car:

- s1: velocity of the car
- s2: position of the car
- a: movement of the car (left, right, neutral)

The rule-base initialization is done according to the suggested expert rule-base merging discussed in section 4. The benefit of the suggested rule-base initialization is measured by the achievable performance gain.

In the first example, the performance of the expert rules extended initial rule-base will be compared to the empty initial rule-base (according to Eq. 12). The effect of the expert rule-base quality will be also studied by comparing a well-formed proper initial expert rule-base in the second example to a partially correct and in the third example to a randomly generated initial "expert" rule-base. During the performance investigation of the well-formed proper initial expert rule-base, the effect of the proper initial rule consequent value $\tilde{Q}_{\text{init}}$ selection will be also discussed.

The performance of the FRIQ-learning can be characterized by the convergence rate of the learning. In this paper, the convergence rate is calculated as the average number of episodes required for adapting the Q-function rule-base to be able to solve the mountain car problem. One episode lasts till the car gets out from the valley, or 1000 iteration steps without solution. The averages, the convergence rate and the number of the required rules, are estimated based on independent runs starting from different initial state space positions. The reward given by the environment is $g_{\max}$=100 (an expert suggested value) for the state if the car reaches the goal position (top of the valley in less than 1000 iteration steps). During the iteration, Eq. (6) was applied for the $\tilde{Q}$ updates. The learning parameters were the followings:

- learning rate (α): 0.5
- discount factor (γ): 0.99

If the system starts from an empty rule-base without the expert-defined initial rules (corner rules only), then the average convergence rate of 10 independent run became 28.3 episodes, with 91.7 rules (at the end of the incremental rule-base creation phase, before the rule-base reduction, see Table 1. for the detailes).

Table 1
The results when starting with the empty knowledge-base

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	23	36	34	35	20	34	25	26	29	21	28.3
Rule-base size	80	85	82	96	105	90	89	98	99	93	91.7

The first task of the suggested Q-function rule-base initiation method, is the estimation of the $\tilde{Q}_{\text{init}}$ values according to Eq. (11). For the $\tilde{Q}_{\text{init}}$ values estimation we have to determine a suitable value of the η discount factor (see Eq. (11)). The effect of the η discount factor is problem dependent. In this paper, we study its effect on the mountain car problem in case of having a properly set initial expert rule-base.

In our example the properly set initial expert rule-base was generated by a single run of the automatic incremental rule-base creation technique introduced in [27], together with the rule-base reduction strategies III and IV introduced in [28] and [24]. The remaining 17 rules (see e.g. on Table 2) became the properly set initial expert rule-base of our example.

Table 2
Rules of the well-formed proper initial expert rule-base, where the **a** is the rule consequent

R#	1	2	3	4	5	6	7	8	9
s_1	-0.5	-0.475	-0.475	-0.27	-0.27	-0.475	-0.065	-0.475	-0.68
s_2	0	-0.014	0.014	0.014	-0.014	0.042	-0.014	-0.042	0.042
a	-1	1	1	1	1	1	0	-1	-1

R#	10	11	12	13	14	15	16	17
s_1	-0.065	-0.065	0.14	-0.27	-0.885	-0.65	-1.09	0.14
s_2	0.042	0.014	-0.014	-0.042	0.042	0.042	0.042	-0.014
a	1	0	-1	-1	-1	0	-1	0

The maximal reward value (g_{max}) is defined by the expert. In this example it is set to 100. From the maximal reward value, the suggested $\tilde{Q}_{\text{init}}$ initial Q-value was calculated according to Eq. (11). Table 3 contains the initial rule-base (expert rules merged with the FRIQ initial rules before the learning phase).

Table 3
The initial Q-values rule-base with the well-formed proper initial expert rules, where the *Q* is the rule consequent

R#	1	2	3	4	5	6	7	8	9
s_1	-0.5	-0.475	-0.475	-0.27	-0.27	-0.475	-0.065	-0.475	-0.68
s_2	0	-0.014	0.014	0.014	-0.014	0.042	-0.014	-0.042	0.042
a	-1	1	1	1	1	1	0	-1	-1
Q	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$

R#	10	11	12	13	14	15	16	17
s_1	-0.065	-0.065	0.14	-0.27	-0.885	-0.65	-1.09	0.14
s_2	0.042	0.014	-0.014	-0.042	0.042	0.042	0.042	-0.014
a	1	0	-1	-1	-1	0	-1	0
Q	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$	$\tilde{Q}_{\text{init}}$

R#	18	19	20	21	22	23	24	25
s_1	-1.5	-1.5	-1.5	-1.5	0.3450	0.3450	0.3450	0.3450
s_2	-0.07	-0.07	0.07	0.07	-0.07	-0.07	0.07	0.07
a	-1	1	-1	1	-1	1	-1	1
Q	0	0	0	0	0	0	0	0

The first 17 rules in Table 3 are the rules of the expert rule-base with the suggested $\tilde{Q}_{\text{init}}$ values estimation. The last 2^3=8 rules (18…25) are the initial corner rules of the empty Q-function rule-base, according to Eq. 13.

The next step is the η discount factor estimation (see Eq. (11), (12)) by checking its effect to the convergence rate having the properly set initial expert rule-base. Table 4 demonstrates the dependency of the convergence rate from the value of the η discount factor with the corresponding initial $\tilde{Q}_{\text{init}}$ values according to Eq. (12).

Table 4

The convergence rate in case of different η discount factor values (γ=0.99, $g_{\max}$ =100)

η	$\tilde{Q}_{\text{init}}$	convergence rate (episodes)
1	10000	23
0.75	7500	23
0.6	6000	30
0.37	3700	29
0.075	750	25
0.00015	1.5	27

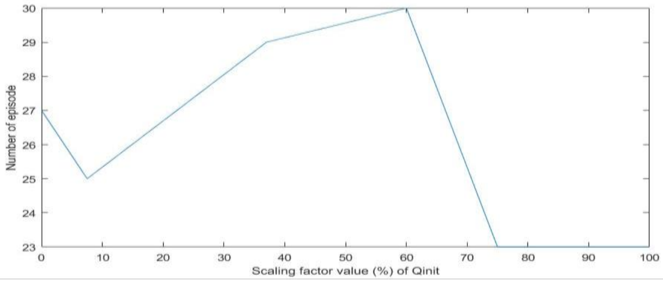


Figure 2

The convergence rate in case of different η discount factor values (γ=0.99, $g_{\max}$ =100)

According to the results (see Figure 2), in this given mountain car example the best convergence rate (23 episodes) can be achieved if the η discount factor is between 0.75 and 1 (γ=0.99, $g_{\max}$ =100).

For checking the performance of the suggested merging of the a priori state-action expert rules with the initial rule-base of the FRIQ-learning, five tests were performed. The first example is the properly defined expert heuristic case, where the initial rule-base of the FRIQ-learning is constructed with all the properly given initial expert rules having the suggested $\tilde{Q}_{\text{init}}$ values (according to Eq. (12)) (Table 5). The second example is a partial lack of knowledge, where the initial rule-base of the FRIQ-learning is constructed from a fragment of the properly given initial expert rules (Table 6). The third example is a partially proper expert knowledge, where some of the expert given initial rules are incorrect (Table 8). The fourth example is a fully incorrect expert knowledge, where all the expert given initial rules are incorrect (Table 10). The fifth example is a full lack of knowledge, where the initial rule-base of the FRIQ-learning is constructed without any expert given initial rules (see the results in Table 1).

The effect of the properly given initial expert rules is demonstrated on Table 5. In this case, the system found the final solution (the car gets out of the valley within 1000 iteration steps) in 10 episodes with 124.3 rules averagely. The expert rule-base contains 17 properly given initial expert rules. This rule-base is merged with the empty rule-base of the FRIQ-learning, forming the 25 rules of the suggested initial rule-base (see Table 3). To reduce the final rule-base size one of the FRIQ-learning reduction strategies [24], [28], [29] could be applied.

Table 5

The results starting with properly given initial expert rules

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	10	20	17	7	11	10	6	5	6	8	10
Rule-base size	108	125	139	109	135	129	107	124	133	134	124.3

The effect of the partial lack of knowledge, where the initial rule-base of the FRIQ-learning is constructed from a fragment of the properly given initial expert rules, is demonstrated on Table 6. In this case, the system found the final solution in 14.4 episodes with 114.3 rules averagely. The expert rule-base contains 10 rules from the original 17 properly given initial expert rules. This partial expert rule-base is merged with the empty rule-base of the FRIQ-learning, forming the 18 rules of the suggested initial rule-base.

Table 6

The results starting with partial lack of initial expert rules

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	20	13	10	7	7	15	29	15	22	6	14.4
Rule-base size	107	85	102	85	98	96	111	107	110	98	114.3

The effect of the partially proper expert knowledge, where some of the expert given initial rules are incorrect, is demonstrated on Table 8. In this case, the system found the final solution in 11.7 episodes with 120.1 rules averagely. The expert rule-base contains 11 rules from the original 17 properly given initial expert rules and 6 rules, where the rule consequents are changed to actions (see Table 7, rules no. 1, 2, 3, 15, 16, 17) which have an incorrect conclusion. This partially proper expert rule-base is merged with the empty rule-base of the FRIQ-learning, forming the 25 rules of the suggested initial rule-base.

Table 7

The partially correct expert rule-base

R#	1	2	3	4…14	15	16	17
s_1	-0.5	-0.475	0.475	…	-0.68	-1.09	0.14
s_2	0	-0.014	-0.014	…	0.042	0.042	-0.014
a	0	1	-1	…	0	0	1

Table 8

The results when starting with the partially correct expert rules

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	8	16	8	13	7	16	10	15	16	7	11.7
Rule-base size	115	134	126	133	135	126	123	135	147	127	120.1

The effect of the fully improper expert knowledge, where all the expert given initial rules are incorrect, is demonstrated in Table 10. In this case, the system found the final solution in 26.6 episodes with 124.4 rules averagely. In this example, the initial "expert" rule-base is 17 randomly generated rules (see Table 9). This improper expert rule-base is merged with the empty rule-base of the FRIQ-learning, forming the 25 rules of the suggested initial rule-base.

Table 9

The randomly generated "expert" rule-base

R#	1	2	3	4	5	6	7	8	9
s_1	-0.475	-0.5	-0.475	-0.475	-0.27	-0.27	-0.27	-0.475	-0.475
s_2	0	0	-0.014	0.014	0	-0.014	0	-0.042	0
a	1	-1	-1	0	-1	0	-1	1	1

R#	10	11	12	13	14	15	16	17
s_1	-0.475	-0.065	0.14	-0.27	-0.885	0.885	-0.065	-1.09
s_2	0	0	-0.014	-0.042	0.042	0.042	0.042	0.042
a	-1	0	1	-1	-1	1	0	-1

Table 10

The results when starting with the randomly generated "expert" rules

Run case	1	2	3	4	5	6	7	8	9	10	Average
Convergence rate	29	56	19	16	24	18	37	29	20	17	26.6
Rule-base size	122	127	118	124	131	120	130	124	127	121	124.4

Table 11 summarizes the results of the five different expert rule-base quality cases.

Table 11

The effect of the expert rule-base quality upon the convergence rate and the rule-base size

Expert rule-base type	Convergence rate	Rule-base size
empty	28.3	91.7
properly given	10	124.3
properly given fragment	14.4	114.3
partially incorrect	11.7	120.1
randomly generated	26.6	124.4

Conclusions

In this paper a methodology is suggested, which is suitable for merging an expert heuristic (an a priori state-action fuzzy production rule-base) to the initial state-action-value (Q-function) rule-base of the FRIQ-learning system. The expert-defined a priori rule-base is a preliminary knowledge about the given RL problem. The suggested merging is based on the reformulation of the expert heuristic given in the form of production (state - action) rules to the rule format (state, action - Q value) of the Q-function representation fuzzy rule-base by adding initial Q-values as consequents to them. The proper initial Q-values of the expert rules must be defined by the expert together with the rule definition. With full confidence, these values cannot be determined independently from the corresponding expert rules. For determining the initial Q-values in case if the expert heuristic contains only the worthy production rules, without any additional information related to the initial Q, or reward values, this paper suggest to set the initial Q-value to be the same for all the rules, as a fraction of the estimated maximal achievable Q-value.

For demonstrating the performance of the suggested initial FRIQ-learning rule-base construction methodology, the quality effect of the merged a priori expert rule-base is discussed. The performance of the FRIQ-learning solution of the “mountain car” benchmark example is studied in the case if the a priori state-action expert rules are fully properly defined, partly properly defined, partly improperly defined and fully improperly defined. The results are compared to the lack of a priori knowledge in average convergence rate and in rule-base size (without rule filtering). The best performer in convergence rate was the initial rule-base constructed with the fully properly defined a priori expert rules. The fully improperly defined a priori expert rules has similar convergence performance as the FRIQ-learning starting from an empty initial rule-base. On the other hand because of the unfiltered incremental manner of the rule-base creation, in rule-base size, the FRIQ-learning starting from an empty initial rule-base has the best performance.

The benefits of the suggested expert heuristic injection to the FRIQ-learning are twofold. The first is the improvement of the convergence speed, as it was discussed in this paper. The second is a way for validating the expert heuristic in given situations. Marking the injected expert heuristic rules during the Q-function initialization and fetching them back after the learning phase, the change of the expert production rules can be determined. Small changes can support, large changes or rule disappearance can disapprove the validity of the expert heuristic in the given situation defined by the environment of the learning phase. This kind of validation of the expert heuristic could be beneficial in application areas, where heuristical rule-based models exists, but the collection of vast data has some difficulties, like adaptive affective [30], or ethorobotical [20] models applied for human-machine interaction.

Acknowledgement

The described article was carried out as part of the EFOP-3.6.1-16-00011 ”Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialisation” project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

References

[1] Appl, M.: Model-based Reinforcement Learning in Continuous Environments. Ph.D. thesis, Technical University of München, München, Germany, dissertation.de, Verlag im Internet (2000)

[2] Baranyi, P., Kóczy, L. T., Gedeon, T. D.:A Generalized Concept for Fuzzy Rule Interpolation, IEEE Trans. on Fuzzy Systems, Vol. 12, No. 6, 2004, pp. 820-837

[3] Berenji, H. R.: Fuzzy Q-Learning for Generalization of Reinforcement Learning. Proc. of the 5th IEEE International Conference on Fuzzy Systems (1996) pp. 2208-2214

[4] Berenji, H. R.: Fuzzy Q-Learning for Generalization of Reinforcement Learning. Proc. of the 5th IEEE International Conference on Fuzzy Systems, pp. 2208-2214, 1996

[5] Bianchi, Reinaldo AC, Carlos HC Ribeiro, and Anna HR Costa. "Accelerating autonomous learning by using heuristic selection of actions." *Journal of Heuristics* 14.2 (2008): 135-168

[6] Bianchi, Reinaldo AC, Carlos HC Ribeiro, and Anna HR Costa. "Accelerating autonomous learning by using heuristic selection of actions." *Journal of Heuristics* 14.2 (2008): 135-168

[7] Bonarini, A.: Delayed Reinforcement, Fuzzy Q-Learning and Fuzzy Logic Controllers. In Herrera, F., Verdegay, J. L. (Eds.) Genetic Algorithms and Soft Computing, (Studies in Fuzziness, 8), Physica-Verlag, Berlin, D, (1996), pp. 447-466

[8] Broekens, Joost, Koen Hindriks, and Pascal Wiggers. "Reinforcement learning as heuristic for action-rule preferences." *International Workshop on Programming Multi-Agent Systems*. Springer Berlin Heidelberg, 2010

[9] D. Vincze, Fuzzy Rule Interpolation and Reinforcement Learning. Proceedings of the IEEE 15th International Symposium on Applied Machine Intelligence and Informatics, Herlany, Slovakia (2017) pp. 173-178

[10] Glorennec, Pierre Yves. "Fuzzy Q-learning and dynamical fuzzy Q-learning."*Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the Third IEEE Conference on*. IEEE, 1994

[11] Horiuchi, T., Fujino, A., Katai, O., Sawaragi, T.: Fuzzy Interpolation-Based Q-learning with Continuous States and Actions. Proc. of the 5th IEEE International Conference on Fuzzy Systems, Vol. 1 (1996) pp. 594-600

[12] Klawonn, F.: Fuzzy Sets and Vague Environments, Fuzzy Sets and Systems, 66, 1994, pp. 207-221

[13] Kovács, Sz., Kóczy, L. T.: Approximate Fuzzy Reasoning Based on Interpolation in the Vague Environment of the Fuzzy Rule base as a Practical Alternative of the Classical CRI. Proceedings of the 7th International Fuzzy Systems Association World Congress, Prague, Czech Republic, 1997, pp. 144-149

[14] Kovács, Sz., Kóczy, L. T.: The use of the concept of vague environment in approximate fuzzy reasoning. Fuzzy Set Theory and Applications, Tatra Mountains Mathematical Publications, Mathematical Institute Slovak Academy of Sciences, Bratislava, Slovak Republic, Vol. 12, 1997, pp. 169-181

[15] Kovács, Sz.: New Aspects of Interpolative Reasoning. Proceedings of the 6th. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain, 1996, pp. 477-482

[16] Kovács, Szilveszter. "Extending the Fuzzy Rule Interpolation" FIVE" by Fuzzy Observation." *Computational Intelligence, Theory and Applications* (2006): 485-497

[17] Krizsán, Z., Kovács, Sz.: Gradient based parameter optimisation of FRI "FIVE", Proceedings of the 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Budapest, Hungary, November 6-8, pp. 531-538, (2008)

[18] Mamdani, Ebrahim H., and Sedrak Assilian. "An experiment in linguistic synthesis with a fuzzy logic controller." *International journal of man-machine studies* 7.1 (1975): 1-13

[19] Matignon, Laëtitia, Guillaume J. Laurent, and Nadine Le Fort-Piat. "Reward function and initial values: better choices for accelerated goal-directed reinforcement learning." *International Conference on Artificial Neural Networks*. Springer Berlin Heidelberg, 2006

[20] Miklosi, A; Korondi, P; Matellan, V; Gacsi, M: Ethorobotics: A New Approach to Human-Robot Relationship, Frontiers in Psychology, Vol. 8, Paper: 958, 8 p. (2017)

[21] Pourhassan, Mojgan, and Nasser Mozayani. "Incorporating expert knowledge in Q-learning by means of fuzzy rules." *Computational Intelligence for Measurement Systems and Applications, 2009. CIMSA'09. IEEE International Conference on*. IEEE, 2009

[22] Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. Vol. 1, No. 1, Cambridge: MIT press, 1998

[23] The original (discrete state- and action space) cart-pole problem can be found at: http://www.jamh-web.appspot.com/download.htm, last access date: 22.10.2019

[24] Tompa, Tamás, and Szilveszter Kovács. "Clustering-based fuzzy knowledge-base reduction in the FRIQ-learning." *Applied Machine Intelligence and Informatics (SAMI), 2017 IEEE 15th International Symposium on*. IEEE, 2017

[25] Tompa, Tamás, and Szilveszter Kovács. "Determining the minimally allowed rule-distance for the incremental rule-base contruction phase of the FRIQ-learning." 2018 19th International Carpathian Control Conference (ICCC) IEEE, 2018

[26] Vincze, D., Kovács, Sz.: Fuzzy rule interpolation-based Q-learning. *Applied Computational Intelligence and Informatics, 2009, SACI'09, 5th International Symposium on*. IEEE, 2009

[27] Vincze, D., Kovács, Sz.: Incremental Rule Base Creation with Fuzzy Rule Interpolation-Based Q-Learning, I. J. Rudas et al. (Eds.), Computational Intelligence in Engineering, Studies in Computational Intelligence, Volume 313/2010, Springer-Verlag, Berlin Heilderberg, 2010, pp. 191-203

[28] Vincze, D., Kovács, Sz.: Reduced Rule Base in Fuzzy Rule Interpolation-based Q-learning, Proceedings of the 10th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, CINTI 2009, November 12-14, 2009, Budapest Tech, Budapest, pp. 533-544

[29] Vincze, D., Kovács, Sz.: Rule-Base Reduction in Fuzzy Rule Interpolation-Based Q-Learning, Recent Innovations in Mechatronics (RIiM) Vol. 2, (2015) No. 1-2

[30] Vircikova, M., Magyar, G., Sincak, P.: The Affective Loop: A Tool for Autonomous and Adaptive Emotional Human-Robot Interaction. In: Kim JH., Yang W., Jo J., Sincak P., Myung H. (eds) Robot Intelligence Technology and Applications 3. Advances in Intelligent Systems and Computing, Vol. 345, Springer, pp. 247-254 (2015)

[31] Watkins, C. J. C. H.: Learning from Delayed Rewards. Ph.D. thesis, Cambridge University, Cambridge, England (1989)

Modeling and Control of Discrete-Event Systems with Partial Non-Determinism using Petri Nets

František Čapkovič

Institute of Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, 845 07 Bratislava, Slovakia, e-mail: Frantisek.Capkovic@savba.sk

Abstract: Discrete-Event Systems (DES) are systems that are discrete in nature. A next state of DES depends on the actual state and on the occurrence of a discrete event. DES are often modeled and controlled by Petri Nets (PN) of different kinds (place/transition PN, timed PN, etc.). However, not always real DES, are purely deterministic. In such cases, the PN-based model contains some uncontrollable and/or unobservable transitions or unmeasurable/unobservable places. In order to control DES with the partial non-determinism, special kinds of PN and control methods/procedures have to be used. That's just it - applications of Interpreted PN (IPN) and Labeled PN (LbPN) for modeling analyzing and control of DES are investigated here.

Keywords: control; discrete-event systems; modeling; non-determinism; Petri nets

1 Introduction

Discrete-Event Systems (DES) are systems where a next state depends only on the actual state and on the occurrence of a discrete event. It can be said that DES are systems discrete in nature. In other words, such a system persists in a state until it is not forced to change its state in consequence of a discrete event occurrence. Many kinds of real systems in practice have the character of DES - e.g. flexible manufacturing systems, robotized working cells, discrete production lines, some kinds of transport systems, communication systems etc. Here, discrete events have the character of starting or ending of particular operations, synchronization of several operations etc., but also external influences. The simple abstract illustrative example presenting the causal relation between the system states x_i, i=1,..., 6 and the occurrence of discrete events u_j, j=1,..., 5 is given in Figure 1. Here, the system response on the discrete event sequence $\{u_1, u_2, u_3, u_4, u_5\}$ is the sequence of states$\{x_3, x_1, x_2, x_5, x_4, x_6\}$, where x_3 is the initial state. A view on DES and their history is presented in [6]. The global overview on modeling, analysis and control of DES is given in [7].

DES are frequently modeled by means of different kinds of Petri Nets (PN) like Place/Transition PN (P/T PN) - alias ordinary PN, timed PN, hybrid PN, colored PN, etc. However, DES are not always purely deterministic. They can be partially nondeterministic. Some newer, specific kinds of PN, like Labeled PN (LbPN), Interpreted PN (IPN) are able to deal better with modeling, analyzing and control of DES with non-determinism like those mentioned above.

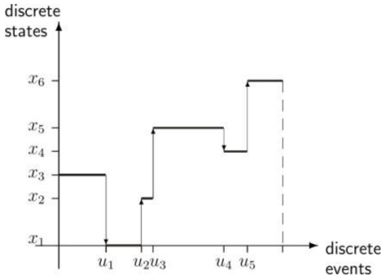


Figure 1

The evolution of states in DES

Non-determinism in DES can be caused, in principle, by means of two different factors: (i) occurrence of *silent events*, that cause a change in the DES state and they are not observable by any external observer; (ii) *indistinguishable events*, whose occurrence at a given DES state yields two or more new states.

1.1 Petri Nets

Essentials of PN were defined by C.A. Petri in his PhD thesis [1] written in German. Later, some novels were brought in [2] - [4]. Nowadays, PN represent a huge variety of PN kinds and methods of their mathematical modeling, analysis and control. The very good survey of PN evolution since [1] is presented in [5].

In principle, P/T PN are bipartite directed graphs with two kinds of nodes and two kinds of edges. Nodes p_i, i = 1, …, n, are named as places and represent elementary states of particular operations in modeled DES. Let $P = \{p_1,...,p_n\}$ is the set of places. Edges t_j, j = 1, …, m, are named as transitions and model the discrete events. Let $T = \{t_1,...,t_m\}$ is the set of transitions. The causality among the states and events is expressed by the sets $F \subseteq P \times \mathrm{T}$, $G \subseteq T \times P$ representing, respectively, the incidence of directed arcs from places to transitions and from transitions to places. Then, the structure of P/T PN can be formally expressed by the following quadruplet

$$<P, T, F, G> \qquad (1)$$

The P/T PN have also their *dynamics* expressing the evolution of the state in the steps k = 0, 1, …, which can be described by the restricted discrete linear vector state equation as follows

$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{B}.\mathbf{u}_k$, $k = 0, 1, \ldots$ (2)

$\mathbf{F}.\mathbf{u}_k \leq \mathbf{x}_k$ (3)

Where,

$\mathbf{x}_k = (x_{p_1}, x_{p_2}, \ldots, x_{p_n})^T$ is the state vector with entries $x_{p_i} \in \{0, 1, \ldots, \infty\}$ being the states of particular places p_i, $i = 1, \ldots, n$, in the step k, namely x_{p_i} represents the actual number of tokens in the place p_i. The vector $\mathbf{x}_0$ is the initial state vector. There are three basic kinds of places: (i) operation places (e.g. in models of flexible manufacturing systems (FMS) they reflect the progress of processing parts in particular steps) being not marked in the initial state, (ii) places representing presence of fix resources (e.g. machine tools, robots, conveyors, buffers, etc. in FMS) being marked in the initial state, (iii) places representing variable resources (e.g. availability of raw materials, semi-products in FMS).

In PT/PN places can be observable/measurable. In IPN and LbPN some places may be unobservable/unmeasurable.

$\mathbf{u}_k = (u_{t_1}, u_{t_2}, \ldots, u_{t_m})^T$ is the control vector with entries $u_{t_j} \in \{0, 1\}$ being the states of particular transitions t_j, $j = 1, \ldots, m$, in the step k. They can be disabled or enabled - namely, when the transition t_j is disabled (i.e. it cannot be fired) then $u_{t_j} = 0$, otherwise it is enabled (i.e. it may be, but not always has to be, fired) and then $u_{t_j} = 1$. In P/T PN enabled transitions represent the occurrence of discrete events which can be observable and/or controllable. However, in IPN and LbPN some events may be spontaneous, i.e. the transitions which model them are uncontrollable and/or unobservable.

$\mathbf{B} = \mathbf{G}^T - \mathbf{F}$ is the structural matrix with $\mathbf{G}$ being the incidence matrix corresponding to the set G and $\mathbf{F}$ being the incidence matrix corresponding to the set F. The term $\mathbf{x}_0$ is the initial state vector.

Starting from $\mathbf{x}_0$ and firing an enabled transition the next state $\mathbf{x}_1$ can be reached. The reachability tree (RT) expresses all possible branches of the development of the system (1) - (2). When all transitions are observable and controllable and all places are measurable/observable, the system development can be controlled without greater problems. For example for a firing sequence of transitions t_a, t_b, …,t_c, the state trajectory will be the following $\mathbf{x}_0 \xrightarrow{u_{t_a}} \mathbf{x}_1 \xrightarrow{u_{t_b}} \ldots \mathbf{x}_{k-1} \xrightarrow{u_{t_c}} \mathbf{x}_k$. Such trajectories represent branches in RT. RT is a standard tree with nodes expressing state vectors $\mathbf{x}_k$, $k = 0, \ldots, K$, where K is the global number of states. K may be infinite too. The states create the state space - the set $\mathcal{R} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k\}$ of reachable states. The edges of RT symbolize the PN transitions.

In the PN theory the term marking M is used instead the term state $\mathbf{x}$. Then, the above introduced trajectory in such a perception and symbolism has the form $M_0[u_{t_a}> \Rightarrow M_1[u_{t_b}> \Rightarrow \cdots \Rightarrow M_{k-1}[u_{t_c}> \Rightarrow M_k$ or $M_0[U> \Rightarrow M_k$, where the set $U = \{u_{t_a}, u_{t_b}, u_{t_c}\}$.

Unfortunately, in PN models of real DES the situation is not so simple because of the above mentioned *silent* and *indistinguishable* transitions. From the DES control point of view it is necessary a deeper view. Thus, it is needful to speak about uncontrollable and/or unobservable transitions and moreover, even about unmeasurable/unobservable places. Namely, when a controller is not allowed to affect some of transitions, the transitions become uncontrollable. Their firing cannot be either inhibited or allowed by any external action. A transition is named to be unobservable when its firing cannot be directly detected or measured. Unobservable transitions model internal events being not observable from outside. Any unobservable transition is implicitly uncontrollable.

Analogically, the state (marking) of an unmeasurable/unobservable place cannot be detected or measured. In such a case the set of observable places is reduced. Thus, the reduced number of entries creates the output vector. It means that an output equation has to be added to the PN model (1), (2).

Consequently, in order to model and control a real DES, new approaches to PN-based modeling and control have to be found. LbPN and IPN are able to help us on such a way.

1.2 A Simple Example of P/T PN

To illustrate the P/T PN structure and dynamics let us introduce a simple example. In Figure 2 the PN structure is presented while in Figure 3a, Figure 3b, respectively, the corresponding RT and RG.

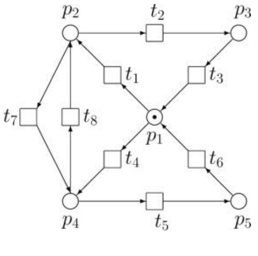


Figure 2
A simple P/T PN

The incidence matrices representing the PN structure are as follows

$$\mathbf{F}=\begin{pmatrix}1&0&0&1&0&0&0&0\\0&1&0&0&0&0&1&0\\0&0&1&0&0&0&0&0\\0&0&0&0&1&0&0&1\\0&0&0&0&0&1&0&0\end{pmatrix};\ \mathbf{G}^T=\begin{pmatrix}0&0&1&0&0&1&0&0\\1&0&0&0&0&0&0&1\\0&1&0&0&0&0&0&0\\0&0&0&1&0&0&1&0\\0&0&0&0&1&0&0&0\end{pmatrix} \qquad (4)$$

and the initial state is

$$\mathbf{x}_0=(1\quad 0\quad 0\quad 0\quad 0)^T. \qquad (5)$$

By means of the model of dynamics development represented by (1) - (2) other four following states can be reached from the initial state

$\mathbf{x}_1 = (0 \quad 1 \quad 0 \quad 0 \quad 0)^T$; $\mathbf{x}_2 = (0 \quad 0 \quad 0 \quad 1 \quad 0)^T$; $\mathbf{x}_3 = (0 \quad 0 \quad 1 \quad 0 \quad 0)^T$ and $\mathbf{x}_4 = (0 \quad 0 \quad 0 \quad 0 \quad 1)^T$

Thus, the space of reachable states $\mathcal{R} = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$. Corresponding RT represents causal relations among states, i.e. among nodes being the state vectors. Such a transit among states is ensured by particular transitions. For drawing P/T PN, analyzing their properties and drawing RT different kinds of graphical simulators can be used. Thus RT of this P/T PN is given in Figure 3. This RT and next RTs are outputs of the graphical simulator of PN. It affects their quality. The corresponding reachability graph (RG) arises by connecting RT nodes with the same name into one node - see Figure 3a. The adjacency matrices in the form **A** or **A**$_d$ (expressing dynamic entries) are the same for both RT and RG, namely

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \text{ or } \mathbf{A}_d = \begin{pmatrix} 0 & u_{t_1} & u_{t_4} & 0 & 0 \\ 0 & 0 & u_{t_7} & u_{t_2} & 0 \\ 0 & u_{t_8} & 0 & 0 & u_{t_5} \\ u_{t_3} & 0 & 0 & 0 & 0 \\ u_{t_6} & 0 & 0 & 0 & 0 \end{pmatrix} \qquad (6)$$

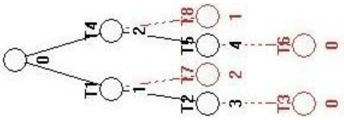


Figure 3a

The corresponding RT (obtained from the graphic simulator of P/T PN)

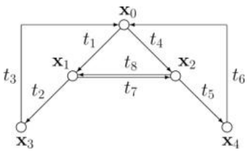


Figure 3b

The reachability graph RG corresponding to RT

1.3 The Paper Organization

While in this Section 1 the PN were defined, in the next Section 2 definitions of LbPN and IPN will be introduced. The particularities of such kinds of PN will be pointed out and shortly illustrated on corresponding examples.

The Section 3 will bring the core issues of the paper concerning the usage of IPN and LbPN in DES modeling and control. Just practical examples of control of real DES working in nondeterministic conditions by means of such kinds of PN represent the main contributions of this paper. Namely, the IPN and LbPN models of such kinds of complicated DES, containing uncontrollable/unobservable transitions and unmeasurable/unobservable places, are able in sequence to show (i) how to describe the non-determinism of such systems; (ii) how they make possible to analyze the nondeterministic DES; (iii) how to synthesize approaches for the control of such DES; (iv) how the control of such DES can be performed.

In the section Conclusion the global evaluation of the paper contributions is introduced.

Finally, the list of References is included.

2 Labeled and Interpreted Petri Nets

LbPN and IPN are purpose-built extended forms of P/T PN. They make possible to deal with the PN containing uncontrollable and/or unobservable transitions and unmeasurable/unobservable places. There are many definitions of LbPN - e.g. in [8]-[12], [14]-[16] - as well as of IPN - e.g. in [13]. It is necessary to say that the primal definition of IPN arose in the Mexican group in CINVESTAV Unidad Guadalajara, where many other papers about IPN besides [13] were written.

2.1 Labeled Petri Nets

In case of LbPN the term net N means the P/T PN structure. Besides N there are additional attributes - the alphabet $\mathcal{L} = L \cup \varepsilon$, where L expresses the observable events and ε represent unobservable events; the labeling function $\ell : T \longrightarrow \mathcal{L}$ assigning events to transitions; the set of reachable states $\mathcal{R}$; the set of finite states $F_x \in \mathcal{R}$ (i.e. $F_x \subset \mathcal{R}$). It means that LbPN can be formally expressed as the following quintuple

$$< N,\ \mathcal{L}, \ell, \mathbf{x}_0, F_x > \qquad (7)$$

There exist three kinds of labeled functions in LbPN, namely: (i) *free* labeling, when ℓ is a one-to-one mapping; (ii) *λ-free* labeling, when two or more transitions share the same label; (iii) *arbitrary* labeling, when $\ell : T \longrightarrow \mathcal{L} \cup \{\lambda\}$ with λ being an empty string.

Although this paper does not deal with the problem of LbPN diagnosability, it is necessary to say that (freely interpreted) the diagnosability of an LbPN, with unobservable transitions, implies [17] that each occurrence of a fault can be detected after a finite number of transition firings.

2.1.1 Illustrative Example

Consider the simple LbPN given in Figure 4. The transitions are labeled by assigned events a, b. The RT of this net is displayed in Figure 5.

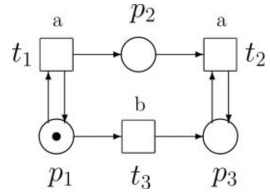


Figure 4
A simple LbPN

Here, in the Figure 4, two kinds of mappings can be seen - namely *free* (at the transitions t_1 and t_3) and *λ-free* (at the transitions t_1 and t_2). The RT of the LbPN is given in Figure 5. As we see there, the RT contains ambiguities. Namely, there are two cycles there, pointing out to two ambiguous states $\mathbf{x}_1$ and $\mathbf{x}_3$. Namely, the set $\mathcal{R}$ consists of the following states

$$\mathbf{x}_0 = (1\ \ 0\ \ 0)^T;\ \mathbf{x}_1 = (1\ \ \omega\ \ 1)^T;\ \mathbf{x}_2 = (0\ \ 0\ \ 1)^T;\ \mathbf{x}_3 = (0\ \ \omega\ \ 1)^T \quad (8)$$

where ω corresponds to the so called self-loops, i.e. when for $p \in P,\ t \in T$, $\{(p, t) \in B \Rightarrow (t, p) \in B\}$. Here the set $B = F \cup G$ corresponds to the matrix $\mathbf{B}$.

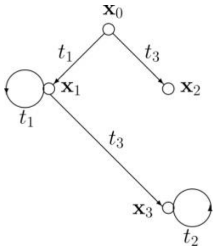


Figure 5
The corresponding RT

In our case the loops are represented by means of the pairs (p_1, t_1) and (p_3, t_2). The analysis of RT, more precisely the coverability graph (CG) in this case, is much more complicated than the previous analysis of RT in Figure 3. Namely, the *deploying* of such CG to the labeled reachability graph (LbRG) leads to the infinite chain given in Figure 6. On the basis of such RT we can analyze the LbPN given on Figure 4 by means of the alphabet $\mathcal{L}$. When all events are observable (i.e. when no ε occurs in $\mathcal{L}$) the situation is relatively simple.

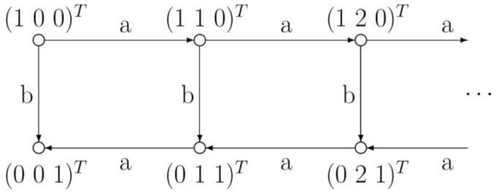


Figure 6
The CG *deployed* to LbRG

When we start from the initial state $\mathbf{x}_0 = (1\quad 0\quad 0)^T$ by firing t_3 the system development will be finished in the state $\mathbf{x}_2 = (0\quad 0\quad 1)^T$. Suppose that $\mathbf{x}_2$ is the desired final state (i.e. $\mathbf{x}_2 \in F_x \subset \mathcal{R}$). The final state can be reached from the initial, by the following firing sequences: {b}, {a, b, a}, {a, a, b, a, a}, {a a a b a a a}, etc. It means that formally written L = {a$^{\mathrm{m}}$ b a$^{\mathrm{m}}$ | m ≥ 0}. When an unobservable event ε occurs in the alphabet $\mathcal{L}$, i.e. when an unobservable transition occurs in the PN, the situation changes. During control synthesis it is necessary to deal with them.

2.2 Interpreted Petri Nets

In case of IPN the symbol N also means the P/T PN. Besides N there are the following additions here - the input alphabet $\Sigma = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ with the input symbols α_i, $i = 1, \ldots, r$; the output alphabet $\Phi = \{\delta_1, \delta_2, \ldots, \delta_s\}$ with the output symbols δ_j, $j = 1, \ldots, s$; the labeling function of transitions $\lambda = T \longrightarrow \Sigma \cup \{\varepsilon\}$ which assigns to each transition either $\alpha_i \in \Sigma$ or an internal event. Here, a constraint asking that a unique label is assigned to each transition, i.e. $\forall\, t_j, t_k \in T$, $j \neq k$ if $\forall\, p_i$ $\mathbf{F}(p_i, t_j) = \mathbf{F}(p_i, t_k) \neq 0$ and both $\lambda(t_j) \neq \varepsilon$, $\lambda(t_k) \neq \varepsilon$ then $\lambda(t_j) \neq \lambda(t_k)$ has to be valid; the labeling function of places $\Psi = P \longrightarrow \Phi \cup \{\varepsilon\}$ which assigns to each place either output symbol $\delta_i \in \Phi$ or ε being a null output signal; the output function $\mathbf{C} : \mathcal{R}$ (IPN, $\mathbf{x}_0$) $\longrightarrow \mathbb{Z}_{\geq 0}^{q\times n}$ (here, $\mathbb{Z}$ is the set of integers) with $q \in \mathbb{Z}_{>0}$ being the positive integer expressing the number of available output signals and $n = |P|$ being the number of places, is the $(q \times n)$-dimensional matrix of integers assigning the output vector to each marking $\mathcal{R}$ (IPN, $\mathbf{x}_0$). Thus, the $(q \times n)$-dimensional matrix $\mathbf{C}$ represents a relation between output vectors and state vectors of IPN. IPN can be formally expressed by means of the following sextuple

$$< \mathrm{N}, \Sigma, \Phi, \lambda, \Psi, \mathbf{C} > \tag{9}$$

Because the definition may seem too complicated, an illustrative example should be introduced for explanation.

2.2.1 Illustrative Example

In Figure 7 the simple example of IPN is displayed. There are two uncontrollable/unobservable places t_3, t_6 and two unmeasurable /unobservable places p$_3$ and p$_5$. Here, the input alphabet $\boldsymbol{\Sigma} = \{\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_4}\}$, correspond to controllable transitions $\{t_1, t_2, t_4, t_5\}$ assigned by the function $\boldsymbol{\lambda}$. By the same function the symbol $\boldsymbol{\varepsilon}$ is assigned to the uncontrollable/unobservable transitions $\{t_3, t_6\}$. The measurable places are $\{p_1, p_2, p_4, p_6\}$. They correspond with the output alphabet $\boldsymbol{\Phi} = \{\boldsymbol{\delta_1}, \ldots, \boldsymbol{\delta_4}\}$. The symbol $\boldsymbol{\varepsilon}$ is assigned to the unmeasurable/unobservable places $\{p_3, p_5\}$. They yield the null output signal. The places $\{p_7, p_8\}$ have the fixed marking because of the self-loops. Therefore, the following output equation

$$\mathbf{y}_k = \mathbf{C}.\, \mathbf{x}_k\ ,\ k = 0, 1, \ldots, K \tag{10}$$

extends the model (1), (2). In case of this illustrative IPN it means that

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{11}$$

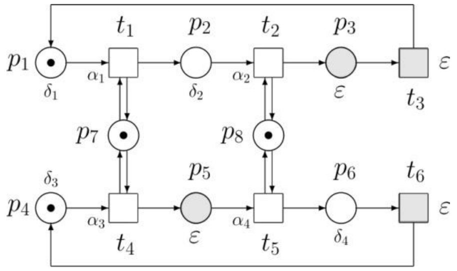


Figure 7
The IPN with two uncontrollable transitions and two unmeasurable/unobservable places

3 Interpreted and Labeled Petri Nets in DES Control

As it was mentioned above, PN are used for modeling, analyzing and control of DES. PN models of real systems may comprehend uncontrollable and/or unobservable transitions and unmeasurable/unobservable places. Therefore, the set of transitions T consists of two subsets $T = T_c \cup T_u$, where T_c includes all controllable transitions and T_u involves all uncontrollable transitions. Some of controllable transitions detecting their firing are named as the *transition sensors*. The set of places P also consists of two subsets $P = P_m \cup P_{um}$, where P_m includes all measurable places and P_{um} involves all unmeasurable/unobservable places. Some places from P_m may represent the so called *place sensors* - see e.g. $\{p_1, p_2, p_4, p_6\}$ in Figure 7. Such places creates the $(q \times n)$-dimensional output vector $\mathbf{y}_k$ in (10).

3.1 The IPN View on the Problem of Control

From the IPN point of view it is possible to utilize a general view. Consider the following simple introduction into the principle of the IPN based control.

3.1.1 The Basic Principle of the IPN-based Control

Consider a controlled segment of a plant. Its IPN model is displayed in Figure 8. There is (i) a control specification represented by the subnet $\{p_4; t_3\}$and; (ii) the model of a segment of the plant (to be controlled) represented by the subnet $\{p_1, p_2, p_3; t_1, t_2\}$. The place p_2 is unmeasurable/unobservable and t_2 is uncontrollable/unobservable.

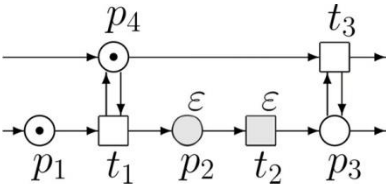


Figure 8
The principle of IPN-based control

Here, on the one hand, the controllable transition t_1 representing a discrete event in the plant is enabled by means of the self-loop with the *sensor* place p_4 of the control specification. On the other hand, the measurable place (*sensor*) p_3 of the plant becomes active by means of the self-loop with t_3. In such a way the uncontrollable transition t_2 (being an internal spontaneous discrete event) and the unmeasurable/unobservable place p_2 are bypassed. Thus, the transition t_2 (labeled practically by ε) can spontaneously fire or not. No interference from outside to t_2 is possible. In spite of this, the situation (i.e. the system dynamics development) is under the control. The RT of the IPN has only one branch - $\mathbf{x}_0 \overset{u_{t_1}}{\longrightarrow} \mathbf{x}_1 \overset{u_{t_2}}{\longrightarrow} \mathbf{x}_2 \overset{u_{t_3}}{\longrightarrow} \mathbf{x}_3$. In spite of this the control system is not able to force transition t_2 directly. However, it is pent inside the control loop and it cannot influence another parts of plant. Of course, the control specifications have to respect the prescribed manufacturing system technology (i.e. the technological process).

The best way how to present the approaches to the real DES control are examples.

3.2 Example 1

Consider the example of the real DES given in Figure 9 being a complex FMS. Here, activities of two automatically guided vehicles AGV 1, AGV 2 are modeled. The activity of AGV 2 is modeled by means of the upper PN subnet $\{p_1, p_2, p_3, p_4; t_2, t_3, t_4, t_5\}$ and the activity of AGV 1 is modeled by means of the lower PN subnet $\{p_6, p_7, p_8, p_9; t_6, t_7, t_8\}$. AGVs cooperate in FMS in such a way that: (i) the role of AGV 2 is to carry two different parts A, B from two input conveyors (each of them feeds the corresponding kind of parts) into a transship center Transfer; (ii) the role of AGV 1 is to carry these products to a robot R which put the parts (by means of its gripper) on a pressure plate of a compactor machine where the parts are pressed down altogether into a final product C (i.e. C = A + B). The PN model of the FMS to be controlled is displayed in Figure 10.

The attendance of both vehicles in marginal points of their movements is checked by sensors indicating these locations on both sides of their runways - in case of AGV 2 at the conveyors (S2b, S2c) as well as at the transship center (S2a) and in case of AGV 1 at the transship center (S1b) as well as at the robot (S1a).

The place p_1 represents AGV2 being right, while p_4 represents AGV 2 being left. The place p_5 represents AGV1 being left, while p_7 indicates AGV 1 being right.

The place p_6 denotes activities of the robot gripper. Unfortunately, there are unmeasurable/unobservable places $\{p_2, p_3, p_8, p_9\}$ in the PN model. They represent states of the unmeasurable/unobservable movements of AGVs on their runways. The place p_9 ensures repeating the working cycle.

The transition t_1 represents a start of the whole process. The sensors S2b, S2c correspond, respectively, to t_2, t_3. Sensors S2ab, S2ac correspond, respectively, to transitions t_4, t_5 depending on the fact whether AGV 2 arrives from S2b or from S2c. The transition t_6 presents the sensor S1a, t_7 denotes the sensor Spl of placing parts on the pressure plate, t_8 denotes the sensor S1b. Finally, t_9 is getting to be practically uncontrollable because of unobservable places p_8, p_9 - it has only deterministic information that AGV 2 is left while information from p_8, p_9 being important for its firing is nondeterministic.

Moreover, while AGV 1 has an ambiguity consisting in p_1 (to fire either t_2 before t_3 or contrariwise) in AGV 2 no such ambiguity occurs. These facts have to be taken into account too.

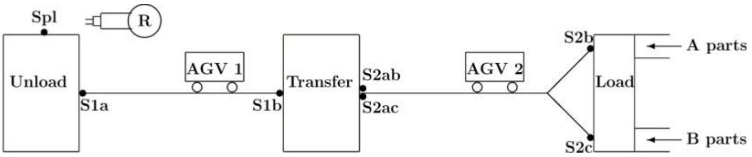


Figure 9

The rough scheme of the real FMS

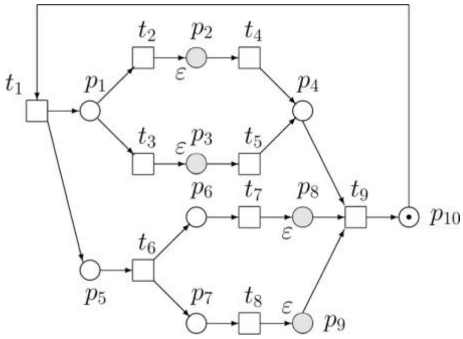


Figure 10

The PN model of the uncontrolled FMS

The RT of the PN model of such a system with controllable and observable transitions and measurable places is displayed in Figure 11.

Of course, building the IPN based model of DES itself is not sufficient. To ensure a desirable behavior of DES, it is necessary to synthesize the control specification able to deal with the non-determinism without affecting the technological process.

Consider that the system behavior under the control should be as direct as possible, it means without any unnecessary turns, but in keeping within the technological process, of course. In such a case, the proposed controller corresponding to demands is added as it is apparent in Figure 12. The system behavior corresponds to the RT displayed in Figure 13.

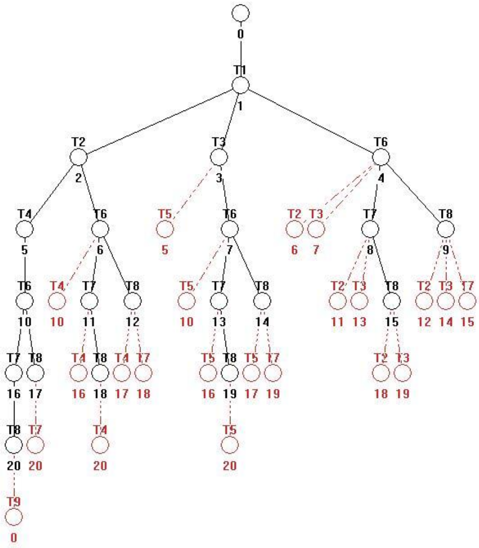


Figure 11

The RT corresponding to the PN model in Figure 10

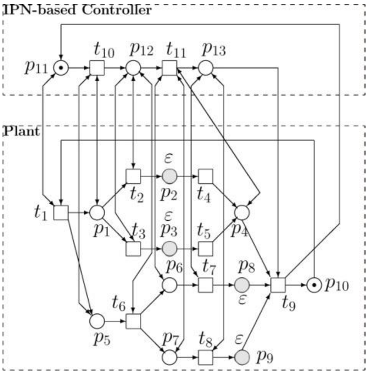


Figure 12

The IPN-based model of the controlled plant

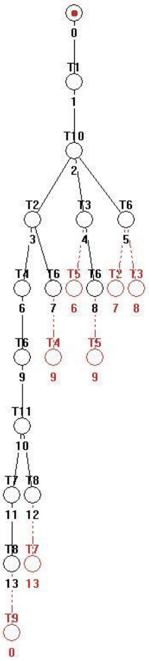


Figure 13

The RT corresponding to the IPN-based model of the controlled plant

3.3 Example 2

Let us introduce the example of DES where the situation is not so light. To illustrate problems, consider the real example of FMS consisting of two production lines L1, L2 producing mutually different parts P1, P2. The parts are assembled together into a final configuration P1 + P2. The robot R machines parts in both lines. The line L1 is modeled by the subnet $\{p_1, p_2, p_3; t_1, t_2, t_3\}$, while the line L2 by the subnet $\{p_5, p_6, p_7; t_4, t_5, t_6\}$. The activities of the robot R are modeled by the subnet $\{p_9, p_{10}, t_7, t_8\}$, while those of the assembly process by the subnet $\{p_4, p_8, t_9\}$. Of course, the subnets are interconnected insomuch that it is impossible to strictly distinguish the confines among subnets. The PN model of the FMS is given in Figure 14. As we can see, firing t_8 is uncertain, because the transition t_8 is uncontrollable (it has the label ε). We do not know whether it, being enabled, will be fired in an actual state k of the dynamics development or not. If not, t_8 as if did not exist (u_{t_8}= 0) and the PN structure has changed - the branch $p_{10} \overset{u_{t_8}}{\longrightarrow} p_9$ is as if dead, impassable, opaque. The RT of such PN structure is different from RT of the full structure together with t_8. In comparison with the RT of the full structure it is cropped - all branches labeled by t_8 are missing. Even, it contains a deadlock after the firing sequence t_7, t_4, t_5, t_6 starting from the initial state $\mathbf{x}_0$. Unfortunately, none of the trees can be introduced because of their size.

Non-determinism consists in the displeasing fact that we do not know whether the enabled transition t_8 will be spontaneously fired - i.e. whether the process of the dynamics development will correctly continue, or not.

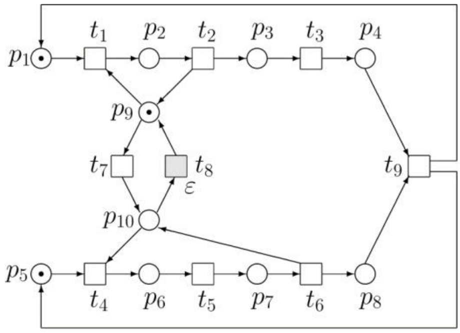


Figure 14
The P/T PN being the model of the FMS

3.3.1 Using IPN-based Control

Let us apply the IPN-based control of the plant given in Figure 14. Consider the same criterion for control as in the previous Example 1 (as direct as possible behavior of DES). Then, the PN model of the controlled plant is given in Figure 15. RT appurtenant to this model is displayed in Figure 17 up. The picture is rotated left (counter clockwise) in order to save space. However, such IPN-based controller needs not be alone. Other schemata can exist too. For example, a counter of production cycles can be added as it can be seen in Figure 16. It is represented by the place p_{15}. RT of such model is given in Figure 17 down. The picture is rotated left too because of saving space.

As we can see, the transition t_8 cannot be bypassed and it induces undesirable cycles in corresponding reachability trees - $\{\mathbf{x}_4, t_8, \mathbf{x}_3, t_7, \mathbf{x}_4\}$ in Figure 17 (up) and $\{\mathbf{x}_{15}, t_8, \mathbf{x}_{14}, t_7, \mathbf{x}_{15}\}$ in Figure 17 (down). It is necessary to say that there is no road to the satisfying solution in this case.

Now, let us investigate a situation when different transitions are uncontrollable or unobservable. Consider the structure of the IPN model given in Figure 18 with the uncontrollable transitions t_2 and t_6. In principle it is intrinsic e.g. in this case of DES when the upper branch $\{p_1, t_1, p_2, t_2, p_3, t_3, p_4\}$ with the fix resource p_9 (expressing the presence of a robot) represents a production line and the lower branch $\{p_5, t_4, p_6, t_5, p_7, t_6, p_8\}$ with the fix resource p_{10} (expressing the presence of the same robot) represents another production line. Both lines are served by the same robot. While p_2 models an activity of a machine in the upper line, $\{p_6, t_5, p_7\}$ models an activity of another machine in the lower line. There the places p_{13}, p_{14} are added to the IPN-based controller in order to define desired priorities of firing transitions - i.e. $t_1 > t_7$ and $t_8 > t_9$.

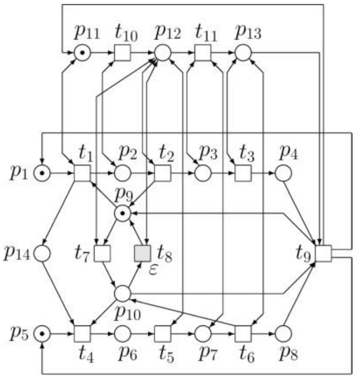


Figure 15

The scheme of the IPN-based control

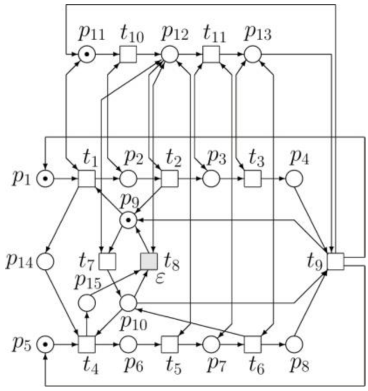


Figure 16

The scheme of the IPN-based control with the counter

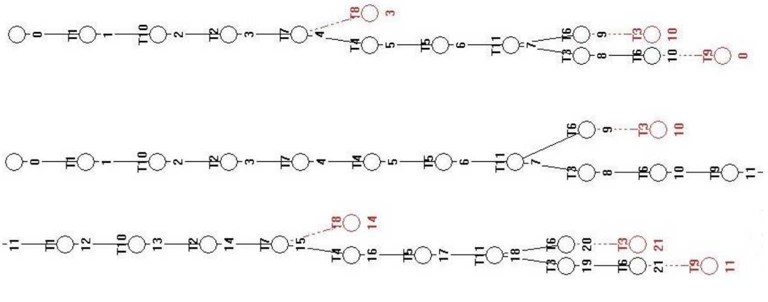


Figure 17

RT of the model in Figure 15 (up) and RT (2 parts) of the model in Figure 16 (down)

The global scheme of IPN-based control reflects a real situation, because finishing of the machining processes modeled by the PN subnets $\{t_1, p_2, t_2, p_9, t_1\}$ and $\{t_4, p_6, t_5, p_7, t_6, p_{10}, t_4\}$ is uncontrollable/unobservable because of the places p_2 and

$\{p_6, p_7\}$ being practically unmeasurable/unobservable during machining. The RT of such controlled plant is given in Figure 19. As we can see, no cycles containing t_7 and t_8 exist there.

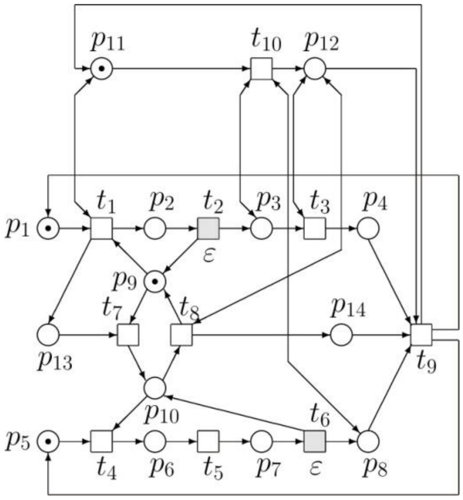


Figure 18

The IPN-based control of the system with uncontrollable transitions t_2, t_6

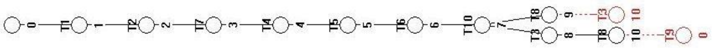


Figure 19

The RT of the system with uncontrollable transitions t_2, t_6

3.1 The LbPN View on the Problem of Control

Consider the LbPN model of a kind of DES, representing the cooperation of two jobs by means of resources p_5, p_6 needed for performing both jobs, given in Figure 20. Corresponding RT (provided that all transitions can be understood to be fired when they are enabled) is displayed in Figure 21.

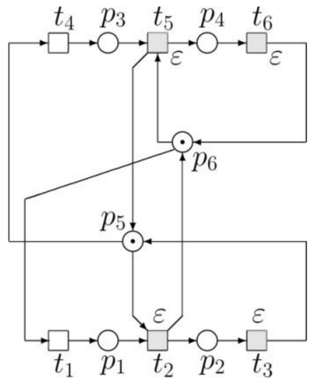


Figure 20

The LbPN model of DES

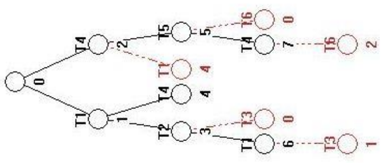


Figure 21

The corresponding RT

The resources must be available to satisfy the needs of both jobs. It means that the components x_{p_5}, x_{p_6} of any state vector $\mathbf{x}_k$, k = 0, 1, …, have to satisfy the following condition

$$x_{p_5}+x_{p_6} \leq 2 \qquad (12)$$

i.e. $\mathbf{L}.\mathbf{x}_k \leq 2$, where $\mathbf{L}$=(0 0 0 0 1 1) and $\mathbf{x}_k$ is an arbitrary reachable state. Then, with respect to the controller synthesis [18], the controller structure $\mathbf{B}_c$ = -$\mathbf{L}.\mathbf{B}$. Hence, $\mathbf{B}_c$ = (-1 0 1 -1 0 1). Because $\mathbf{B}_c = \mathbf{G}_c^T\text{-}\mathbf{F}_c$, where $\mathbf{G}_c^T$= (0 0 1 0 0 1) and $\mathbf{F}_c$= (1 0 0 1 0 0), the controller is represented by the place p_c together with its interconnections with the uncontrolled model - i.e. directed arcs from p_c to transitions of the uncontrolled model by means of $\mathbf{F}_c$, as well as from transitions of the uncontrolled model to p_c by means of $\mathbf{G}_c^T$. Consequently, the controlled LbPN model is given in Figure 22 and its RT is displayed in Figure 23.

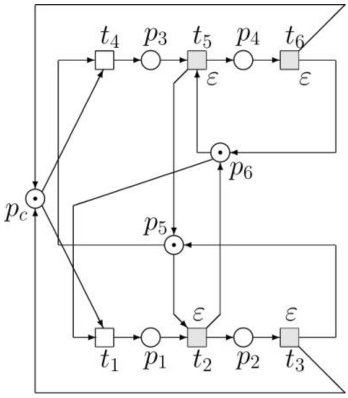


Figure 22
The LbPN model of DES

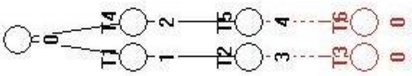


Figure 23
The RT of the LbPN model of DES

As we can see, the controller is able (i) to react on unobservable transitions t_3, t_6 when they are fired; (ii) to affect only the observable transitions t_1, t_4; (iii) to deal with the unobservable transitions t_2, t_5 when they are fired. Of course, the unobservable transitions cannot be either avoided or their influence eliminated. Moreover, both jobs are relatively autonomous - i.e. (i) $\mathbf{x}_0 \xrightarrow{t_1} \mathbf{x}_1 \xrightarrow{t_2(\varepsilon)} \mathbf{x}_3 \xrightarrow{t_3(\varepsilon)} \mathbf{x}_0$ and (ii) $\mathbf{x}_0 \xrightarrow{t_4} \mathbf{x}_2 \xrightarrow{t_5(\varepsilon)} \mathbf{x}_4 \xrightarrow{t_6(\varepsilon)} \mathbf{x}_0$.

3.2 A Comment on Computational Complexity

The computational complexity at using IPN and LbPN is nothing anomalous in comparison with P/T PN. Namely, the marking evolution of IPN and LbPN is the same like that of P/T PN and leads to computation of RT which may be a main source of problems concerning the computational complexity. Moreover, the structure of IPN and LbPN is prepared before the RT computation. The computational complexity of P/T PN is investigated in several papers concerning theoretical informatics - see e.g. [19]-[21]. For LbPN and IPN the total number of possible firing vectors is upper bounded by a polynomial function in k, i.e. $O(k^b)$, where k is the length of the sequence of labels and b is a parameter depending on the structure and the labeling function of the net. Namely $b = c(d - 1)$, where c is the number of nondeterministic labels and d is the maximum number of transitions corresponding to a label in the net.

Conclusions

This paper describes the possibility of how DES, with non-determinism, can be modeled and controlled by means of special kinds of PN, namely, IPN and LbPN. Such nets were defined and illustrative examples were introduced, in order to demonstrate their structure and dynamic behavior. The kernel of the paper is devoted, especially to the control synthesis of actual DES modeled, by means of these kinds of PN. It was shown, thru examples, that, in spite of non-determinism, real DES, in practice, can be controlled, albeit, with some limitations or restrictions.

Acknowledgement

This work was supported by the Slovak Grant Agency for Science VEGA under Grant No. 2/0029/17.

References

[1] C. A. Petri: Communication with Automata. Ph.D. Thesis. Technical University of Darmstadt, 1962, 128 pages (in German)

[2] J. L. Peterson: Petri Net Theory and the Modeling of Systems. Englewood Cliffs, NJ: Prentice-Hall, 1981

[3] T. Murata: Petri Nets: Properties, Analysis and Applications, Proceedings of the IEEE, Vol. 77, No. 4, 1989, pp. 541-580

[4] J. Desel, W. Reisig: Place/Transition Petri Nets. In: W. Reisig, G. Rozenberg (Eds.): Advances of Petri Nets, Lecture Notes in Computer Science, Vol. 1491, Springer, Heidelberg, 1998, pp. 122-173

[5] M. Silva: Half a century after Carl Adam Petri's Ph.D. thesis: A perspective on the field, Annual Reviews in Control, Vol. 37, 2013, pp. 191-219

[6] M. Silva: On the history of Discrete Event Systems, Annual Reviews in Control, Vol. 45, 2018, pp. 213-222

[7] A. Giua, M. Silva: Modeling, Analysis and Control of Discrete Event Systems: A Petri Net Perspective, In: 20th IFAC World Congress, Toulouse, France, IFAC PapersOnLine 50-1, 2017, pp. 1772-1783

[8] Z. Ma, Y. Tong, Z. Li, A. Giua: Marking Estimation in Labelled Petri Nets by the Representative Marking Graph, In: 20th IFAC World Congress, Toulouse, France, IFAC PapersOnLine 50-1, 2017, pp. 11175-11181

[9] M. P. Cabasino, C. N. Hadjicostis, C. Seatzu: Marking Observer in Labeled Petri Nets with Application to Supervisory Control, IEEE Transactions on Automatic Control, Vol. 62, No. 4, April 2017, pp. 1813-1824

[10] M. P. Cabasino, A. Giua, M. Pocci, C. Seatzu: Discrete Event Diagnosis Using Labeled Petri Nets. An Application to Manufacturing Systems, Control Engineering Practice, Vol. 19, 2011, pp. 989-1001

[11] Z. Ma, Z. He, Z. Li, A. Giua: Design of Monitor-based Supervisors in Labelled Petri Nets, In: 14th IFAC Workshop on Discrete Event Systems WODES, Sorrento, Italy, IFAC PapersOnLine 51-7, 2018, pp. 374-380

[12] M. P. Cabasino, C. N. Hadjicostis, C. Seatzu: State Feedback Control of Labeled Petri Nets with Uncertainty in the Initial Marking, In: 19th International Conference on Emerging Technologies and Factory Automation ETFA, Barcelona, Spain, 2014, pp. 1-7

[13] A. Ramírez-Treviňo, I. Rivera-Rangel, E. López-Mellado: Observability of Discrete Event Systems Modeled by Interpreted Petri nets, IEEE Trans. on Robotics and Automation, Vol. 19, No. 4, 2003, pp. 557-565

[14] A. Giua: Analysis and Control of Petri Nets, In: DES School, Sorrento, Italy. Available in: http://schooldes2018.unisa.it/Giua_18wodes_school.pdf

[15] J. O. Moody, P. J. Antsaklis: Petri Net Supervisors for DES with Uncontrollable and Unobservable Transitions, IEEE Transactions on Automatic Control, Vol. 45, No. 3, March 2000, pp. 462-476

[16] Z. Achour, N. Rezg, X. Xie: Supervisory Controller of Petri Nets under Partial Observation, In: 7th International Workshop on Discrete Event Systems WODES, Reims, France, Sept. 22-24, 2004, IFAC Proceedings Vol. 37, No. 18, 2004, pp. 51-56

[17] A. Boussif, M. Ghazel, K. Klai: DPN-SOG: A Software Tool for Fault Diagnosis of Labeled Petri Nets Using the Semi-Symbolic Diagnoser, HAL Id: hal-01653191, Available on: https://hal.archives-ouvertes.fr/hal-01653191, 2017, 14 pages

[18] F. Čapkovič: Timed and Hybrid Petri Nets at Solving Problems of Computational Intelligence. Computing and Informatics, Vol. 34, No. 4, 2015, pp. 746-778

[19] J. Esparza: Decidability and Complexity of Petri Net Problems - An Introduction. In: Reisig, W., Rozenberg, G. (Eds.): Lectures on Petri Nets I: Basic Models, Advances in Petri Nets, Springer, 1998, pp. 374-428

[20] H.-C. Yen: Introduction to Petri Net Theory. In: Recent Advances in Formal Languages and Applications. Book series Studies in Computational Intelligence (SCI), Vol. 25, Springer, 2006, pp. 343-373

[21] L. Li, C. Hadjicostis: Minimum Initial Marking Estimation in Labeled Petri Nets. In: Proceedings of the American Control Conference - ACC'2009, St. Louis, USA, paper FrB12.2, 2009, pp. 5000-5005

Counting the Number of Shortest Chamfer Paths in the Square Grid

Laith Alzboon, Bashar Khassawneh, Benedek Nagy

Department of Mathematics, Eastern Mediterranean University, Famagusta,
North Cyprus, via Mersin-10, Turkey, Laith.Khalaf@emu.edu.tr,
bashar.khassawneh@emu.edu.tr, benedek.nagy@emu.edu.tr

Abstract: In this paper, the number of shortest paths between any point pairs of the square grid, based on weighted distances, is computed. We use two types of steps on the gridlines and diagonal steps. Consequently, we use an 8-adjacency square grid, that is, one where a first weight is associated with the horizontal and vertical movements, while a second weight (not necessarily different from the first) is assigned to the diagonal steps. The chamfer distance of two points depends on the numbers and weights of the steps in a 'shortest path'. In special cases, the cityblock and the chessboard distances, the two most basic and widely used digital distances (they are also referred as L_1 and L_∞ distances, respectively) of the two-dimensional digital space occur. Although our combinatorial result is theoretical, it is closely connected to applications, such as communication networks, path counting in digital images, traces and trajectories in 2D digital grids. We consider all seven cases with non-negative weights and also the case when negative weights are allowed.

Keywords: traces; trajectories; weighted distances; shortest paths; digital distances; combinatorics

1 Introduction

In this paper we are interested in answering a combinatorial question about the number of cheapest solutions of a problem. More precisely, a solution is considered as a path between two grid points. The number of cheapest solutions is, then, the number of minimal weighted paths between the points.

Digital grids and their applications in various fields play important roles in science and technology. Digital grids are used in applications such as image processing [11], computer graphics, communication networks, crystallography and physical simulations. The space, in this case, the considered grid, is discrete, so theoretic tools from discrete mathematics, graph theory, combinatorics and, especially, from digital geometry can be used. In most cases only coordinates with integer values are used to address points (nodes). The square grid (also called rectangular grid) is the most usual digital grid, as it is the most frequently applied grid in two dimensions (2D). It is essential in image processing, cellular automata and other fields of applied information technology as well as 2D physical simulations. The Cartesian coordinate frame describes it, and, consequently, two well-known neighbourhood relations are defined based on those coordinate values. One of the benefits of working on the square grid is that it is self-dual: the square grid can be seen by connecting (the midpoints of) neighbour pixels to each other, and also by using the original grid, with points where the grid lines cross. In discrete grids, neighbour relation is of high importance. Opposite to discrete space, Euclidean space is continuous space, and there is no neighbour relation. The natural, Euclidean distance has several well-known and beneficial properties. However, using Euclidean distance on a discrete space may not be the best option. For example, one topological paradox is that the grid points having an exact Euclidean distance of seven from the origin do not really form a circle in any usual sense; the determined four pixels are not even connected. When working with computers, one may prefer digital distances, i.e., distances based on paths through neighbour points. In most cases, these distances have integer values, and it is easy to work with them. The shortest paths between any two points are computed depending on the types of the steps of the paths: they could be horizontal, vertical or diagonal between any two adjacent points. In this context, movements to 4- and 8-neighbourhoods are defined. In discrete spaces, especially in the square grid, digital distances, have been proposed in [18]. Two basic distance functions are defined based on the two neighbour relations. The cityblock (also called the Manhattan taxicab) distance and the chessboard distance are related to the 4- and 8-neighbourhood movements, respectively. Since these two digital distances give very rough approximations of the Euclidean distance, it was recommended to use them alternating along a path (the obtained distance is called octagonal distance). From the end of 1980's, extending this idea more formally, digital distances based on predefined neighbourhood sequences have been introduced and used in which both chessboard and cityblock neighbourhoods are combined in a sequence that can be periodic [19, 4, 5] or non-periodic [14]. Distances based on neighbourhood sequences on other grids have also been defined, see, e.g., [13]. Other digital distances, the weighted distances give another way to have distances on a grid with integer values [2]. They are also called chamfer distances. Given the associated weights for the used neighbourhoods, the chamfer distance between p and q relative to these weights is the length (i.e. the sum of the weights) of the shortest digital arc (path) from p to q with respect to the weights of movements. One of the advantages of weighted distances is that one can approximate the Euclidean distance by them in a smooth way (especially allowing larger steps and more weights). Another reason to prefer weighted distances versus distances based on neighbourhood sequences is that the former ones are always metric, while there are plenty of neighbourhood sequences that do not provide metrics [14] (since the triangular inequality may fail).

Counting the number of shortest paths is a theoretical, combinatorial problem which has various connections for applied fields. The number of the cheapest solutions may give significant information about the problem to be solved. In networks, communication-related fields, computer simulations and in theory of algorithms some concepts and algorithms from graph theory play important roles. In communication networks the transmitters, receivers, etc. can be represented by nodes of a graph, and their connections, the possible ways of communication, can be shown by the edges. Concepts such as paths, shortest paths and distances in these graphs are understood and give some important features of the communication network. The number of shortest paths also has importance in applications for transmitting messages over networks since they refer to the width of the connection channel between the given points. Any shortest path can be useful and used to increase the performance, in this case, the amount of information transmitted during a unit of time (the width of the network), speeding up communication [3]. Opposite to the Euclidean space, the shortest path is usually not unique over discrete spaces (i.e., there is no unique cheapest solution). These networks are usually artificial, meaning that graphs with special properties can be considered, such as the square grid. In social networks, the various graph measures, such as eccentricity, are defined based on the number of some shortest paths [8]. In some physical simulations connected to random walks, percolations, trajectories and traces [9, 10, 17], it is also important to count the number of shortest paths. Several related applications have already been detailed in [12]. Path counting in discrete spaces is closely related to graph theory. It is well-known that the number of paths with a given length can be computed by the appropriate powers of the adjacency matrix of the graph. Path counting (for cityblock and chessboard distances) in digital images (i.e., finite subgraphs of the square grid) is used to infer properties of images [18]. It was also considered in [7] based on matrix multiplication with various neighbourhood relations. In [6], the numbers of shortest paths are computed for the two above mentioned basic distances and also to the octagonal distance, which is a special neighborhood sequence-based distance. For neighbourhood sequences in general, the problem was considered in [15]. In this paper, a similar combinatorial problem, the path counting for weighted distances considering the basic two types of steps is presented with enumerative combinatorial calculation. In most cases, we assume that both weights are non-negative. Moreover, we solve all the cases of the problem by providing the solutions by closed formulae. As we will see that there are five entirely different cases based on (the relation of) the used weights if both weights are positive; and there are two cases with 0 weights. Two of the cases, actually, provide the same result as the corresponding results for cityblock and chessboard distances, however, our proof technique is different from the technique used in [6]. We also present 3D charts to show how the number of shortest paths grows when the distance grows. Thus, the significance of the paper is not only to consider and summarize all the possible cases, but also to give solutions for cases which were not analyzed before, e.g., the last three cases shown in this paper. Few of the cases presented in this paper were already presented in [1]. Those results are complemented here with various results of non-traditional weight settings, e.g., allowing the diagonal steps to have less weight than the cityblock steps, or to have zero or negative weight.

2 Preliminaries

Now we recall some definitions and concepts that are necessary to understand the results of this paper. As we briefly mentioned before, there are two popular types of neighbourhood relations in the square grid: the cityblock and the chessboard neighbourhood. The cityblock's four neighbours of a point of the square grid are defined as the edge-adjacent points: two horizontal and two vertical neighbours (see Figure 1, left). The chessboard neighbourhood of a point, in addition to the previously given four points, also contains the four closest points in diagonal directions (see Figure 1, middle). These neighbourhood relations are also called von Neumann and Moore neighbours, respectively, in cellular automata theory.

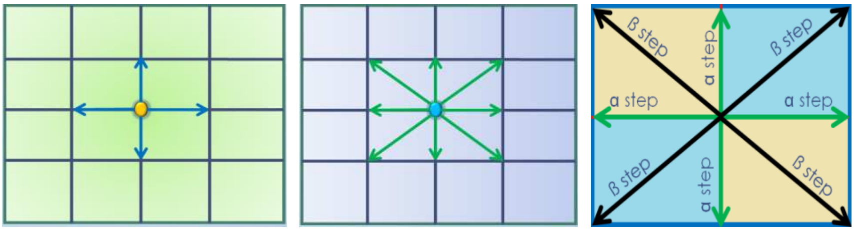


Figure 1

Cityblock neighbours (left, four neighbours for each point). Chessboard neighbours (middle, eight neighbours for each point). Weighted steps for chamfer distance (right, α steps, β steps) from the point in the centre.

Weighted distance (or chamfer distance per the original terminology [2]) is used in the square grid to get a better approximation to the Euclidean distance in two-dimensional digital space than the distances based on the sole use of the chessboard or cityblock neighbourhood. According to the movements from a point to one of its neighbours, we associate a weight for each type of movement: we give weight α for cityblock movements and weight β for each diagonal movement, as shown in Figure 1 (right). Formally, we can describe it as follows:

Let $p = (x_1,y_1)$ and $q = (x_2,y_2)$ be two points in the square grid; let $W = (w_1,w_2)$ be the absolute difference vector between the points, where $w_1 = |x_1 - x_2|$ and $w_2 = |y_1 - y_2|$. Then, as was previously computed [18], the cityblock distance (Manhattan taxicab distance) can be computed as $d(p,q) = w_1 + w_2$ and the chessboard distance can be computed as $d(p,q) = max\{w_1,w_2\}$.

The number of diagonal steps in a shortest path of the chessboard distance is $min\{w_1,w_2\}$, and the number of cityblock steps (i.e. the number of vertical or horizontal movements) in a shortest path with the chessboard distance is $max\{w_1,w_2\} - min\{w_1,w_2\}$. These values become important when calculating the number of shortest paths of chamfer distances.

When using both types of neighbours, but with different weights, in order to calculate the length of a shortest path (i.e. the chamfer distance between p and q), we must find how many α and β steps are in the given path. Their numbers in a shortest path depend on the respective values of α and β as well. According to their numbers and values, we will compute not only the length but also the number of shortest paths between any two points. The actual computation depends on the used weights. In this paper, as usual, both in graph theory and in digital geometry, we assume that both α and β are non-negative, and actually, in the first some cases we assume that they are positive. In some applications, there is also an assumption that $0 < \alpha \leq \beta$. We show Example 1 below, for this case. With this condition, subcases are defined by the relation of 2α and β. However, in this paper, we do not restrict our studies to these cases. We will also do computations when $0 < \beta < \alpha$ (see Subsection 3.5), and as we will see this case is the most interesting among all. For the sake of completeness, we also present the cases, when one or both the weights have value 0.

Now, as an example, we show how to compute the distance, or the length of the shortest path if $0 < \alpha \leq \beta$ holds. Let N be the number of α steps and M be the number of β steps in a shortest path; then, the weighted distance between p and q is $d_w(p,q) =N\,\alpha + M\,\beta.$

Example 1. Let $p = (5,6)$, $q = (7,1)$ and $\alpha = 3,\beta = 4$. Then $w_1 = 2$ and $w_2 = 5$. Thus, the cityblock distance of these points is 7, their chessboard distance is 5 (note that in these distances unit weight is used). Now, computing the chamfer distance, since $\alpha < \beta < 2\alpha$, it is worth using the path defined by the chessboard distance, i.e. with 2 diagonal and 3 cityblock steps: the chamfer distance equals to: $2\cdot4+3\cdot3=17$.

3 Results: Formulae for the Number of Shortest Paths

According to the values of α and β, we can compute weighted distances, and, consequently, we can compute the number of shortest paths. In this context, we have various cases depending on the respective ratio of the used weights, letting $f(w_1,w_2)$ be the function calculating the number of shortest paths between two points with an absolute difference vector (w_1,w_2). The cases are listed in the following subsections. The first two cases are equivalent to obvious discrete mathematical exercises (and have been proven also in [6] by a recursive method), and we explain them only for the sake of completeness using enumerative combinatorial techniques in our proofs. In the first five subsections, cases with positive weights are studied, while in the last subsection we deal with the cases when one or both weights is/are zero.

3.1 Case of $\beta > 2\alpha > 0$

Theorem 1. Let α and β be the weights for cityblock and diagonal movements, respectively, such that $\beta > 2\alpha > 0$. Let $p = (x_1,y_1)$ and $q = (x_2,y_2)$ be points of the square grid and $w_1 = |x_1 - x_2|$ and $w_2 = |y_1 - y_2|$ be the absolute differences between the corresponding coordinates of the points. Then, the number of shortest paths between p and q, denoted by $f(w_1,w_2)$, is given as $f(w_1,w_2)=\binom{w_1+w_2}{w_1}$.

Proof. We have $\beta > 2\alpha$ such that the weights are positive, which means that in the shortest path between p and q, no diagonal steps occur since diagonal steps can be substituted by two consecutive cityblock (i.e. a vertical and a horizontal) steps to produce a path with a smaller weight. Thus, all shortest paths contain only cityblock steps. The number of α steps between points p and q in the shortest weighted paths is computed in the same way as in the cityblock distance: w_1+w_2. The distance between p and q is α times more since each step has weight α. Moreover, in each shortest weighted path between p and q, the numbers of horizontal and vertical steps are w_1 and w_2, respectively. However, the order of these steps is arbitrary; thus, the number of shortest paths is given by the number of ways that we can arrange w_1 or w_2 steps among the total w_1+w_2 steps; it is given by the binomial coefficient $\binom{w_1+w_2}{w_1}$. Actually, $\binom{w_1+w_2}{w_1}$ and $\binom{w_1+w_2}{w_2}$ give the same value. ■

Example 2. The number of shortest paths between the points $p(5,12)$ and $q(8,13)$ with $\alpha = 3$ and $\beta = 7$ is computed as follows: $w_1 = |\,8 - 5| = 3$ and $w_2 = |\,13 - 12\,| = 1$, $w_1+w_2 = 4$. Thus, the result is $f(1,3) = \binom{4}{3} = \binom{4}{1} = 4$. See also Figure 2.

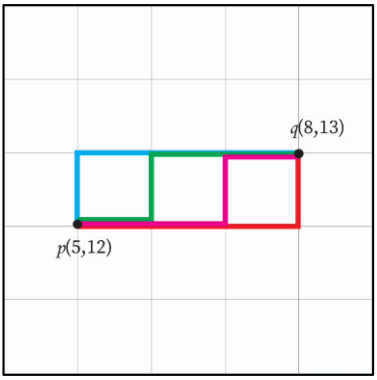


Figure 2

All shortest paths in case $\beta > 2\alpha > 0$; the four colors show the four shortest paths of Example 2

The paths of this case are also called grid-paths since only the edges of the grid are used. Since these results are exactly the binomial coefficients (almost as they form the Pascal's triangle), we do not give them in a table form, only sketch them in Figure 3 as a 3D chart for the number of shortest weighted paths from point (0,0) to all points in the represented region. In the figure, the origin is placed in the middle to show the symmetry distribution of the values. The formula grows rapidly in the corner directions when the coordinate differences are (almost) equal.

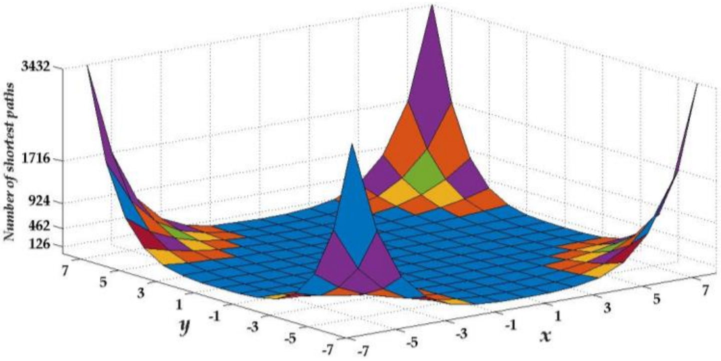


Figure 3
The number of shortest weighted paths from point (0,0) to other points in a 14 × 14 window with corners (−7, −7), (7, −7),(7,7) and (−7,7) in case $\beta > 2\alpha > 0$.

3.2 Case $0 < \alpha < \beta < 2\alpha$

Theorem 2. Let α and β be the weights for cityblock and diagonal movements, respectively, with the condition $\alpha < \beta < 2\alpha$. Further, let $p = (x_1,y_1)$ and $q = (x_2,y_2)$ be points, and let $w_1 = |x_1 - x_2|$ and $w_2 = |y_1 - y_2|$. Then, the number of the shortest paths between p and q is given as $f(w_1,w_2)=\binom{max\,\{w_1,w_2\}}{min\{w_1,w_2\}}$.

Proof. Both weights are positive and $\alpha < \beta < 2\alpha$. Thus, we may move in the shortest path from p to q using both α-steps and β-steps. We use β-steps as much as possible to get closer to the endpoint, which means that we will move diagonally by the minimum number of differences between the point coordinates, and the remaining steps are α-steps. According to this, the number of α-steps and β-steps in the shortest paths will be computed in the same way, as in a chessboard path from p to q (i.e. the number of steps is $max\{w_1,w_2\}$). Since the number of β-steps is $min\{w_1,w_2\}$, the number of α-steps is ($max\{w_1,w_2\} - min\{w_1,w_2\}$). The order of the steps is arbitrary; thus, the number of shortest weighted paths equals to the number of ways the β-steps can be arranged in the path with $max\{w_1,w_2\}$ steps altogether, which is exactly the binomial coefficient $\binom{max\,\{w_1,w_2\}}{min\{w_1,w_2\}}$. ■

Example 3. Let $p(-2,3)$, $q(2,0)$, and let α=3, β=4.Then w_1=4, w_2= 3 ,further $min\{w_1,w_2\}$ = 3 and $max\{w_1,w_2\}$= 4. Applying the formula for this case, the number of shortest paths from p to q is$\binom{max\,\{4,3\}}{min\,\{4,3\}}$ =$\binom{4}{3}$= 4. Actually these four shortest weighted paths are illustrated in Figure 4 with various colors.

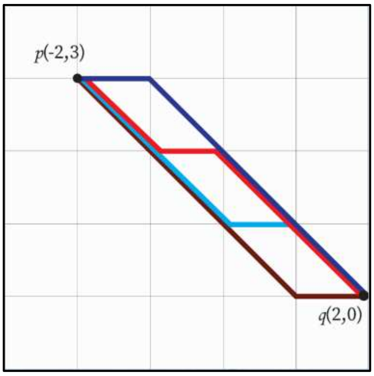


Figure 4
Example for all shortest paths of case 0 < α < β < 2α as in Example 3

Again, the values of Pascal's triangle appear, but in a different arrangement than in the previous case. Figure 5 gives a 3D chart for values for the number of shortest weighted paths from point (0,0) to all points in a 14 × 14 window. To show the symmetry of the distribution the origin is in the middle. This graph is already more interesting than the previous one, with more growing directions: the value grows fastest when one of the absolute coordinate differences is (approximately) half of the other one.

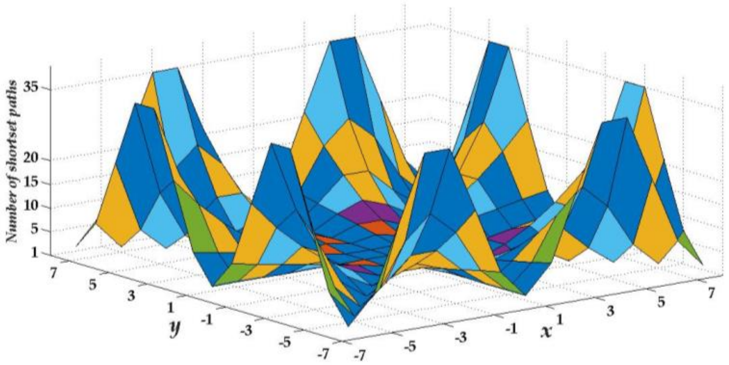


Figure 5
The number of shortest weighted paths in case 0 < α < β < 2α in a 14 × 14 window with corners (−7, −7), (7, −7),(7,7) and (−7,7). The minimums on the axes and on the diagonals can be seen.

3.3 Case of $\beta = 2\alpha > 0$

Theorem3. Let α and β be (positive) weights for cityblock and diagonal movements, respectively, such that $2\alpha = \beta$. Let $p = (x_1,y_1)$, $q = (x_2,y_2)$, $w_1 = |x_1 - x_2|$ and $w_2 = |y_1 - y_2|$. The number of shortest paths $f(w_1,w_2)$ between p and q is

$$f(w_1, w_2) = \sum_{i=0}^{min\{w1,w2\}} \frac{(\, w_1 + w_2 - i)!}{i!\,(w_1 - i)!\,(w_2 - i)!} \tag{1}$$

Proof. In this case, a diagonal step has exactly the same weight as two consecutive movements to cityblock neighbours. The number of shortest weighted paths between p and q depends on the number of used diagonal steps (β-steps) between the two points, which is at most the minimum difference of the two coordinate values of p and q. Since each diagonal step can be substituted by two consecutive α-steps (a horizontal and a vertical one), the number of diagonal steps may be less, potentially equalling zero, meaning that the points are connected by only cityblock steps. (In special cases, when the two points p and q share a coordinate value, the shortest path cannot contain diagonal steps. Thus, the number of shortest weighted paths is exactly one in this case.) Let i be the number of diagonal steps in the shortest path (these steps can be replaced by α-steps); then, i has a range between $0 \le i \le min\{w_1,w_2\}$. Because each diagonal step can be replaced by two consecutive α-steps, we need to sum up the cases, such as the number of shortest weighted paths corresponding to various value of i. This can be done as follows:

$i = 0$, then all steps in the path are α-steps:$\binom{w_1+w_2}{w_1}$: vertical and horizontal steps in any order;

$i = 1$, then 1 diagonal step and remaining steps are α-steps(horizontal and vertical steps, accordingly): $\frac{(w_1+w_2-1)!}{1!(w_1-1)!(w_2-1)!}$;

i in general, the number of steps is $w_1 + w_2 - i$ from which i are diagonal, $w_1 - i$ and $w_2 - i$ are the number of horizontal and vertical steps. The number of such paths is $\frac{(w_1+w_2-i)!}{i!(w_1-i)!(w_2-i)!}$; $i = min\{w_1,w_2\}$, (the same formula applies for this special case, as we have used in the case $\alpha < \beta < 2\alpha$) : $\binom{max\{w_1,w_2\}}{min\{w_1,w_2\}}$.

To sum these numbers up, the number of shortest weighted paths is computed:

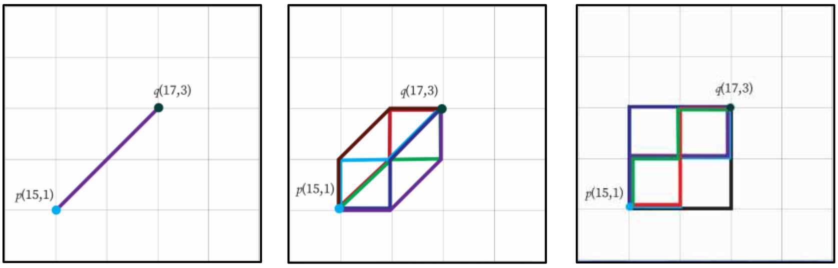
$$\sum_{i=0}^{min\{w_1,w_2\}} \frac{(\, w_1 + w_2 - i)!}{i!\,(w_1 - i)!\,(w_2 - i)!}$$

As is shown in the formula, each time we increment i by 1, the number of diagonal steps is increased by 1 and the number of α-steps is decreased by 2 (1 vertical step, 1 horizontal step); then, the overall number of steps in the shortest path is decreased by i for each i, where in this shortest path we have i diagonal steps, $w_1 - i$ horizontal steps and $w_2 - i$ vertical steps. Therefore, the number of shortest weighted paths according to the value of i of diagonal steps is given as $f(w_1, w_2, i) = \frac{(w_1 + w_2 - i)!}{i!(w_1 - i)!(w_2 - i)!}$ using the fact that the order of steps is arbitrary. ∎

Example 4. Let $p(15,1)$ and $q(17,3)$ and weights α=3, β=6 be given. Then, w_1 = 2, w_2 = 2 and $min\{w_1, w_2\}$ = 2, thus, the number of shortest paths is:

$$f(2,2) = \sum_{i=0}^{2} \frac{(2 + 2 - i)!}{i!\,(2 - i)!\,(2 - i)!} = 13.$$

As we can see, the number of shortest weighted paths, in this case, can be computed by various numbers of diagonal steps with a maximum of $min\{w_1, w_2\}$. Figure 6 (a), (b) and (c) shows all the possible shortest paths between points p and q of Example 4 separated by the possible number of diagonal steps.



(a) (b) (c)

Figure 6

The shortest paths between $p(15,1)$ and $q(17,3)$, with α= 3 and β= 6 (case 2α=β>0), when i= 0,1 and 2 is the number of diagonal steps (starting from i=0 to i= $min\{w_1, w_2\}$ in the path from p to q

Summarizing the results of this case, Figure 7 shows the 3D chart for the values of the number of shortest weighted paths from point (0,0) to all points in a 14 × 14 window with the origin in the middle. The function grows most rapidly on the diagonal directions.

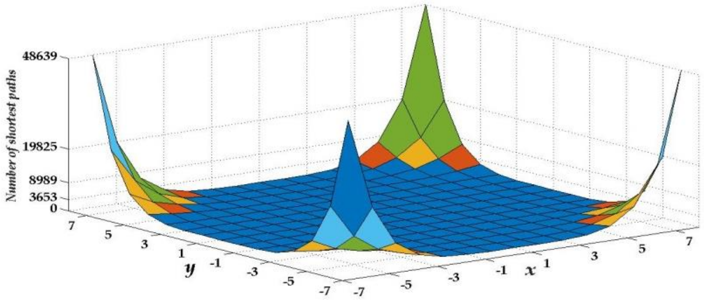


Figure 7

The number of shortest weighted paths from the point (0,0) to other points in a 14 × 14 window with corners (−7, −7), (7, −7),(7,7) and (−7,7), in case $\beta = 2\alpha > 0$

3.4 Case of $\beta = \alpha > 0$

Theorem 4. Let α and β be (positive) weights for cityblock and diagonal movements, respectively, with $\alpha = \beta$. Let $p = (x_1,y_1)$, $q = (x_2,y_2)$, $w_1 = |x_1 - x_2|$ and $w_2 = |y_1 - y_2|$. The number $f(w_1,w_2)$ of the shortest paths between p and q is counted as

$$f(w_1,w_2) = \sum_{i=0}^{\left\lfloor\frac{max\{w_1,w_2\}-min\{w_1,w_2\}}{2}\right\rfloor} \binom{max\{w_1,w_2\}}{i}\binom{max\{w_1,w_2\}-i}{min\{w_1,w_2\}+i} \qquad (2)$$

Proof. In this case, the weight of a diagonal step equals the weight of an α step (i.e. a vertical or horizontal step). The number of steps in a shortest path is clearly given by $max\{w_1,w_2\}$ (as in chessboard distance). Since one does not need to pay any extra for diagonal steps, it is possible, for example, that instead of having two consecutive α steps in the same direction, two diagonal steps are applied, reaching the same point after the two steps. In these paths, there are diagonal steps that are in an unnecessary direction (i.e. shortest paths can be obtained without any such direction steps). Let i denote the number of such unnecessary direction diagonal steps. Evidently, the minimum value of i is 0. For any such step, we need to have an extra diagonal step (in the other diagonal direction, to equalize its effect) instead of an α step. Originally, without any unnecessary diagonal steps ($i = 0$), there are exactly $max\{w_1,w_2\} - min\{w_1,w_2\}$ number of α steps in a shortest weighted path. Thus, the number of α steps decreases by two when an unnecessary diagonal step is introduced. Thus, the maximum of i will be $\left\lfloor\frac{max\{w_1,w_2\}-min\{w_1,w_2\}}{2}\right\rfloor$, where the floor function is used. When i is fixed, we know the number of various steps in the shortest path(s): there are $max\{w_1,w_2\}$ steps, from which i are unnecessary diagonal steps, and we have also $min\{w_1,w_2\}+ i$ number of diagonal steps in the other diagonal direction. The remaining steps are α steps, and their number is $(max\{w_1,w_2\} - i) - (min\{w_1,w_2\} + i) = max\{w_1,w_2\} - min\{w_1,w_2\} - 2i$.

Thus, the number of shortest paths with various values of i can be computed as follows:

$i = 0$, then all steps in the path are in the right direction diagonal and α-steps, and their number is $\binom{max\{w_1,w_2\}}{min\{w_1,w_2\}}$;

for i in general:

$$\frac{max\{w_1,w_2\}!}{i!\,(min\{w_1,w_2\}+i)!\,(max\{w_1,w_2\}-min\{w_1,w_2\}-2i)!} \qquad (3)$$

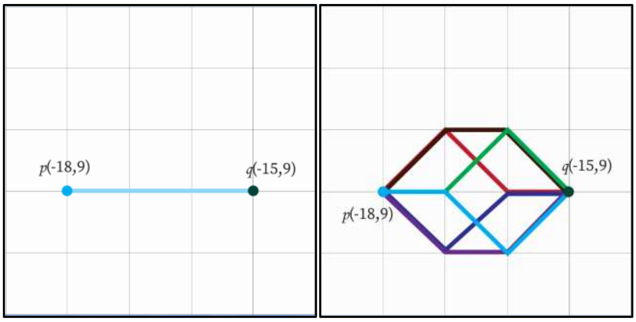
Where $i=\left\lfloor\frac{max\{w1,w2\}-min\{w1,w2\}}{2}\right\rfloor$is the maximum value for i. Thus, the total number of shortest paths is the sum of those:

$$f(w_1,w_2) = \sum_{i=0}^{\left\lfloor\frac{max\{w_1,w_2\}-min\{w_1,w_2\}}{2}\right\rfloor} \binom{max\{w_1,w_2\}}{i}\binom{max\{w_1,w_2\}-i}{min\{w_1,w_2\}+i} \qquad \blacksquare$$

Example 5. Let us use the points $p(-18,9)$ and $q(-15,9)$ with weight values $\alpha=1$, $\beta=1$. Then $w_1=3$, $w_2 = 0$ and thus, $min\{w_1,w_2\} = 0$, $max\{w_1,w_2\} =3$. Further, the number of shortest weighted path (where the distance is 3) is:

$$f(3,0) = \sum_{i=0}^{1} \binom{max\{3,0\}}{i} \binom{max\{3,0\} - i}{min\{3,0\} + i} = 7$$

These paths are also illustrated in Figure 8 (a) and (b), with $i = 0$ and $i = 1$, respectively.



$i = 0, f(3,0,i = 0) = 1$ $i = 1, f(3,0,i = 1) = 6$

(a) (b)

Figure 8

Shortest paths between $p(-18,9)$ and $q(-15,9)$ with $\alpha = \beta$, when $i = 0$ and 1, respectively in (a) and (b), where i is the number possible diagonal steps to an unnecessary direction in the path from p to q

To show how these numbers are changing in the function of the coordinate differences, in Figure 9 we present a 3D chart for the number of shortest paths from the origin to other points in a 14 × 14 window when the diagonal and cityblock steps have the diagonals at the minimum places of this curve while it grows rapidly on the axes.

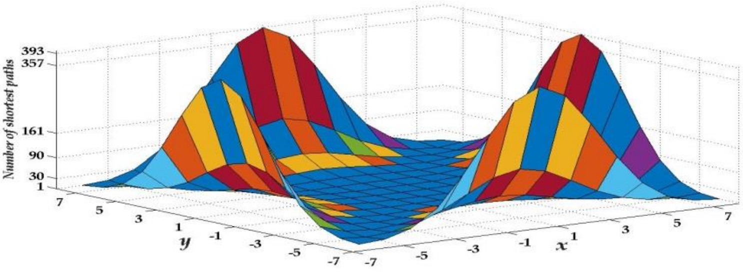


Figure 9

The number of shortest weighted paths from point (0,0) to other points in a 14 × 14 window with corners (−7,−7), (7, −7),(7,7) and (−7,7), in case $\beta = \alpha > 0$

3.5 Case of $0 < \beta < \alpha$

In this case, β steps (diagonal steps) have less weight than α steps (i.e. horizontal and vertical steps); therefore, it will be more convenient and shorter to move from one point to another by diagonal steps, the shortest path between two points relying on the parity of the sum S of the absolute differences of the coordinates of the points. Therefore, we discuss two sub-cases in the two subsections below.

3.5.1 Sub-Case of $0 < \beta < \alpha$ for Points with Even Sum of Differences

Theorem 5.1. Let α and β be the weights for cityblock and diagonal movements, respectively, with $\alpha > \beta$. Let $p = (x_1,y_1)$, $q = (x_2,y_2)$, $w_1= |x_1 - x_2\,|$ and $w_2= |y_1 - y_2\,|$. If $S = w_1+w_2$ is an even number, then the number of the shortest paths between p and q, denoted by $f(w_1,w_2)$, is computed as

$$f(w_1, w_2) = \binom{\max\{w_1, w_2\}}{\frac{\min\{w_1, w_2\} + \max\{w_1, w_2\}}{2}} \qquad (4)$$

Proof. The number of steps between two points is given as $max\{w_1,w_2\}$; moreover, all of them can be diagonal steps. As we showed previously in Subsection 3.2 (case $\alpha < \beta < 2\alpha$), $min\{w_1,w_2\}$ is the number of original diagonal steps in a shortest path. The remaining number of steps, $max\{w_1,w_2\}-min\{w_1,w_2\}$, can also be expressed by diagonal steps in this case; we call these diagonal steps 'added' diagonal steps. These added diagonal steps are used instead of the α-steps of the case $\alpha < \beta < 2\alpha$. These added diagonal steps are of two directions. One of them is the one we have called 'unnecessary' direction. We must have them in this case if $w_1 \neq w_2$. (In case of equality, the shortest path is built up by original diagonal steps to the same direction.) We need to add the same number of unnecessary direction diagonal steps and other (original) direction steps. Thus, the number of unnecessary direction diagonal steps is $\frac{max\{w_1,w_2\}-min\{w_1,w_2\}}{2}$, and the same number of added diagonal steps is needed. Therefore, the number of diagonal steps in a shortest path is $(min\{\{w_1,w_2\} + \frac{max\{w_1,w_2\}-min\{w_1,w_2\}}{2}) + \frac{max\{w_1,w_2\}-min\{w_1,w_2\}}{2}$. The first term gives the number of original direction diagonal steps (both the original and the added ones), while the second term gives the unnecessary direction diagonal steps. The sum equals $max\{w_1,w_2\}$.

Since the order of these steps is arbitrary, the number of shortest weighted paths between points p and q is the number of possible arrangements of these diagonal steps in the shortest path. Consequently, their number can be expressed by the following equation:

$$f(w_1, w_2) = \binom{max\{w_1, w_2\}}{\frac{max\{w_1, w_2\} + min\{w_1, w_2\}}{2}}$$

Equivalently, it can be written as the following binomial coefficient:

$$f(w_1, w_2) = \binom{max\{w_1,w_2\}}{min\{w_1,w_2\}+\frac{max\{w_1,w_2\}-min\{w_1,w_2\}}{2}} \qquad (5)$$ ■

Let us analyse a special case. When w_1 or w_2 equals zero, the number of original diagonal steps is $min\{w_1, w_2\} = 0$, and the shortest path contains only added diagonal steps: one (any) of the directions is then unnecessary, and we have the same number of other added diagonal steps. In this case, the previous formula, the number of shortest weighted paths is simplified as follows:

$$f(w_1, w_2) = \binom{max\{w_l, w_2\}}{\frac{max\{w_1, w_2\}}{2}} \qquad (6)$$

Example 6. Let $p(0,0)$, $q(3,1)$, $\alpha = 2$ and $\beta = 1$ be given. Then, $w_1 = |3 - 0| = 3$ and $w_2 = |1-0| = 1$, then the number of shortest paths between p and q is:

$$f(3,1) = \binom{max\{3,1\}}{\frac{max\{3,1\} + min\{3,1\}}{2}} = \binom{3}{2} = 3$$

These paths are illustrated in Figure 10.

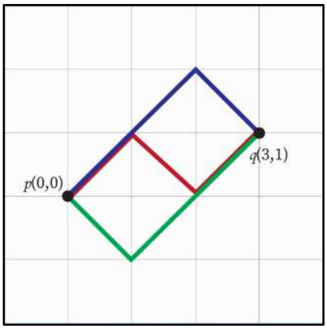


Figure 10
The shortest paths between $p(0,0)$ and $q(3,1)$ in case $0 < \beta < \alpha$ with even sum of differences

3.5.2 Sub-Case of $0 < \beta < \alpha$ for Points with Odd Sum of Differences

Theorem 5.2. Let α and β be the weights for cityblock and diagonal movements, respectively, with $\alpha > \beta$. Let $p = (x_1,y_1)$ and $q = (x_2,y_2)$ be two points in the square grid and $w_1 = |x_1 - x_2|$ and $w_2 = |y_1 - y_2|$. If $S=w_1+w_2$ is an odd number, then the number $f(w_1,w_2)$ of the shortest paths between the points p and q is determined as

$$f(w_1, w_2) = \binom{max\{w_1, w_2\} - 1}{\frac{min\{w_1, w_2\} + max\{w_1, w_2\} - 1}{2}} \cdot max\{w_1, w_2\} \qquad (7)$$

Proof. In this case, we must have a cityblock step (α step) in the path because there is no way to have a shortest path with only diagonal steps (with β steps, two coordinates are always modified by ±1, and thus odd difference cannot be eliminated). The number of diagonal steps, thus, is $max\{w_1,w_2\}-1$, since the number of steps in this shortest path is $max\{w_1,w_2\}$. Let *q'* be the cityblock neighbour of *q* that is the closest to *p* (i.e. a shortest path between *p* and *q'* can be obtained by $max\{w_1,w_2\} - 1$ diagonal steps). Actually, a shortest path from *p* to *q* contains exactly the same number of various direction steps as the shortest path from *p* to *q'* plus an additional cityblock step in the direction that is the same as from *q'* to *q*. The number of shortest paths is counted as the number of possible arrangements of the diagonal steps and the cityblock steps. Applying Theorem 5.1, the number of ways to have the diagonal steps between *p* and *q* is as follows:

$$\binom{max\{w_1,w_2\}-1}{\frac{min\{w_1,w_2\}+max\{w_1,w_2\}-1}{2}} \tag{8}$$

Then, the number of ways to locate one cityblock step (which may not necessarily be the last step of the shortest path, but can be anywhere) is as follows:

$$\binom{max\{w_1,w_2\}}{1} = max\{w_1,w_2\}$$

From these, the number of shortest weighted paths is given by the following equation:

$$f(w_1,w_2) = \binom{max\{w_1,w_2\}-1}{\frac{min\{w_1,w_2\}+max\{w_1,w_2\}-1}{2}} \cdot max\{w_1,w_2\} \qquad \blacksquare$$

Example 7. Let the points *p*(2,0) and *q*(4,3), and the weights *α* = 3 and *β* = 2 be given. Then $w_1 = |4-2| = 2$ and $w_2 = |3-0| = 3$, therefore $w_1 + w_2 = 5$ (which is odd number), then the number of shortest paths between *p* and *q* is:

$$f(2,3) = \binom{max\{2,3\}-1}{\frac{max\{2,3\}+min\{2,3\}-1}{2}} \cdot max\{2,3\} = \binom{2}{2} \cdot 3 = 3$$

Figure 11 shows all these shortest paths.

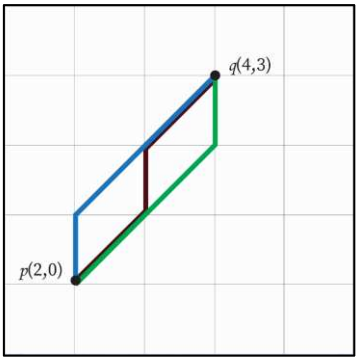


Figure 11
The shortest paths between *p*(2,0) and *q*(4,3), in case 0 < *β* < *α* with odd sum of differences

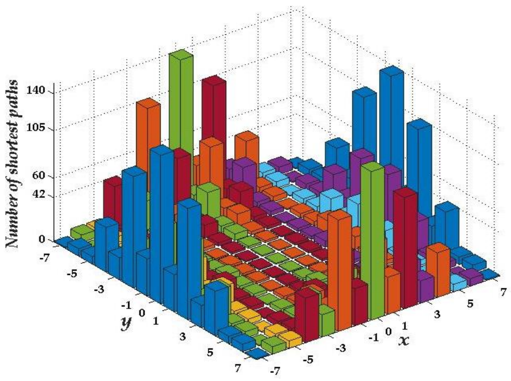


Figure 12
The number of shortest weighted paths from the origin (0,0) to other points in a 14 × 14 window with corners (−7, −7), (7, −7),(7,7) and (−7,7), in case 0 < $\beta < \alpha$.

Finally, for summarizing the case when diagonal steps have lower weights than cityblock steps, Figure 12 shows the number of shortest paths between (0,0) and other points in a 14 × 14 window. For the subcases, we also separately show the values: Figure 13 represents the cases ($\beta < \alpha$ for points with even coordinate sum *S*) and ($\beta < \alpha$ for points with odd coordinate sum *S*) for the number of shortest weighted paths from point (0,0) in a 14 × 14 window. One can observe that minimal values are given on the diagonals, while the function is growing with different speeds for the points with odd and even coordinate sums. For odd coordinate sums, it grows more rapidly. The largest growth values are on the axes with a growing coordinate difference.

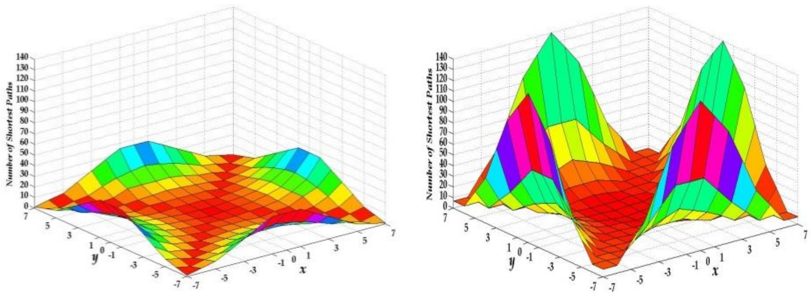


Figure 13
The number of shortest weighted paths from the origin (0,0) to other points in a 14 × 14 window with corners (−7, −7), (7, −7), (7,7) and (−7,7), in case $\beta < \alpha$ for points with even coordinate sum S shown on the left and with odd coordinate sum S shown on the right

3.6 Cases with Zero Weights

In this subsection we consider the special scenarios when one or both of the weights is/are 0. In subsection 3.6.1 we consider the cases when $\alpha = 0$, while in 3.6.2 we consider the case when α is positive, but β has zero value. Up to our knowledge, these cases were never considered before.

3.6.1 Case of $\alpha = 0$

Theorem 6.1. Let $\alpha = 0$ and $\beta \geq 0$ be the weights for cityblock and diagonal movements, respectively. The distance of points $p = (x_1,y_1)$ and $q = (x_2,y_2)$ is 0, since there are paths between any two points built up only by cityblock steps. Moreover, there are infinitely paths between p and q with sum of the weights 0.

Proof. Consider the path built up by cityblock steps along the line from $p = (x_1,y_1)$ to $r = (x_1,y_2)$ with fixed first coordinate concatenated with the path from r to $q = (x_2,y_2)$ on the line with fixed second coordinate. (One or both of these paths could be empty, i.e., with 0 steps, depending on the fact if the points $p = (x_1,y_1)$ and $p = (x_1,y_1)$ share one or two or no coordinates.) The cost of this path is 0 and thus, the distance of the points with the condition $\alpha = 0$ is also zero.

Now, w.l.o.g., assume that $x_1 \leq x_2$. Let us consider the paths defined as follows: cityblock steps along the lines from $p = (x_1,y_1)$ to $p' = (x_1-n,y_1)$ (for any positive integer n), then from p' to $r' = (x_1-n,y_2)$, and from r' to q. Since the sum of the weights of this path is 0 for any value of n, all of these paths are considered as shortest paths between the two mentioned points, thus, there are infinitely many of them. ■

Theorem 6.2. Let $\alpha = 0$ and $\beta \geq 0$ be the weights for cityblock and diagonal movements, respectively. The weighted distance defined by these weights is not metrical, but it is a pseudometric.

Proof. A pseudometric is a distance function which has non-negative values, is symmetric, fulfils the triangular inequality, and such that the distance from any point to itself is 0. All of these properties are easy to check since all distance values are 0. A distance is metric if it is a pseudometric, moreover if the distance of two points is 0, then the points coincide. This additional property is dropped by the considered distance function, thus it is not a metric. ■

Actually, the given pseudometric is the trivial pseudometric, since all the distance values are zero.

3.6.2 Case of $\alpha > \beta = 0$

Theorem 6.3. Let $\beta = 0$ and $\alpha > 0$ be the weights for cityblock and diagonal movements, respectively. The distance of points $p = (x_1,y_1)$ and $q = (x_2,y_2)$ is 0 if and only if the sum of coordinate differences, $w_1 + w_2 = |x_1 - x_2| + |y_1 - y_2|$ is even. On the other hand, the distance of points p and q is α if and only if the sum of coordinate differences, $w_1 + w_2$ is odd. The number of paths between the points with the given length is infinite in both cases.

Proof. Consider, first, the case, when the sum of the coordinate differences is even. There are paths between any two points built up only by diagonal steps. For instance, consider the diagonal line with slope 1 containing point p and the "antidiagonal" line, the line with slope −1 going through on q. These two lines will intersect each other at a point r with coordinates $x_1 + n = x_2 - m$ and $y_1 + n = y_2 + m$ for a pair of integers n and m, where these integers give the number (and the direction) of the diagonal steps from p to r and from r to q, respectively. Thus, it is clear that the distance of the points becomes zero. Furthermore, the given path can be easily modified to contain more and more diagonal steps (in a similar manner as we have shown in the proof of Theorem 6.1), thus the number of paths with length 0 becomes infinite.

Now, let us consider the case when the sum of the coordinate differences is an odd number. Since in every diagonal step, both of the coordinates change by ±1, we cannot reach from one (of the points p and q) the other point only by diagonal steps. However, we can reach any of its cityblock neighbours by only diagonal zero-weight steps, thus, we need one extra cityblock step in the path resulting in the distance of the points being α in this case. As the number of zero length paths between p and a given cityblock neighbour of q is (according to the first part of the proof) is infinite, each of them produces a shortest, i.e., α length path between p and q by adding the last cityblock step, hence the proof. ■

By a similar proof as the proof of Theorem 6.2, one can also establish the following result.

Theorem 6.4. Let $\alpha > 0$ and $\beta = 0$ be the weights for cityblock and diagonal movements, respectively. The weighted distance defined by these weights is not metrical, but it is a pseudometric.

Finally, in the last subsection, we consider the case, when at least one of the weights is negative.

3.7 Case of Negative Weight(s)

Table 1
The discussed cases for the weights α and β

Condition	Only positive weights						
	$2\alpha < \beta$	$2\alpha = \beta$	$\alpha < \beta < 2\alpha$	$\alpha = \beta$	$\beta < \alpha$	$\alpha = 0 \leq \beta$	$\alpha > \beta = 0$
Case /subsection	**3.1**	**3.3**	**3.2**	**3.4**	**3.5**	**3.6.1**	**3.6.2**

If one or both of the weights α and β are negative, then it is easy to see that one can find a path between any two points with sum of the weights that is less than any number. In this case the "distance" of the points does not exist, i.e., it could be written as $-\infty$. Since such a path (with finite length) does not exist, there is no shortest path. Actually, in these cases the obtained weighted distance is not distance since the condition of the non-negativeness of its value does not hold. With this note we have finished checking all possible cases, and we give a short summary and discussion in the last section.

4 Discussion and Conclusion

In fact, our shortest paths can be represented by trajectories on the digital grid. A combinatorial problem, the number of cheapest solutions, i.e., the number of shortest weighted paths is computed in various scenarios. The numbers of shortest paths with the cityblock and chessboard metrics were already known [6]. However, we have presented results for a much larger class of digital distances, for chamfer distances, in this way our study can be seen as a generalisation of these previous results. Digital distances can be used in various ways in communication networks [16], and they are also related to combinatorial problems. For example, the number of shortest paths gives important features of a network. Results on the number of shortest paths for neighbourhood sequence distances were presented in [15], in this sense we have completed the picture by presenting here analogous results for the other type of widely used digital distance family. In this paper, we have analysed rigorously all the cases to find the number of minimum weighted paths between any two points in a square grid. The cases depend on the value of weights given to cityblock steps (α steps) and diagonal steps (β steps). We have discussed five cases with positive weights and two cases when weight zero is allowed. We have seen that the results obtained in them are pairwise different. By Table 1 one can also be sure that there are no more cases, all the possibilities to have positive and/or zero weights for both cityblock and diagonal steps are discussed. Our results with positive weights are also displayed in 3D graphs, which show how the resulting functions grow. In most cases. the functions have strong monotonic behaviour as one goes further from the origin. The cases of $\beta > 2\alpha$ and $\beta = 2\alpha$ show very similar behaviour (see Figure 4 and 8). The case of $\alpha < \beta < 2\alpha$ does not seem to relate to any other cases, (Figure 6). Case of $\beta = \alpha$, displayed in Figure 10, shows some relation to the case of $\beta < \alpha$, however, this latter is more complicated than the others, see Figure 13. We highlight the results of this letter case, i.e., when the diagonal steps have less weight than cityblock steps. As we have seen, the result is described by two different functions depending on the parity of the sum of the coordinate differences of the points, thus it does not behave in a monotonic way. We have shown also the cases when zero weight is allowed. If the cityblock step has zero weight, all the distances become 0. Contrary, if only the diagonal movements without cost, but the cityblock steps have a positive weight, then somewhat similarly to the case of $\beta < \alpha$, the result is not monotonous but given by two different values alternating for the points of the grid. We have also complemented our results by providing the case when negative weights are allowed. In those cases, we cannot really refer to distances based on the smallest weighted paths. Our results are useful in network analysis, in digital image processing and in shape analysis. We believe that it is important also for the application point of view to consider all the possible cases depending on the possible values of the weights. The number of shortest weighted paths between points that contain a given point or a set of given points can be discussed in the future. For example, if we have path *s,…,b,…,t*, then it can be computed how many shortest paths between *s* and *t* contain *b*. Extensions to higher dimensional or other grids (architectures) can also be done.

References

[1] Alzboon, L., Khassawneh, B., Nagy, B. (2017) On the Number of Weighted Shortest Paths in the Square Grid. Proceedings IEEE 21st Int. Conf. Intelligent Engineering Systems (INES), Larnaca, Cyprus, pp. 83-90, DOI: 10.1109/INES.2017.8118533

[2] Borgefors, G. (1986) Distance transformations in digital images. Computer vision, graphics, and image processing, 34(3), 344-371, DOI: 10.1016/s0734-189x(86)80047-0

[3] Cheng, E., Grossman, J., Qiu, K., Shen, Z. (2013) The number of shortest paths in the arrangement graph. Information Sciences, 240, 191-204, DOI: 10.1016/j.ins.2013.03.035

[4] Das, P., Chakrabarti, P., Chatterji, B. (1987) Distance functions in digital geometry. Information Sciences, 42(1), 113-136, DOI:10.1016/0020-0255(87)90019-3

[5] Das, P., Chakrabarti, P., Chatterji, B. (1987) Generalized distances in digital geometry. Information Sciences, 42(1), 51-67, DOI: 10.1016/0020-0255(87)90015-6

[6] Das, P. (1991) Counting minimal paths in digital geometry. Pattern Recognition Letters, 12(10), 595-603, DOI: 10.1016/0167-8655(91)90013-c

[7] Das, P. P. (1989) An algorithm for computing the number of the minimal paths in digital images. Pattern Recognition Letters, 9(2), 107-116, DOI: 10.1016/0167-8655(89)90043-3

[8] Hanneman, R., Riddle, M. Introduction to social network methods, Riverside, CA, University of California, Riverside, 2005 (published in digital form at http://faculty.ucr.edu/~hanneman/ [accessed 03.03.2019])

[9] Kari, L., Konstantinidis, S., Sosik, P. (2004) Substitutions, trajectories and noisy channels. CIAA, Kingston, Canada, LNCS, 3317, 2004, 202-212, ISBN: 3-540-24318-6, 978-3-540-24318-2

[10] Kari, L., Konstantinidis, S., Sosik, P. (2005) Operations on trajectories with applications to coding and bioinformatics. International Journal of Foundations of Computer Science, 16(3), 531-546, DOI: 10.1142/S0129054105003145

[11] Klette, R., Rosenfeld, A. Digital geometry: Geometric methods for digital picture analysis; Morgan Kaufmann, 2004, ISBN: 1-55860-861-3, 978-1-55860-861-0

[12] Mohanty, G. Lattice Path Counting and Applications, Academic Press, 1979, ISBN: 978-0-12-504050-1

[13] Nagy, B. (2002) Metrics based on neighbourhood sequences in triangular grids. Pure Math. Appl., 13(1), 259-274

[14] Nagy, B. (2003) Distance functions based on neighbourhood sequences. Publ. Math., 63(3), 483-493

[15] Nagy, B. (2015) On the number of shortest paths by neighborhood sequences on the square grid. Joint Austrian-Hungarian Math Conf, Széchenyi István Univ, Győr, Hungary (abstract)

[16] Nagy, B. (2017) Application of neighborhood sequences in communication of hexagonal networks. Discrete Applied Mathematics, 216, 424-440, DOI: 10.1016/j.dam.2015.10.034

[17] Oyama, T., Morohosi, H. (2004) Applying the shortest-path-counting problem to evaluate the importance of city road segments and the connectedness of the network-structured system. International Transactions in Operational Research, 11, 555-573, DOI: 10.1111/j.1475-3995.2004.00476.x

[18] Rosenfeld, A., Pfaltz, J. (1968) Distance functions on digital pictures. Pattern Recognition, 1(1), 33-61, DOI: 10.1016/0031-3203(68)90013-7

[19] Yamashita, M., Ibaraki, T. (1986) Distances defined by neighborhood sequences. Pattern Recognition, 19(3), 237-246, DOI: 10.1016/0031-3203(86)90014-2

Resource-Aware Network Topology Management Framework

Aaqif Afzaal Abbasi[1], Shahaboddin Shamshirband[2,3*], Mohammed A. A. Al-qaness[4], Almas Abbasi[5], Nashat T. AL-Jallad[6,7], Amir Mosavi[8,9,10,11]

[1]Department of Software Engineering, Foundation University, Islamabad, 44000 Pakistan; aaqif.afzaal@fui.edu.pk

[2]Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh 758307, Vietnam

[3]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh 758307, Vietnam; shahaboddin.shamshirband@tdtu.edu.vn

[4]School of Computer Science, Wuhan University, Bayi Road No. 299, Wuhan 430072, China; alqaness@whu.edu.cn

[5]Department of Computer Science, International Islamic University, Islamabad 44000, Pakistan; almas.abbasi@iiu.edu.pk

[6]School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430079, China

[7]School of Computer Science, Palestine Technical University, Tulkarem 44864, Palestine; jallad@whut.edu.cn

[8]School of the Built Environment, Oxford Brookes University, No.5, Jack Straws Lane, Oxford OX3 0BP, UK; a.mosavi@brookes.ac.uk

[9]Kandó Kálmán Faculty of Electrical Engineering, Óbuda University, Bécsi út 94-96, 1034 Budapest, Hungary; amir.mosavi@kvk.uni-obuda.hu

[10]Institute of Structural Mechanics, Bauhaus University Weimar, Marienstraße 15, D-99423 Weimar, Germany; amir.mosavi@uni-weimar.de

[11]Department of Mathematics and Informatics, J. Selye University, Hradná str. 21, 94501 Komarno, Slovakia; a.mosavi@ujs.sk

Abstract: Cloud infrastructure provides computing services where computing resources can be adjusted on-demand. However, the adoption of cloud infrastructures brings concerns like reliance on the service provider network, reliability, compliance for service level agreements (SLAs), etc. Software-defined networking (SDN) is a networking concept that suggests the segregation of a network's data plane from the control plane. This *concept improves networking behavior. In this paper, we present an SDN-enabled resource-aware topology framework. The proposed framework employs SLA compliance, Path Computation Element (PCE) and shares fair loading to achieve better topology features. We also present an evaluation, showcasing the potential of our framework.*

Keywords: cloud computing; big data; fog computing; software-defined; networking; network management; resource management; topology

1 Introduction

Cloud computing technology provides computing and networking services over the internet [1]. Computational and I/O resource management in cloud computing is a challenging task. Different methods have been adopted to address the computation and I/O management challenges in cloud systems. Therefore, the success of any cloud management software depends on the efficiency of the system through which it can utilize the underlying networking resources [2, 3, 4]. Figure 1 shows a generic view of services provided by a conventional cloud resource management system. It includes a set of virtual machines (VMs) operating on a network operating system (hypervisor). A VM is an emulation of a physical computing machine. It provides the functions of a computer system by using the resources of the underlying hardware and software resources. User applications are hosted as applications (APPs) on guest operating systems that operate on these VMs. For a detailed study of the topology and virtualization system technology, literature research is available in [1, 3].

There have been many attempts to make networks more manageable and secure. Various methods have been adopted to deliver resource management features in a cloud environment. However, one of the drawbacks in cloud services delivery is that consumers are kept unaware of the details of how cloud services and features are provided. In effect, users focus on what really matters to them, i.e. consuming a service. Similarly, the cloud service providers focus only on aspects of their domain that are largely nontransparent to the end consumers.

In software-defined networking (SDN) [5], the control plane of a network element is separated from its data plane functions. SDN technology is used in data centers to effectively manage network traffic. The SDN principles can also be applied to other areas such as storage, security and service level agreement. Software-defined cloud computing (SDCC) in this term in an approach where all aspects of a data center providing services to the users are software-defined [6]. The principles and concepts of SDCC provide an easy way for reconfiguration and adaptation of physical resources to adjust QoS demands [7, 8]. Figure 2 shows the architecture of an SDN enabled cloud resource management system.

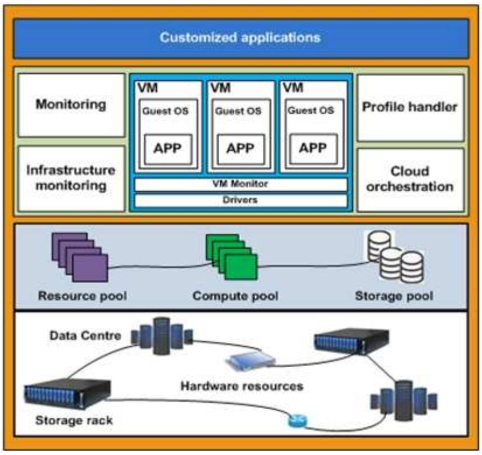


Figure 1

Cloud resource management infrastructure

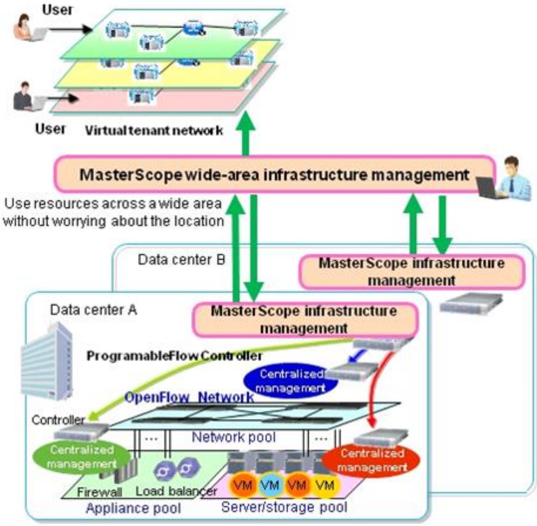


Figure 2

SDN-enabled cloud resource management

In this paper, a network topology management framework is presented. We explain the structure of its topology administration features. The paper also includes a discussion on the architectural developments made in traditional and SDN enabled cloud resource management systems. A thorough analysis of topology management functions has also been presented in Section 4. Finally, we present the results of our framework.

The paper is organized as follows. Section 2 describes the related work. Section 3 discusses the architecture and working of the presented framework. Section 4 presents its working and Section 5 presents the evaluation and analysis of results. Finally, Section 6 concludes the paper.

2 Related Work

2.1 Cloud Resource Management Issues

A gossip-based protocol is presented in [9]. The protocol uses a heuristic solution to solve resource allocation challenges. By delivering a simulation-based approach to solve this problem, the protocol can be employed to achieve fairness in resource allocation scenarios. The issues pertaining to network utility maximization (NUM) in SDN controllers can be addressed by using variable resource decomposition methods which can also look after the update rules in parallel.

The SLA based resource allocation challenges have been discussed in [10]. Due to the recent emergence of complex IT technologies the user applications are becoming complicated. In particular, the cloud management framework in [11] discusses the resource provisioning in a cloud environment. Among the multiple issues faced by cloud services, the most persistent problem is that users are unaware of the services provisioning methodology. The recent surveys on the topic indicate that cloud system developers should develop tools to automate cloud operator tasks. This will improve cloud services delivery and will bring transparency to the technology audit mechanism. Resource management strategies for improving network overheads are discussed in [12]. The research analyzes the pros and cons of these strategies particularly in terms of performance costs and services stability. Other techniques used for controlling and managing network traffic across WAN includes the use of multiprotocol label switching (MPLS) over SDN managed carrier connections that can handle incoming network traffic from multiple locations.

In cloud-based systems, multiple user services are entertained simultaneously. This is possible due to the recent advancements in efficient parallel data processing techniques. Efficient parallel data processing is described in [13]. It presents Nephele, a framework that uses the benefits of resource provisioning services offered by IaaS clouds. The on-demand service provisioning framework for grid computing is presented in [14] where the system allocates resources to users on the basis of their profile and service usage. A profile-based approach is presented in [15] where the user profile is used to evaluate resource usage. A service optimization framework for risk-aware resource provisioning of dynamic resource allocation is presented in [16] where the workload of multiple clients is evaluated under the uncertainty of workloads. SDNs are extensively used for risk assessment to redefine network operations at runtime. This is due to their resilience when used as a control parameter to administer the underlying hardware infrastructure.

2.2 SDN-based Cloud Resource Management

The software-defined cloud functions of a data center are administered by an open-access user interface. This helps in discouraging proprietary software from handling network resource management functions. A software-defined resource manager automatically manages network data, offering easier administration. It can work with existing resource management solutions allowing applications to share common resource management platforms.

Harmony [17] is proposed to manage various aspects of software-defined clouds. It reduces workload dependencies between different tasks. In order to achieve fault tolerance, a model framework has been presented in [18] which realize the true benefits of SDNs in data centers. SDN-based orchestration technologies coordinate together to provide balanced composite cloud and network services. These technologies also ensure that VM allocations in the network topologies are based on estimations of switch/link and server loads.

2.3 SDN as Enabling Technology to Administer Resource Management

SDNs bring network awareness to network control features. SDN controllers can read the entire network topology through subnets. Subnets use available network resources to constitute a logical topology within a network. In [19], a software-defined interface is presented which uses pluggable modules for scheduling and fault management of a network. This enables SDN applications to deploy network control functionalities in a practical multi-tier cloud infrastructure as shown in Figure 3. A software-defined resource manager discovers the inventory of links in a given network by plotting all possible paths across the network. Therefore, if all the applications use the best available path strategy for performing network functions, it can result in bringing greater resource administration and agility. Research studies conducted in [4] are aimed at reducing network operation costs by either combination with virtualization of network services through the use of SDN-enabled resource allocation techniques. Network operation and resource allocation through centralized data streams also benefit network users in simplifying network software upgrades.

3 System Architecture

The proposed framework consists of the SDN application programming interface (API), a cloud resource infrastructure and underlying computational resources as shown in Figure 3. The application management APIs consist of the cloud management console, topology manager and admission controller. The SDN-based API manages topology functions in a cloud environment. Cloud management console acts as an interface between the user and applications.

The SDN API lies next to a network of underlying cloud resources, which logically control resource management operations. It is followed by a layer of virtual and physical components. The physical part typically includes the servers, storage media, and network peripherals. The virtual layer consists of the software deployed across the physical layer. The arrangement of the network entities is similar to [20]. Below we provide a short description of each component.

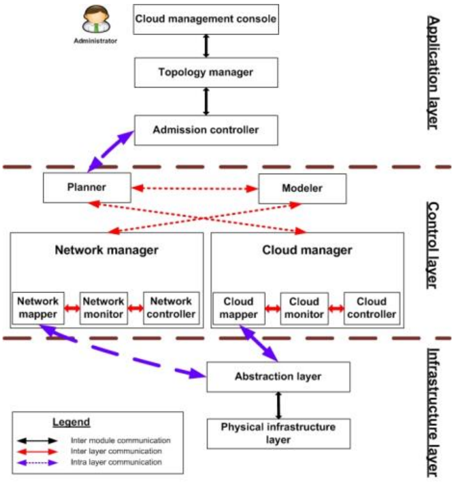


Figure 3
SDN-enabled cloud resource management

Cloud management console: The console facilitates cloud resource management functions to express high-level policies on the underlying network infrastructure.

Topology manager: It maintains a directory of overall network mappings. It enables access to a seemingly infinite pool of computing resources using open federated cloud computing architecture like RESERVOIR [21]. This facilitates the administrator in gaining an understanding of available network resources through a service provider network.

Admission controller: The admission controller receives and evaluates the user requests on the basis of prescribed factors that are defined through SLAs.

Planner: A planner in consultation with the modeler, network manager, and cloud manager determines the location of hosting the services.

Modeler: The modeler compares the planner request with the network manager and cloud manager to plot a resource utilization scope that can be sent to the topology manager for updating the network directory.

Cloud and network manager: Cloud and network managers consolidate data center infrastructure resources. This consolidation is performed at both physical and logical levels. They monitor network status and manage both physical and virtual infrastructures.

Abstraction layer: The layer provides an abstraction of logical deployment of physical hardware for all devices in the cloud infrastructure.

Physical infrastructure layer: It consists of the physical resources of a data center network including computing devices, storage devices, servers, and other networking equipment.

The proposed topology manager provides an insight into the hierarchical structure and state of the cloud. By using SDN's flexible nature, the topology manager helps in organizing software and hardware into zones, blocks, servers, nodes, resource pools, and software deployments. In the next section, we explain the working of the topology manager in detail.

4 Topology Management Framework

4.1 Topology Management

Route analytics and SDN together improve the availability of network routes and traces in data centers. SDNs allow network administrators to manage real-time network-wide abstractions into topology control. This helps the SDN-enabled data center infrastructure to use the latest traffic status and workload profiles. In this regard, TOSCA is presented in [22]. It is an industrially-endorsed standardization for topology specifications. By using SDN control functions, it creates a blend of network traffic requirements. This helps network and cloud applications to derive the relationship between a service and its behavior in the network. A topology manager's implementation provides a complete view of the availability of devices and resources. Its function is independent of the vendor-specific monitoring technologies. The proposed topology manager uses a combination of techniques for monitoring and customization of third party services. Below we give a short description of its components and their functions.

Application handler: When topology requests are scheduled for deployment, the application handler registers these services. It performs multiple checks on topology requests and later sends them for the compatibility checks.

Service handler: Once a topology request confirms SLA requirements, it is accessed by a service handler that compares its requirements with available virtual infrastructure.

Path Computation: PCE [23] helps in reducing resource-based computational constraints during path computation by considering multiple constraints. In our model, we first ensure that the experimental test-bed has sufficient resources for task execution. Then, we use PCE for path computation.

The topology management operation initiates when an application requests for grant of resources. A graphical portrayal of the topology manager scheme is given in Figure 4. The application handler performs a check operation on SLA violations. If SLA violations are made the process is terminated. If not, a check is performed for available resources. The applications with limited resources are dropped. All dropped processes are prompted to the application handler for rejection. Those processes with enough resources are forwarded to the PCE module. The PCE module is used to define the suitable path between the traffic source and destination. It then assigns optimal resources to the process/ request. The use of PCE helps in reducing the processing overheads as it uses the previously calculated paths from the path allocation table.

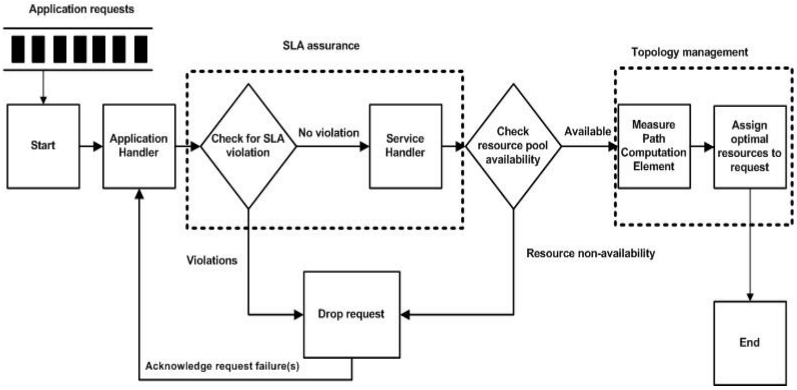


Figure 4
Topology management - Process flow diagram

We evaluate the proposed cloud management console prototype on an Intel Xeon CPU E5-2650 v2 with a clock speed of 2.60 GHz and 8 cores (16 threads). It is developed in Python version 2.7.9. By using Ubuntu 12.04 environment, we use two 8 slot SDN enabled HP 12508E switches, having a 10.8 Tbps maximum switching capacity. We measure the framework for throughput and CPU resource utilization efficiency.

Our results (Figure 6) show that the proposed framework can handle an increasing number of requests. This is because the allocation scheme used in share fair systems is not the same as utilization. Normally a request allocating 50-60 percent of CPU resources to process an application, might only use a part of these resources. Regardless of how big resource allocation is, a resource requesting query always receives 100 percent of the processing capacity. Maintaining a balance among resource usage and allocation is complicated sometimes. Therefore, the allocation of either a small or big chunk of CPU share to a busy workload might not solve the problem, yet it may result in slower performance.

We conducted a performance evaluation of our proposed framework by submitting admission control requests and then measuring the system's CPU and memory utilization. We used conditional statements to ensure that SLA conditions are not violated for the topology management scheme. In our experiments, we used the Batch workload. By submitting the Batch workload jobs through a job queue, it was ensured that the submitted data load runs unconstrained, and is free from bottlenecks.

Figures 5 and 6 show the CPU and memory resources utilization. We compared the proposed framework efficiency with Realistic and Capacity-aware admission control schemes for 9 instances. The realistic approach employs product logic for modeling requests. A detailed study of product logic is presented in [24]. The capacity-aware admission control scheme uses optimized risk values for CPU and I/O functions and employs real-time values for memory mapping of received admission control requests. From the results can be seen that Realistic and Capacity aware schemes are constraining network resources. These techniques achieved better performance for one capacity (CPU usage) at the expense of others (memory usage). This results in presenting a more asymmetric behavior. However, the results show small signs of jitter due to the continuous increase in the amount of resource utilization.

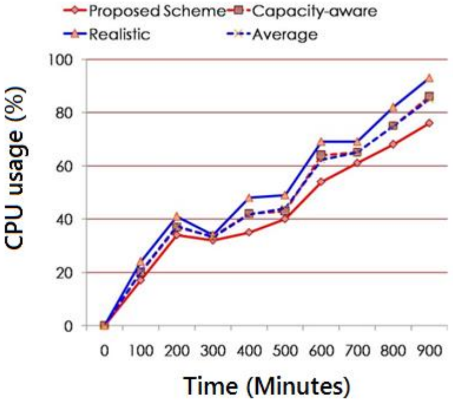


Figure 5
CPU resources utilization

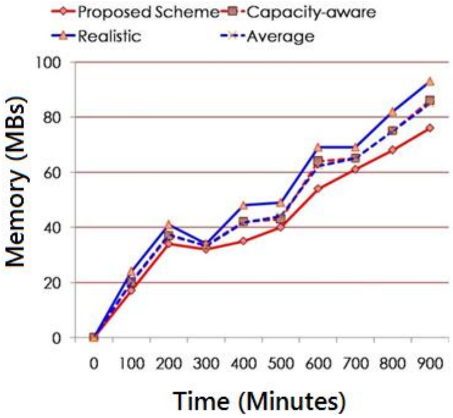


Figure 6
Memory resources utilization

In Figure 7, we present the overall resource utilization comparison. Results demonstrate that by using our proposed methodology, resource utilization is even lower than the average of Realistic and Capacity aware schemes. Due to these schemes, we can confirm the effectiveness of our proposed framework.

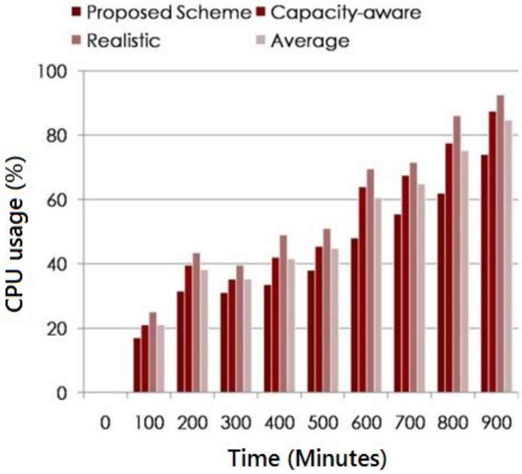


Figure 7
Overall resource utilization comparison

Conclusions

SDN concepts bring great potential and efficiency to reduce the complexity of network control. To efficiently manage and configure the network, it needs to have up-to-date information about the state of the network, in particular, its topology.

The paper presents a network topology framework. It utilizes SDN–enabled infrastructure for improving topology management functions. The proposed framework employs SLAs, PCE and shares fair loading as a means for improving resource administration. The framework facilitates in reducing complexities for resource allocation problems. We evaluate our system on limited real-time controlled data center traffic. We believe that the presented work provides a foundation for developing a more efficient topology management infrastructure. In the future, we plan to improve our system's planner design so that it can effectively handle VM placement and allocation related challenges. We also plan to reduce VM overheads by improving the topology discovery feature in a data center environment.

Acknowledgement

We acknowledge the financial support of this work by the Hungarian State and the European Union under the EFOP-3.6.1-16-2016-00010 project and the 2017-1.3.1-VKE-2017-00025 project.

References

[1] Jin, Hai, Aaqif Afzaal Abbasi, and Song Wu. "Pathfinder: Application-aware distributed path computation in clouds." *International Journal of Parallel Programming* 45.6 (2017): 1273-1284

[2] Abbasi, Aaqif Afzaal, and Hai Jin. "v-Mapper: An Application-Aware Resource Consolidation Scheme for Cloud Data Centers." *Future Internet* 10.9 (2018): 90

[3] Gulati, Ajay, Arif Merchant, and Peter J. Varman. "mClock: handling throughput variability for hypervisor IO scheduling." Proceedings of the 9th USENIX conference on Operating systems design and implementation. USENIX Association, 2010

[4] Mitzenmacher, Michael. "The power of two choices in randomized load balancing." *IEEE Transactions on Parallel and Distributed Systems* 12.10 (2001): 1094-1104

[5] Kim, Hyojoon, and Nick Feamster. "Improving network management with software defined networking." *IEEE Communications Magazine* 51.2 (2013): 114-119

[6] Abbasi, Aaqif Afzaal, Hai Jin, and Song Wu. "A software-defined cloud resource management framework." Asia-Pacific Services Computing Conference, Springer, 2015

[7] Fundation, Open Networking. "Software-defined networking: The new norm for networks." ONF White Paper 2 (2012): 2-6

[8] Nunes, Bruno Astuto A., et al. "A survey of software-defined networking: Past, present, and future of programmable networks." *IEEE Communications Surveys & Tutorials* 16.3 (2014): 1617-1634

[9] Wuhib, Fetahi, Rolf Stadler, and Mike Spreitzer. "Gossip-based resource management for cloud environments." IEEE International Conference on Network and Service Management (CNSM), 2010

[10] Buyya, Rajkumar, et al. "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility." *Future Generation computer systems* 25.6 (2009): 599-616

[11] Cunningham, S. J. "Developing innovative applications of machine learning." Proc. Southeast Asia Regional Computer Confederation Conference, 1999

[12] Juve, Gideon, and Ewa Deelman. "Resource provisioning options for large-scale scientific workflows." IEEE Fourth International Conference on eScience, 2008

[13] Warneke, Daniel, and Odej Kao. "Exploiting dynamic resource allocation for efficient parallel data processing in the cloud." *IEEE Transactions on Parallel and Distributed Systems* 22.6 (2011): 985-997

[14] Leivadeas, Aris, Chrysa Papagianni, and Symeon Papavassiliou. "Efficient resource mapping framework over networked clouds via iterated local search-based request partitioning." *IEEE Transactions on Parallel and Distributed Systems* 24.6 (2013): 1077-1086

[15] Yang, Jie, Jie Qiu, and Ying Li. "A profile-based approach to just-in-time scalability for cloud applications." IEEE International Conference on Cloud Computing, 2009

[16] Kusic, Dara, and Nagarajan Kandasamy. "Risk-aware limited lookahead control for dynamic resource provisioning in enterprise computing systems." *Cluster Computing* 10.4 (2007): 395-408

[17] Grandl, Robert, et al. "Harmony: Coordinating network, compute, and storage in software-defined clouds." ACM Proceedings of the 4th annual Symposium on Cloud Computing, 2013

[18] Baset, Salman A., et al. "Toward achieving operational excellence in a cloud." *IBM Journal of Research and Development* 58.2/3 (2014): 4-1

[19] Lin, Thomas, et al. "Enabling SDN applications on software-defined infrastructure." IEEE Network Operations and Management Symposium (NOMS), 2014

[20] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011)

[21] Rochwerger, Benny, et al. "The reservoir model and architecture for open federated cloud computing." *IBM Journal of Research and Development* 53.4 (2009): 4-1

[22] Mousavi, S., Mosavi, A., Varkonyi-Koczy, A. R. and Fazekas, G., 2017, Dynamic resource allocation in cloud computing. Acta Polytechnica Hungarica, 14(4), pp. 83-104

[23] Farrel, Adrian, J-P. Vasseur, and Jerry Ash. A path computation element (PCE)-based architecture. No. RFC 4655. 2006

[24] Vázquez, Carlos, et al. "A fuzzy approach to cloud admission control for safe overbooking." International Workshop on Fuzzy Logic and Applications. Springer, 2013

Dynamic Simulation of a Prototype Race Car Driven by Series Wound DC Motor in Matlab-Simulink

Attila Szántó, Sándor Hajdu, Gusztáv Áron Sziki

University of Debrecen, Doctoral School of Informatics, Kassai út 26, 4028 Debrecen, Hungary, szanto.attila@zimbra.inf.unideb.hu

University of Debrecen, Faculty of Engineering, Ótemetú u. 2-4, 4028 Debrecen, Hungary, hajdusandor@eng.unideb.hu, szikig@eng.unideb.hu

Abstract: In the last decade student teams at the Faculty of Engineering of the University of Debrecen have designed and constructed several race cars with alternative (electric or pneumatic) drives and took part and achieved successes in domestic and international competitions. For more successful racing a simulation program has been developed in Matlab for the calculation of the vehicle dynamic functions of the cars from their technical data. Currently, it has become a prerequisite of successful racing because of the large number of the possible values of technical data we can select the optimal ones by applying the above program combined with an optimizing procedure. The detailed description of the developed dynamic model and simulation program, together with the vehicle dynamics functions generated by the program are presented here. The description of the optimizing procedure will be presented elsewhere.

Keywords: dynamics modelling; simulation; race car; series DC motor; Matlab

1 Introduction

The Faculty of Engineering, University of Debrecen (FEUD) has years of experience in the designing, development, and construction of vehicles with alternative (electric or pneumatic) drives. As a part of the above activity, student teams at our faculty have designed and constructed several alternative driven race cars in the last decade and took part and achieved successes in domestic and international competitions. One of the above-mentioned competitions is the MVM Race which is sponsored and organized by the Hungarian Electrical Works Ltd (Magyar Villamos Művek) in Budapest from 2014, annually. With an electric driven race car – that was designed and constructed by the Department of Mechanical Engineering of FEUD – we scored first and second place in the above race in 2014 and 2015 respectively.

With the aim of more conscious design and development, thus more effective racing, we have realised a simulation program in Matlab-Simulink for the calculation of the vehicle dynamic functions (acceleration-, velocity-, and covered distance-time functions) of a car from its technical data. The above technical data have to be measured or sometimes they can be found in the specification of a given vehicle component [1, 2, 3]. Applying the above program, together with an optimizing procedure, the optimal values of technical data – with which the best performance can be achieved (e.g. a given distance can be covered in the shortest possible time) – can be determined. For example, determining the optimal gear ratio in the chain drive is of key importance to reach the above aim.

It was important to develop a simulation program with modular structure, so that the different structural units of the car (motor, powertrain, front and back wheels, vehicle body) can be modelled and simulated separately. It was important because we design and construct vehicles with very different vehicle components (e.g. the motor can be pneumatic, electric, or combustion type, and the power train can also be very different). Restructuring the whole program when, e.g. a new type of motor is applied would be a tremendous amount of work, thus we wanted to avoid this situation.

In Section 2 a detailed technical description of the car – which won the first and second place in the MVM race in 2014 and 2015 – is given, while in Section 3 the developed dynamic model for the same car is presented. In section 4 the latest version of our simulation program – which is based on the model in Section 3 – is presented. Finally, in Section 5 the dynamics functions – generated by the simulation program for the modelled car – are presented.

2 The Race Car

The race car was designed considering the rules of MVM competition (prototype category), and it raced first in 2014. The vehicle has a welded frame structure, which is made up of aluminium hollow profiles and tubes (Figure 1) [4].

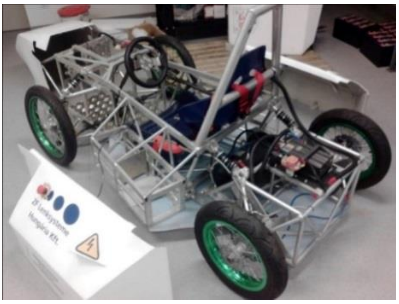


Figure 1
The race car without housing and batteries

The car has a double-wishbone front suspension, while the rear part of the vehicle is suspended independently. This part involves the complete drive-train (Figures 2 and 3) with the electric motor, motor control unit and chain drive, and also the rear wheels. The frame structure also contains battery holders at both sides of the pilot and a safety rollover hoop as well. The vehicle is driven by a series wound DC motor. The motor (1) is connected to the rear shaft (5) through a chain-drive (2, 3, 4). Bevel roller bearings (6) are applied to support the rear shafts. The wheels (8) are connected to the rear shaft rigidly. For effective braking, a divided brake system with four brake discs (7) is applied. The mass of the car – without the driver – is 176.2 kg [4].

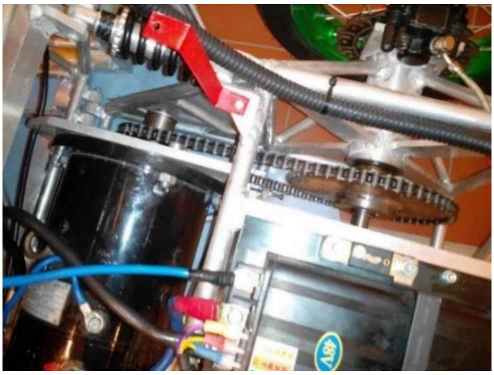


Figure 2
The drive-train of the car

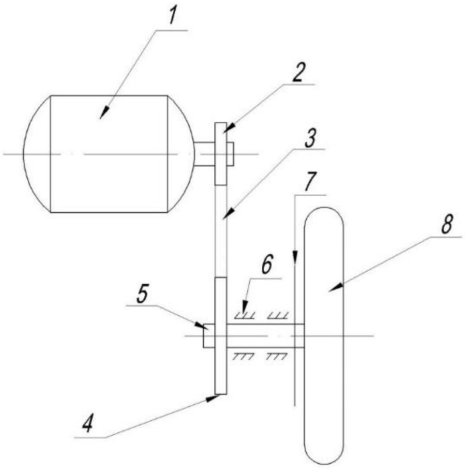


Figure 3
Schematic drawing of the drive-train

The type and technical data of the applied electric motor are summarized in Table 1:

Table 1

Type and technical data of the applied electric motor

Type	Nominal voltage [V]	Nominal power [kW]	Nominal current intensity [A]	Nominal speed [1/min]	Nominal torque [Nm]	Mass [kg]
DC/T4-48	48	4	104	2800	14,7	28

3 The Vehicle Dynamics Model

In our model the car is divided into four structural units (Figure 4).

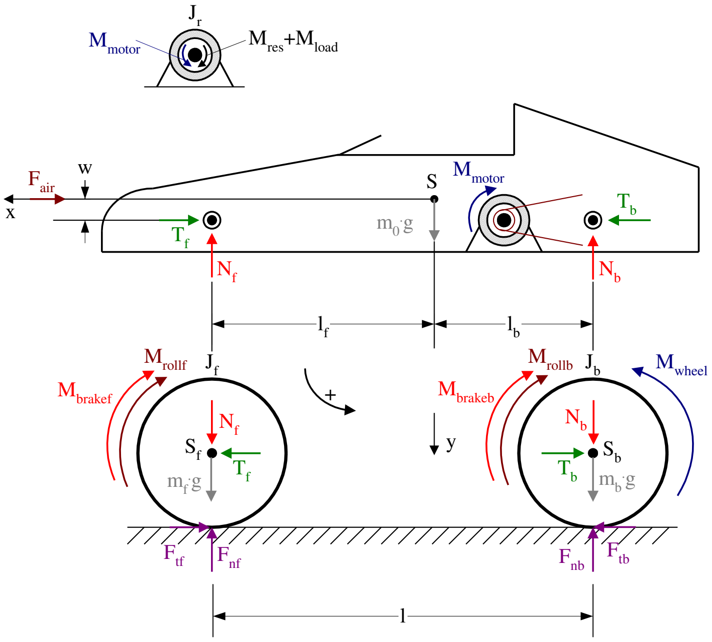


Figure 4

The developed dynamic model for the car together with the forces and torques acting on its different structural units

These units are:

1) The driven back wheels with the rotating machine parts connected to them;

2) The freely rotating front wheels with the rotating machine parts connected to them;

3) The body of the car including the strator of the electric motor;

4) The rotor of the electric motor [4]

In the model dynamic equations are written for the above construction parts taking into consideration the different external and internal forces and moments acting on them on the basis of Newton's laws. Thus, from the above equations, we can calculate not only the external forces and moments but also the internal ones (e.g. the loads on the front and back shafts).

In our model we applied the following assumptions:

- The mass distribution of the car is symmetric to the xy plane in Figure 4. (It was checked with measurement and it is valid with a good approximation.)
- The vehicle is regarded as a rigid system. (This assumption is valid since the stiffnesses of the front and back suspensions are high, and the race track has a flat and smooth surface)
- The resultant of the air resistance force goes through the centre of mass of the vehicle and it is parallel to the vehicle's velocity.

In references [5, 6, 7] other examples for longitudinal vehicle modells are presented.

3.1 The Model of the Electric Motor. Dynamic and Electromagnetic Equations

For the description of the series wound DC motor we applied the following model:

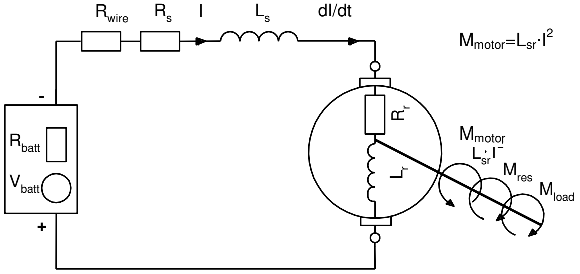


Figure 5

The applied model for the simulation of the series wound DC motor

With the aim of describing the motion of the different structural units of the car (Figure 4) quantitatively, we established dynamics equations taking into account all the external and internal (between the different units) forces and torques.

On the basis of Figure 4 and 5 the dynamic and electromagnetic equations for the different structural units are:

Equations for the freely rotating front wheels with the rotating machine parts connected to them:

$$\text{I.}\quad \sum_i \mathrm{F_{ix}} = \mathrm{T_f} - \mathrm{F_{tf}} = \mathrm{m_f} \cdot \mathrm{a_s} \rightarrow \mathrm{T_f} = \mathrm{m_f} \cdot \mathrm{a_s} + \mathrm{F_{tf}}$$

$$\text{II.}\quad \sum_i \mathrm{F_{iy}} = -\mathrm{F_{nf}} + \mathrm{N_f} + \mathrm{m_f} \cdot \mathrm{g} = 0$$

$$\text{III.}\quad \sum_i \mathrm{M_{i(S_f)}} = -\mathrm{M_{rollf}} + \mathrm{F_{tf}} \cdot \mathrm{R} - \mathrm{M_{brakef}} = \mathrm{J_f} \cdot \mathrm{\varepsilon_f}$$

Equations for the driven back wheels with the rotating machine parts connected to them:

$$\text{IV.}\quad \sum_i \mathrm{F_{ix}} = \mathrm{F_{tb}} - \mathrm{T_b} = \mathrm{m_b} \cdot \mathrm{a_s} \rightarrow T_b = -\mathrm{m_b} \cdot \mathrm{a_s} + \mathrm{F_{tb}}$$

$$\text{V.}\quad \sum_i \mathrm{F_{iy}} = -\mathrm{F_{nb}} + \mathrm{N_b} + \mathrm{m_b} \cdot \mathrm{g} = 0$$

$$\text{VI.}\quad \sum_i \mathrm{M_{i(S_b)}} = -\mathrm{M_{rollb}} - \mathrm{M_{brakeb}} - \mathrm{F_{tb}} \cdot \mathrm{R} + \mathrm{M_{wheel}} = \mathrm{J_b} \cdot \mathrm{\varepsilon_b}$$

Equations for the vehicle body:

$$\text{VII.}\quad \sum_i \mathrm{F_{ix}} = \mathrm{T_b} - \mathrm{T_f} - \mathrm{F_{air}} = \mathrm{m_0} \cdot \mathrm{a_s}$$

$$\text{VIII.}\sum_i \mathrm{F_{iy}} = -\mathrm{N_f} - \mathrm{N_b} + \mathrm{m_0} \cdot \mathrm{g} = 0$$

$$\text{IX.}\quad \sum_i \mathrm{M_{i(S)}} = -\mathrm{M_{motor}} - \mathrm{N_f} \cdot \mathrm{l_f} + \mathrm{N_b} \cdot \mathrm{l_b} + \mathrm{T_f} \cdot \mathrm{w} - \mathrm{T_b} \cdot \mathrm{w} = 0$$

Equation for the rotor of the electric motor (only for rotational motion):

$$\text{X.}\quad \overbrace{\mathrm{L_{sr}} \cdot \mathrm{I}^2}^{\mathrm{M_{motor}}} - \mathrm{M_{res}} - \mathrm{M_{load}} = \mathrm{J_r} \cdot \varepsilon_{mot} = \mathrm{J_r} \cdot \mathrm{i_{12}} \cdot \mathrm{\varepsilon_b} \qquad \left(\mathrm{\varepsilon_b} = \mathrm{\varepsilon_{mot}} \cdot \frac{1}{\mathrm{i_{12}}}\right)$$

Connection between the driving torque on the back shaft ($\mathrm{M_{wheel}}$) and the loading torque on the electric motor:

$$\text{XI.}\quad M_{\text{wheel}} = -\eta \cdot i_{12} \cdot M_{\text{load}} \qquad \left(i_{12} = \frac{z_2}{z_1}\right)$$

Electromagnetic equation for the electric motor:

$$\text{XII.}\quad V_{\text{batt}}\text{-}(R_s + R_r + R_{\text{batt}} + R_{\text{wire}}) \cdot I\text{-}\big(L_s(I) + L_r(I)\big) \cdot \frac{dI}{dt}\text{-}L_{sr}(I) \cdot \omega \cdot I = 0$$

The parameters in the above equations are:

M_{wheel}: magnitude of the torque exerted by the chain drive on the back shaft

M_{motor}: magnitude of the torque exerted by the strator of the motor on its rortor

$M_{rollf,}$ M_{rollb}: magnitude of the rolling resistance torque on the front and back wheels

F_{air}: resultant of air resistance force

$F_{tf,}$ F_{tb}: magnitude of the force exerted by the road on the front and back wheels in tangential direction

$F_{nf,}$ F_{nb}: magnitude of the force exerted by the road on the front and back wheels in normal direction

T_f, T_b: load on the front and back shaft in tangential direction

N_f, N_b: load on the front and back shaft in normal direction

S_f, $S_{b,}$ S: centre or gravity of the front and back wheels and the whole vehicle

M_{brakef}, M_{brakeb} : braking torques on the front and back wheels

η: efficiency of the chain drive

$z_{1,}$ z_2: number of teeth on the driver and driven sprockets

i_{12}: gear ratio in the chain drive

ε_{mot}: angular acceleration of the rotor of the motor

ε_b: angular acceleration of the driven back wheels

R_{batt}: the battery's internal resistance

V_{batt}: the battery's electromotive force

R_{wire}: resultant resistance of the wires connecting the battery to the motor

R_s, R_r: electric resistances of the rotor and stator windings

I: the intensity of current flowing through the motor

L_s: self dynamic inductance of the stator winding

L_r: self dynamic inductance of the rotor winding

L_{sr}: mutual dynamic inductance

M_{res}: sum of the bearing resistance and brush friction torques on the rotor of the motor

M_{load}: loading torque on the rotor of the motor [8]

l: distance between the front and back shafts in tangential direction

l_f, l_b: distance of the centre of mass of the vehicle from the front and back shaft respectively in tangential direction

w: distance of the centre of mass of the vehicle from the front and back shafts in normal direction

m_0: mass of the vehicle body including the driver

m_f, m_b: mass of the front and back wheels with the rotating machine parts connected to them

J_f, J_b: moment of inertia of the front and back wheels with the rotating machine parts connected to them

J_r: moment of inertia of the rotor of the motor

R: effective wheel radius

The electromagnetic and dynamic characteristic of the applied electric motor (R_{batt}, V_{batt}, R_{wire}, R_s, R_r, L_s, L_r, L_{sr}, M_{res}, J_r) were measured previously by the authors [9]. Some of the other vehicle parameters (M_{rollf}, M_{rollb}, l, l_f, l_b, w, m_0, m_f, m_b, J_f, J_b, R) were also measured, while the remaining ones were approximated on the basis of literature data.

3.2 Additional Formulas

Tyre slip:

During acceleration the longitudinal slip of the front and back tyres is calculated as:

$$slip = \frac{\omega \cdot R - v_s}{\omega \cdot R} \cdot 100\%$$

while during deceleration,

$$slip = \frac{\omega \cdot R - v_s}{v_s} \cdot 100\%$$

where ω is the wheel's angular speed, R is the effective wheel radius and v_s is the speed of the vehicle. [4]

Coefficient of friction:

The coefficient of friction (the ratio of the tangential and normal force exerted by the road on the tyres) was calculated by the Pacejka formula [2, 3, 4, 9]:

$$\frac{F_{tb}}{F_{nb}} = \mu = c_1 \cdot \sin\left(c_2 \cdot \operatorname{atan}\left(c_3 \cdot slip - c_4 \cdot \left(c_3 \cdot slip - \operatorname{atan}(c_3 \cdot slip)\right)\right)\right)$$

Where c_1,c_2, c_3 and c_4 are constants based on the tire experimental data. The constants above are referred to as the stiffness, shape, peak and curvature coefficients respectively. [1, 7]

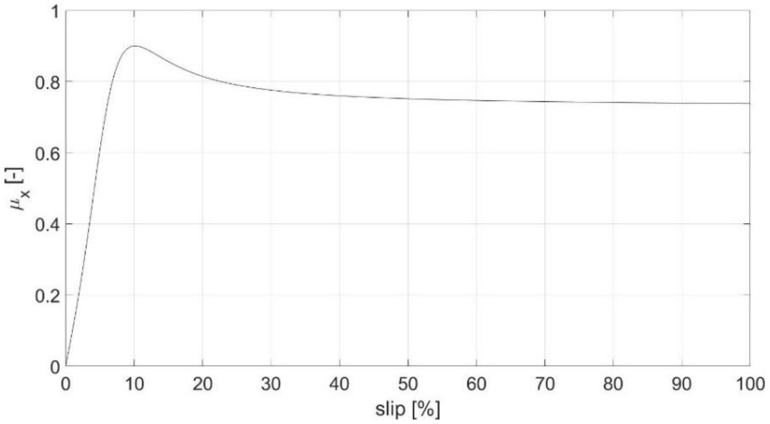


Figure 6

The graph of the Pacejka formula applying values c_1=0.9, c_2=1.4, c_3= 7.936, c_4= -10

Air resistance force:

The magnitude of air resistance force is calculated as [3]:

$$F_{air} = \frac{1}{2} \cdot C \cdot A \cdot \rho \cdot v_s^2$$

where C and A are the coefficient of air resistance and maximum cross section area of the car normal to its moving direction, ρ is the air density and v_s is the speed of the car relative to air [4].

Rolling resistance forces and torques:

The magnitude of rolling resistance forces and torques in case of the front and back wheels of the vehicle are calculated as [9]:

$$F_{rollf} = \mu_{rollf} \cdot F_{nf}, \qquad F_{rollb} = \mu_{rollb} \cdot F_{nb}$$

$$M_{rollf} = R \cdot F_{rollf}, \qquad M_{rollb} = R \cdot F_{rollb}$$

Where μ_{rollf} and μ_{rollb} are the coefficients of rolling resistance for the front and back wheels at given F_{nf} and F_{nb} normal forces exerted by the road on them. During intensive speed-up or braking the load on the back or front wheels increases significantly. We face the same situation applying a hevyer driver instead of a lighter one. Because of that the dependence of rolling resistance coefficient on vertival load ($\mu_{roll}(F_n)$) was taken into consideration in the simulation program applying lookup tables (Figure 9 and 11). On the other hand at 3 bar tyre pressure – which is the value regularly applied in a race situation – the dependence of rolling resistance coefficient on vehicle speed is negligible in the whole speed range of the car [17]. The procedure for the measurement of rolling resistance coefficient as a function of tyre pressure and normal (vertical) load is presented in references [10, 11] in details.

Bearing resistance torques:

On the basis of our simulations, it can be stated, that the contribution of bearing resistance torque to the simulation results – in the case of the front and rear shafts of the vehicle – is negligible compared to the other factors. Additionally, taking it into consideration the running time of the simulation program increases significantly. Thus we neglected this factor. However, at the shaft of the rotor of the electric motor the bearing resistance and brush friction torques can not be neglected, thus we built it into our simulation program in the form of a look-up table (M_{res}, Figure 8).

4 The Simulation Program

On the basis of our vehicle model we developed a simulation program in Matlab which is able to calculate the vehicle dynamics functions of a vehicle from its technical data. The program's block diagram is shown in *Figure 7*. [4]

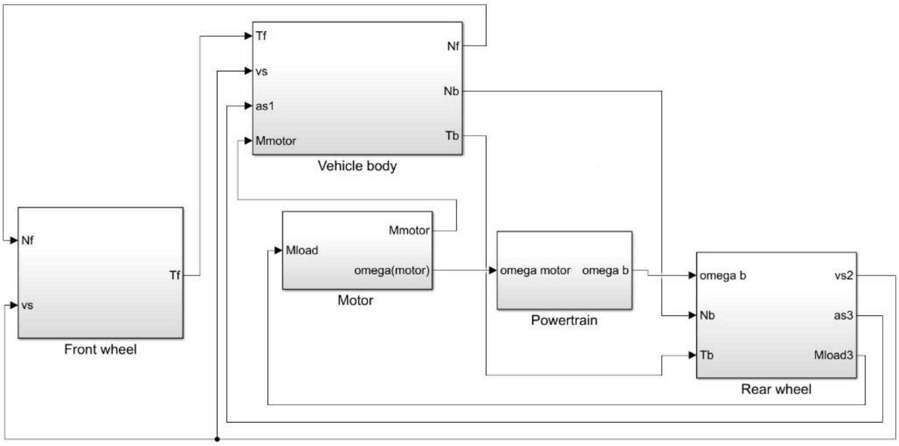


Figure 7
The block diagram of the simulation program in MATLAB Simulink

The simulation program is built up of five modules in accordance with our dynamics model. The modules are:

- vehicle body
- front wheels and connected rotating machine parts
- back wheels and connected rotating machine parts
- motor
- powertrain

In the following we describe the structure and operation of the above modules in detail.

In references [12, 13, 14, 15, 16] other examples for vehicle dynamics simulation programs are presented.

4.1 The Motor

The simulation module for the applyed series wound DC motor [8] is presented in Figure 8.

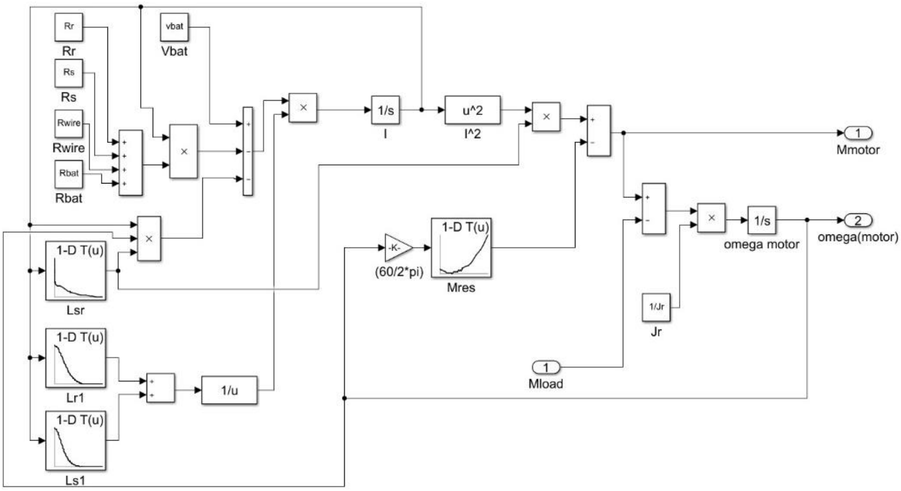


Figure 8

The motor simulation program module

The self and mutual dynamic inductances – together with the resistance torque of the shaft of the motor – are given as look-up tables in the program module. All the other parameters are given as constants. The program module is based on equations X and XII in Section 3. A more detailed description of the motor simulation program module – together with the description of the measurement process of the different input characteristics – can be found in reference [8].

In references [17, 18, 19] other examples for the modeling of a series wound DC motor are presented.

4.2 Front Wheels and Connected Rotating Machine Parts

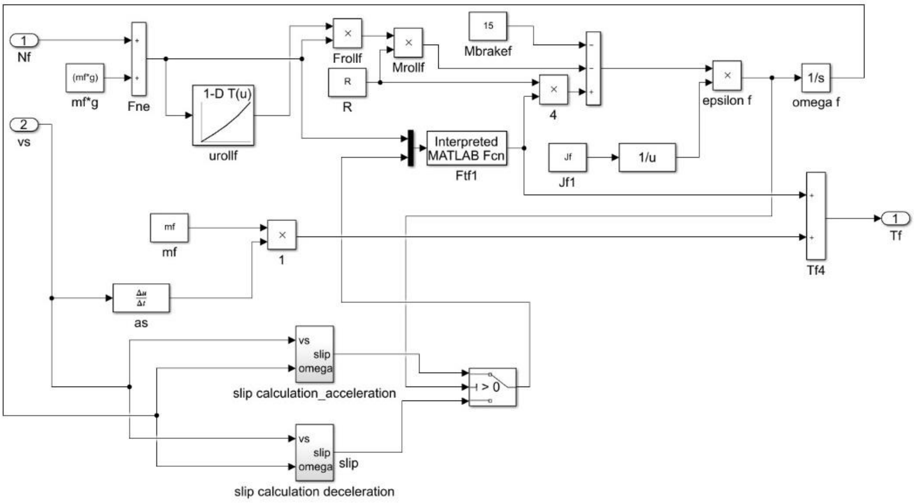


Figure 9
Simulation program module describing the front wheels and connected rotating machine parts

The program module – describing the front wheels together with the connected rotating machine parts – calculates the load on the front shaft in tangential direction (T_f) from the load on it in normal direction (N_f) and also from the velocity of the car (v_s) on the basis of equations I, II and III (section 3.1). Parameters N_f and v_s are calculated by the program modules describing the vehicle body and back wheels respectively (Figures 14 and 10). The module calculates the angular velocity (omega(f)) of the front wheels independent of vehicle velocity (v_s), and then the slip of the tyres from omega(f) and v_s in case of acceleration and deceleration applying different formulas.

4.3 Back Wheels and Connected Rotating Machine Parts

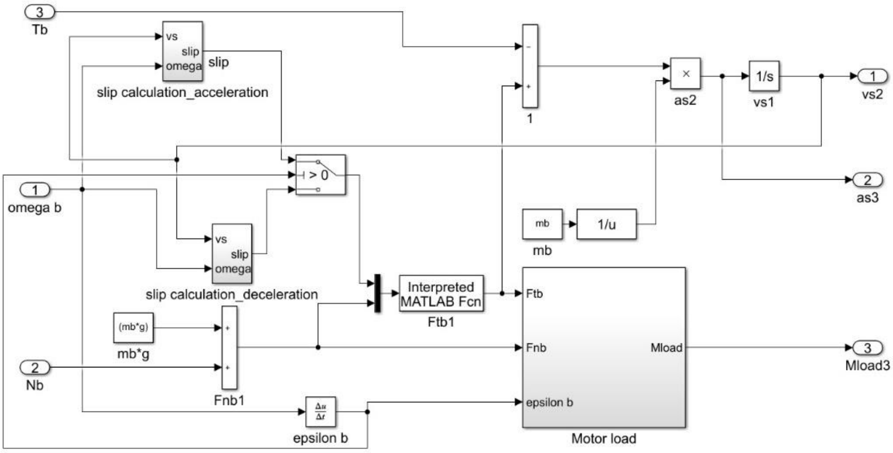


Figure 10

Simulation program module describing the back wheels and connected rotating machine parts

The program module – describing the back wheels with the rotating machine parts connected to them – calculates the load on the motor (M_{load}) together with the acceleration (a_s) and velocity (v_s) of the vehicle from the load on the back shaft in tangential (T_b) and normal (N_b) direction and also from the angular velocity of the back wheels (omega(b)) on the basis of equations IV, V and VI (Section 3.1).

Inside the module, there are „sub-modules" for the calculation of the load on the motor (M_{load}, Figure 11) and tyre slip in case of acceleration (Figure 12) and deceleration (Figure 13). The load on the motor is calculated from the forces exerted by the road on the back wheels in tangential (F_{tb}) and normal (F_{nb}) direction and also from the angular acceleration of the wheels (ε_b).

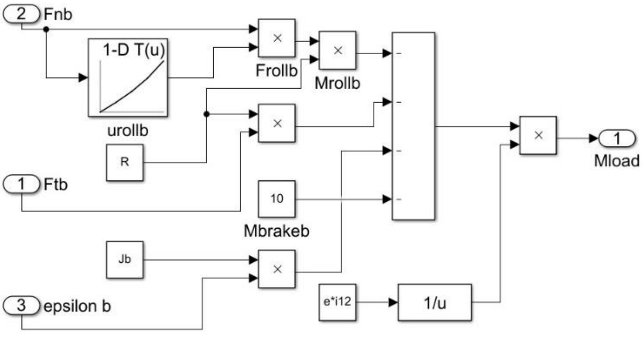


Figure 11

Program module for the calculation of the load on the motor

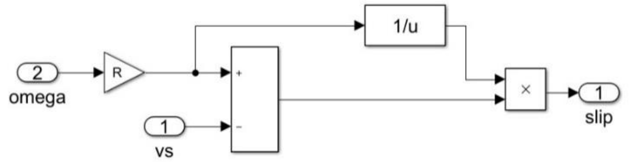


Figure 12
Program module for the calculation of tyre slip during acceleration

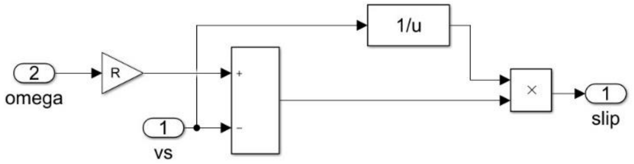


Figure 13
Program module for the calculation of tyre slip during deceleration

4.4 Vehicle Body

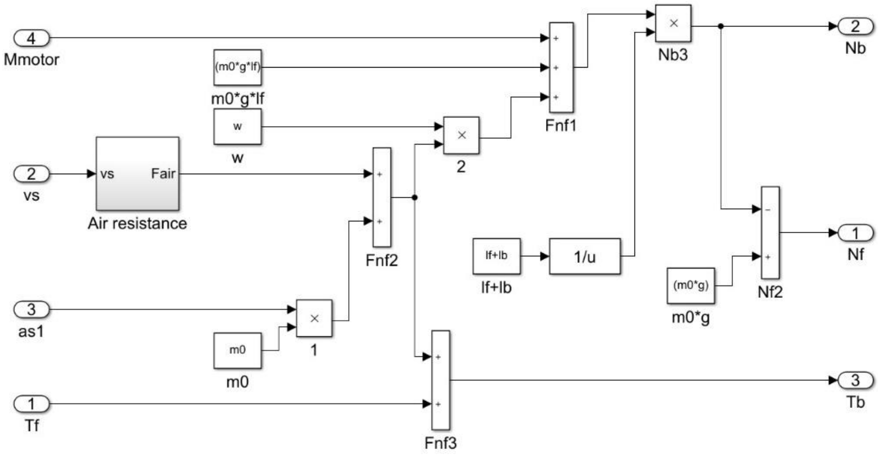


Figure 14
Program module for the description of vehicle body

The program module – describing the vehicle body – calculates the load on the back shaft in normal and tangential direction (N_b, T_b) and also the one on the front shaft in normal direction (N_f) from the torque that the motor exerts on the vehicle body (M_{motor}) and from the velocity (v_s) and acceleration (a_s) of the vehicle, and also from the load on the front shaft in the tangential direction (T_f) on the basis of equations VII. VIII. and IX. (section 3.1). The load on the back shaft in normal direction is calculated by the formula

$$N_b = \frac{(T_b - T_f) \cdot w + m_0 \cdot g \cdot l_f + M_{motor}}{l_b + l_f}$$

which is derived from equations VIII. and IX. Inside the module, there is a "sub-module" for the calculation of the magnitude of air resistance force (Figure 15).

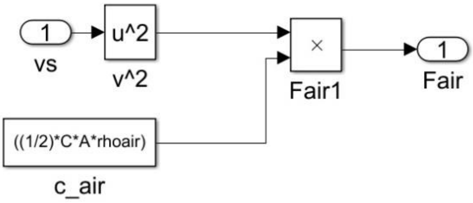


Figure 15
Program module for the calculation of air resistance force

4.5 Powertrain

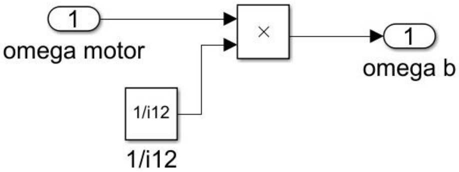


Figure 16
Program module for the description of the power train

The module calculates the angular velocity of the back wheel (omega(b)) from the angular velocity of the motor (omega(motor)) and the gear ratio (i_{12}) of the chain drive.

5 Output Dynamics Functions of the Program

The program generates the following quantities as a function of time:

- The acceleration, velocity and covered distance of the vehicle
- The angular velocity and acceleration of the front and back wheels
- The forces the road exerts on the tyres in tangential and normal direction
- Front and back shafts loading in tangential and normal direction
- Rolling resistance torques
- Air resistance force

- Tyre slip
- The intensity of the current flowing through the motor
- The torque and angular speed of the motor

It is important to emphasise that the simulation program stops with an error message if we give zero initial values in the integrators, which are: calculate the angular speed of the motor and the velocity of the car in the “Motor” and “Rear wheel” program modules. So small initial values have to be given to these quantities.

Figure 17 shows the (tractive) force exerted by the road on the back wheels in tangential direction (F_{tb}) vs. time applying different gear ratios in the chain drive. At gear ratio 8 a sharp peak appears on the graph of the tractive force-time function. It is happening because meanwhile, the car is starting, the driven wheels are spinning up, and then – with increasing vehicle speed – the wheels stick to the road, which results in the sharp increase of the tactive force. The values of all the other parameters are the same in the presented three cases. [4]

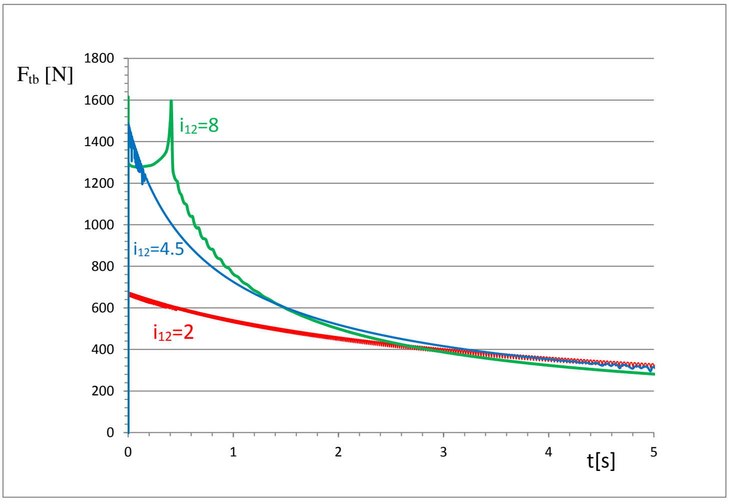


Figure 17

Tractive force on the back wheels vs. time [4]

Figure 18 shows the velocity vs. time functions of the car applying the same gear ratios as the ones in Figure 17.

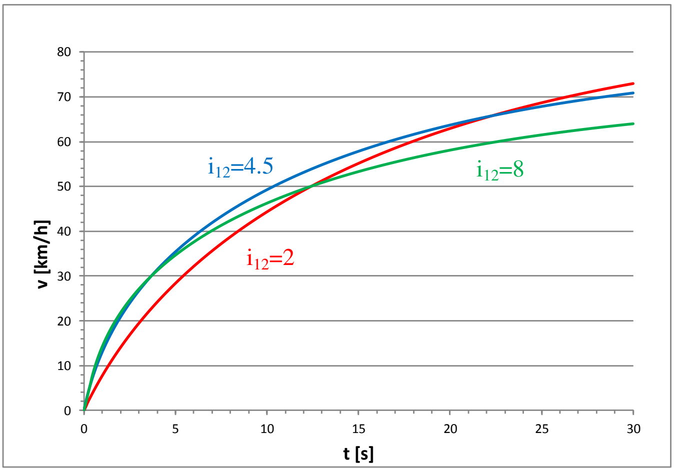


Figure 18

Velocity vs. time functions of the car applying different gear ratios [4]

Figure 19 presents the forces exerted by the road on the front and back wheels of the car in normal direction vs. time.

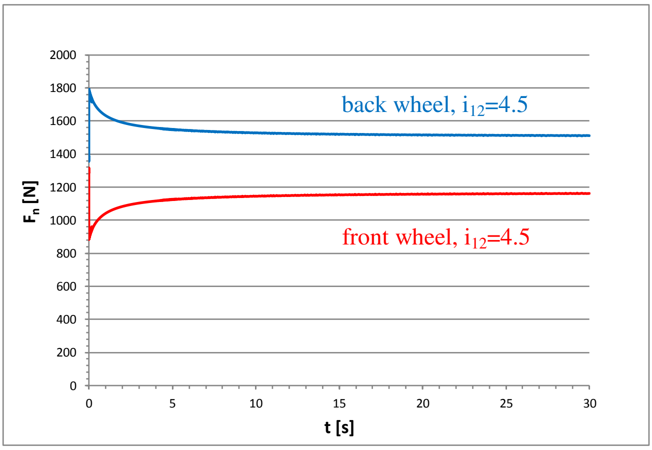


Figure 19

Forces exerted by the road on the front and back wheels of the car in normal direction vs. time [4]

Figure 20 presents the simulated velocity vs. time function of the car in case of 5s acceleration and consequently that 5s deceleration.

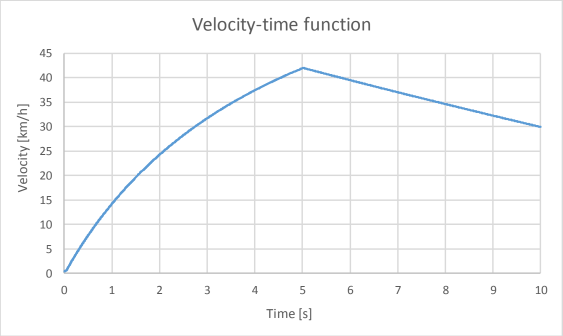


Figure 20

Velocity-time function of the car during during acceleration and braking

The supply voltage and braking torque belonging to the above simulation are presented in Figure 21.

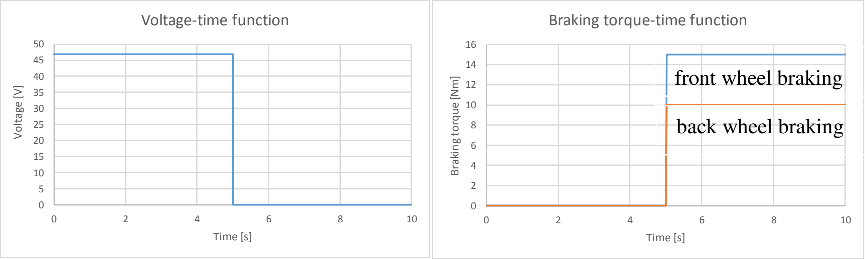


Figure 21

Supply voltage and braking torque as a function of time

Conclusions

A vehicle dynamic simulation program has been developed which is capable of the complete simulation of the motion of a vehicle on a linear track. The program has a modular structure, so the different construction parts (motor, vehicle body, drive train, front and back wheels and connected rotating machine parts) are modeled and simulated in separate program modules. The program takes into account all the factors which have a significant effect on the motion of the vehicle and can calculate all its vehicle dynamic functions.

In the recent work, the program was applied for the simulation of the motion of a prototype race car driven by a serious wound DC motor. The calculated dynamic functions give a good approximation of the real ones, but for the precise quantitative verification of the program, we intend to measure the values of all the missing input technical data, and then to perform precise telemetry measurements on the car in the near future.

Acknowledgement

"The research was financed by the Thematic Excellence Programme of the Ministry for Innovation and Technology in Hungary (ED_18-1-2019-0028), within the framework of the (Automotive Industry) thematic programme of the University of Debrecen."

References

[1] Hans B. PACEJKA, Igo BESSELINK: Tire and Vehicle Dynamics (Thirdedition) – Published by Elseiver Ltd. (2012) ISBN 978-0-08-097016-5

[2] Jörsen REIMPELL, Jürgen W. BETZLER, BÁRI Gergő, HANKOVSZKI Zoltán, KÁDÁR Lehel, LÉVAI Zoltán, NAGYSZOKOLYAI Iván: Gépjárműfutóművek I. (2012) ISBN 978-963-279-606-2

[3] Bernd HEISSING, Metin ERSOY: Chassic Handbook (2011) ISBN 978-3-8348-0994-0

[4] SZIKI Gusztáv Áron, HAJDU Sándor, SZÁNTÓ Attila: Vehicle dynamics modelling of an electric driven race car, Proceedings of the 3rd International Scientific Conference on Advances in Mechanical Engineering, ISBN 978-963-473-917-3, 2015

[5] KOST, F. (2014) Basic principles of vehicle dynamics. Brakes, Brake Control and Driver Assistance Systems, 12-27, doi:10.1007/978-3-658-03978-3_2

[6] Bengt JACOBSON et al: Vehicle Dynamics. Department of Applied Mechanics, Chalmers University of Technology. 2016

[7] Reza N. JAZAR: Vehicle Dynamics: Theory and Application. Springer. 2008. ISBN: 978-0-387-74243-4

[8] SZIKI Gusztáv Áron; SARVAJCZ Kornél; KISS János; GÁL Tibor; SZÁNTÓ Attila; GÁBORA András; HUSI Géza: Experimental investigation of a series wound dc motor for modeling purpose in electric vehicles and mechatronics systems. In: MEASUREMENT (ISSN: 0263-2241) 109: pp. 111-118 (2017) IF: 1.742

[9] Dr. ILOSVAI Lajos Prof. Emeritus (2013) Járműdinamika

[10] F. GRAPPE, R. CANDAU, B. BARBIER, M. D. HOFFMAN, A. BELLI, J.-D. ROUILLON: Infuence of tyre pressure and vertical load on coefficient of rolling resistance and simulated cycling performance, ERGONOMICS, 1999, VOL. 10, 1361-1371

[11] SZESZÁK B., SÜTŐ T., Nagyné KONDOR R., SZÍKI G., JUHÁSZ G.: Analysis of the Rolling Resistance of Pneumobils for Vehicle Dynamics Modelling Purpose. Proceedings of the 2nd Agria Conference on Innovative

Pneumatic Vehicles ACIPV 2018 / ed. Pokorádi László, pp. 17-20, ISBN 978-963-449-089-0

[12] József POLÁK and István LAKATOS. Examination of drive line mathematical model. Machine design 8(1):33-36, 2016

[13] József POLÁK and István LAKATOS. Efficiency optimization of electric permanent magnet motor driven vehicle. Machine design 7(1):11-14, 2015

[14] SHAKOURI, P., LAILA, D. S., ORDYS, A., & ASKARI, M. (2010) Longitudinal vehicle dynamics using Simulink/Matlab. UKACC International Conference on Control 2010, doi:10.1049/ic.2010.0410

[15] SZAKÁCS T.: Pneumatic modelling of a pneumobil. In: Pokorádi, László (szerk.) Proceedings of the 2nd Agria Conference on Innovative Pneumatic Vehicles ACIPV 2018 Eger, Magyarország : Óbudai Egyetem, (2018) pp. 25-30, 6 p.

[16] SZAKÁCS T.: Modelling and Validation of a Pneumobil. In: Pokorádi, László (szerk.) Proceedings of the 3rd Agria Conference on Innovative Pneumatic Vehicles – ACIPV 2019 Eger, Magyarország : Óbudai Egyetem Mechatronikai és Járműtechnikai Intézet (2019) pp. 31-35, 5 p.

[17] Farhan A. Salem: Dynamic Modeling, Simulation and Control of Electric Machines for Mechatronics Applications. International Journal of Control, Automation and Systems Vol. 1, No. 2, April 2013 ISSN 2165-8277 (Print) ISSN 2165-8285 (Online)

[18] Zeina BITARA, Samih AL JABIA, Imad KHAMISB (2014): Modeling and Simulation of Series DC Motors in Electric Car, The International Conference on Technologies and Materials for Renewable Energy, Environment and Sustainability, TMREES14, Energy Procedia 50 (2014) 460-470

[19] P. ZÁSKALICKÝ (2006): Modelling of a serial wound DC motor supplied by a semi controlled rectifier, Advances in Electrical and Electronic Engineering (AEEE) Vol. 5 (2006) 110-113

Analysis of Edge Detection on Compressed Images with Different Complexities

Vladimir Maksimović[1], Branimir Jakšić[1], Mile Petrović[1], Petar Spalević[1], Mirko Milošević[2]

[1] Faculty of Technical Sciences, University of Pristina in Kosovska Mitrovica, Knjaza Milosa 7, 38220 Kosovska Mitrovica, Serbia,
vladimir.maksimovic@pr.ac.rs, branimir.jaksic@pr.ac.rs, mile.petrovic@pr.ac.rs, petar.spalevic@pr.ac.rs, mirko.milosevic@viser.edu.rs

[2]School of Electrical and Computer Engineering of Applied Studies, Vojvode Stepe 283, 11000 Belgrade, Serbia

Abstract: This paper provides edge detection analysis on images, which consist of different numbers of details (small, medium and high number of details) and which are compressed by different compression algorithms - JPEG, JPEG2000 and SPIHT. Images from the BSD (Berkeley Segmentation Database) database were used and compressed with different number of bits per pixel. The analysis was performed for five edge detectors: Canny, LoG, Sobel, Prewitt, and Roberts. The fidelity of the detected edges was determined using the objective measures Figure of Merit (FOM), F measure and Performance Ratio (PR), where the reference value was taken from the GroundTruth image. Based on the results presented in the tables, it can be concluded that edge detection behaves differently depending on the number of bits per pixel and applied compression algorithm, as well as, the number of details in the image. Roberts operator has been proven to be the best solution, when it is necessary to perform better edge detection over compressed images with small a number of details, but Canny shows better results for images with a high number of details.

Keywords: edge detection; compression; image processing; Figure of Merit (FOM); F measure; Performance Ratio (PR); image complexity; bit per pixel (BPP)

1 Introduction

In today's multimedia systems, it is almost impossible to find a system that does not use image, video or audio compression. However, the development of technology has also brought an increasing use and processing of images, from use in daily life to those more serious professional uses such as image analysis in medicine, sensor networks, smart and security systems, television and so on. An uncompressed image requires more storage space for storage and processing, as well as, transmission via telecommunication channels. Considering this fact, there is a great deal of interest among researchers regarding image processing and compression. The size of the image can be large so that it is very impractical to store or transfer, especially when it comes to real-time image processing systems. For this reason, many image compression methods have been developed, but we can divide them all into lossy and lossless ones [1-3].

Depending on the need, various compression techniques and compression algorithms are applied, and as a result, the most popular are JPEG and JPEG2000. JPEG standard compression is based on Discrete Cosine Transform (DCT), while JPEG2000 compression is based on Discrete Wavelet Transform (DWT) [4-9]. Also, the compression algorithm based on Embedded Zero Tree Wavelet (EZW) is the SPIHT algorithm [10-12]. As mentioned, compression and coding techniques are used in many systems, i.e. where image processing is performed, so there are techniques for medical images [13-17], radar images [18-20], satellite images [21-23] and for many other smart systems combining different compression and coding techniques [24-26]. Image processing is an integral part of machine learning and artificial intelligence, where there are classifiers and neural network models that can be used as in [27], [28]. Also, the mathematical models presented in [29], [30] provide ideas for improving the algorithms for estimating image complexity used in this paper.

We are also witnessing an increase in the use of smart networks, the use of artificial intelligence to analyze, collect and process data. Such systems are mainly based on image processing and data processing, where the main processes are the extraction of a particular object from the scene, where edge detection and segmentation play an important role [31-34]. However, all of this gain particular weight and interest with the emergence and implementation of such systems on devices like Raspberry Pi and Arduino, which very often use real-time image processing, object detection and segmentation [35-39]. Many techniques and enhancements have been proposed to maximize the quality of edge detection and segmentation [39-43]. Given that the resolutions and image quality are increasing, thus occupying a large storage space, it is important to do compression so as not to impair the quality. Compression will affect edge detection, as examined in [44] using a wavelet transform, which underlies some compression algorithms, as well as facial recognition [45-47]. Therefore, the effect of compression on edge detection is presented in [44] where the authors examined only the influence of wavelet-based compression. The authors in [45-47] examined the effect of compression on face recognition using the JPEG and JPEG2000 algorithms, while the effect on edge detection was not examined. In this paper, the idea is to examine the impact of compression on edge detection using the most common compression algorithms.

The rest of the paper is divided as follows: Section 2 explains the system model, that is, the basic setting on which a detailed edge detection analysis was made. The images that were used for analysis are given, followed by tabulated values obtained during compression using different algorithms. In this section can be seen the method used by the authors to perform the analysis. Section 3 presents the obtained results of edge detection for five edge detectors over compressed images using different compression algorithms. The tables show three objective measures, and based on the results the discussion was made. In Section 3, there are sub sections for each operator. Finally, the conclusion of this paper is given, as well as the direction of future research.

2 System Model

This paper analysis the impact of JPEG, JPEG2000, and SPIHT algorithm on edge detection, where images are compressed with different number of bits per pixel (BPP), namely: 0.1, 0.3, 0.5, 1, 1.5, and 3 BPP. Images from the BSD. Used images are from the BSD database with the corresponding GroundTruth [48]. The images were selected to meet the three complexity criteria of small, medium and high complexity [49], that is, each image consists of a different level of detail: small, medium, and high level of details [49]. Table 1 shows the obtained values on the basis of which are the selected images from the database BSD, which meet the defined criteria.

Complexity in an image shows information about how much details exists in that image, and this can be observed for both static images and video formats. The simplest way of determining complexity is on the basis of observer’s visual assessment. However, it is not an objective measure to confirm the credibility of that assessment [50-52]. Since this paper looks at the effect of compression on edge detection, there are also methods that measure image complexity based on compression and thus make a link between compression, quality and complexity. One way of doing this is shown in [50]. JPEG, JPEG200 and SPIHT algorithms are based on the DCT and DWT techniques, so the number of details was calculated by making DCT and DWT on the high-frequency components (details), which are divided into four quadrants, along both directions (x and y). After that, the mean absolute value of the amplitude of the components belonging to the quadrants is calculated according to [49]: DCT in quadrant 1 (DCTD); DCT in quadrants 2 and 3 (DCTM); DWT in quadrant 1 (DWT); DWT in quadrants 2 and 3 (DWTM).

Edge detection and analysis were performed on the selected and compressed images for five edge detectors, namely: Canny, LoG, Sobel, Prewitt, and Roberts. Gradient and Laplace edge detection algorithms were written in Matlab, while image compression was performed using VcDemo. So, first, the images extracted from the BSD database with the corresponding GroundTruth were selected to satisfy the criteria in [49] using the technique from that paper. After that, the images were compressed in VcDemo using JPEG, JPEG2000 and SPIHT algorithm with different BPP. In the end, edge detection over compressed images was performed using five operators and objective measures are calculated in Matlab.

Table 1

Complexity criteria

	Images	DCTD	DCTM	WVTD	WVTM
Criterion L	**#238011**	**<2**	**<3.5**	**<0.8**	**<1.2**
		0.75	1.69	0.17	0.44
Criterion M	**#245051**	**3-4**	**4.5-6.5**	**1.4-1.8**	**2-2.8**
		3.11	7.02	1.12	2.09
Criterion H	**#231015**	**>4.9**	**>9**	**>1.9**	**>3.9**
		5.48	10.97	2.14	7.29

The authors have created a repository [53] containing used images for analysis, obtained images and codes.

Objective measures that were used are:

F measure (F1 score) which ranges from 0 to 100 and can be calculated [54]:

$$F = \frac{2 \times \mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision \times \mathrm{Re}\,call} \times 100 \tag{1}$$

F is within the limits of 0 ≤ F ≤ 1, ideally, F is equal to 1. The precision, also known as the positive predictive value is calculated [31]:

$$\mathrm{Pr}\,ecision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FP)} \tag{2}$$

while Sensitivity (Recall):

$$Sensitivity = \frac{TruePositive(TP)}{TruePositive(TP) + FalseNegative(FN)} \tag{3}$$

Where is TP - True Positive, pixels correctly segmented as foreground; FP - False Positive, pixels falsely segmented as foreground; TN - True Negative, pixels correctly detected as background and FN - False Negative, pixels falsely detected as background. Figure of Merit (FoM) which also ranges from 0 to 100, respectively represents the percentage value and can be calculated [55]:

$$FoM = \frac{1}{\max\{I_d, I_i\}} \sum_{k=1}^{I_d} \frac{1}{1 + \delta e^2(k)} \times 100 \tag{4}$$

where Id is the number of points on the detected edge, and Ii is the number of points on the ideal edge, represents the distance between the detected edge and the ideal edge, and is scaling constant and is usually 1/9.

Performance Ratio (PR) which ranges from 0 to infinite [56]:

$$PR = \frac{TrueEdge(EdgePixelsIdentifiedAsEdges)}{FalseEdges(NonEdgePixelsIdentifiedAsEdges) + (EdgePixelsIdentifiedAsNon - EdgePixels)} \times 100 \quad (5)$$

Table 2 shows the Peak Signal to Noise ratio (PSNR) [49] values which show how the number of bits per pixel (BPP) is affecting image compression. It can be seen from Table 2 that the increase in BPP contributes significantly to image quality, especially with JPEG compression, when the number of image details is small. JPEG2000 and SPIHT obtained similar results, but the number of details noticeably affects the compression.

Table 2

PSNR values for three compression algorithms with different BPP and level of details

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	28.2	40.3	52.7	62.2	61.8	61.8
	JPEG2000	36.7	42.5	46.4	52.4	52.4	52.4
	SPIHT	36.6	42.7	46.6	53.5	58.5	70.7
MD	**JPEG**	20.7	26.5	28.9	31.7	41.5	56.4
	JPEG2000	24.3	28.6	31.6	37.5	42.1	50.8
	SPIHT	24.0	28.4	31.7	37.5	42.1	53.9
HD	**JPEG**	19.1	22.8	24.7	27.6	28.8	50.4
	JPEG2000	21.4	24.2	26.3	30.3	33.7	43.6
	SPIHT	21.3	24.2	26.1	30.1	33.6	43.5

Fig. 1, Fig. 2, and Fig. 3 show a compressed image for a different number of BPP and small number of details (SD) when using JPEG, JPEG2000, and SPIHT compression, respectively. In Fig. 4, Fig. 5 and Fig. 6, images with medium level of details compressed with JPEG, JPEG2000, and SPIHT compression and different BPP are shown, respectively. When it comes to a high number of details in an image (HD), using the JPEG, JPEG2000 and SPIHT algorithm, the resulting compressed images for different BPP are shown in Fig. 7, Fig. 8 and Fig. 9, respectively.

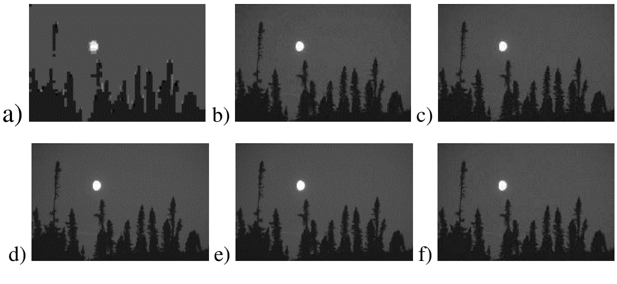


Figure 1

SD image with JPEG compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

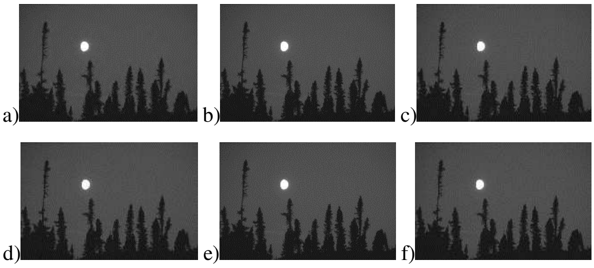


Figure 2
SD image with JPEG2000 compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

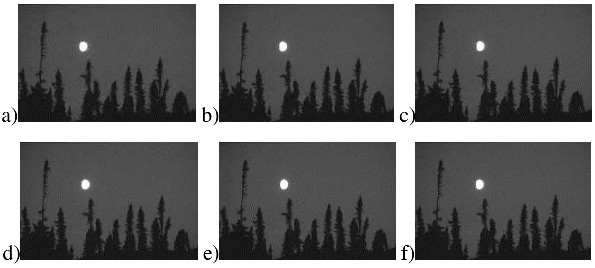


Figure 3
SD image with SPIHT compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

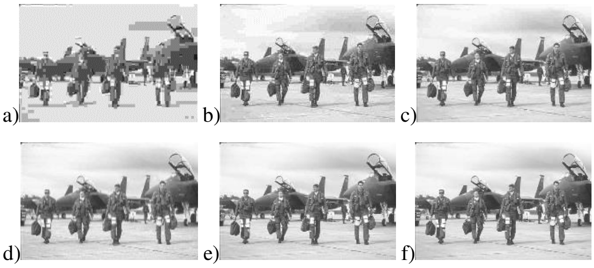


Figure 4
MD image with JPEG compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

From the shown figures can be seen that the quality is usable by applying all kinds of compression for all levels of detail in the image. However, degradation is greatest with a high number of details and at lower BPP, which is confirmed by results in Table 1.

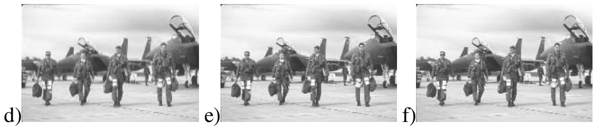


Figure 5
MD image with JPEG2000 compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

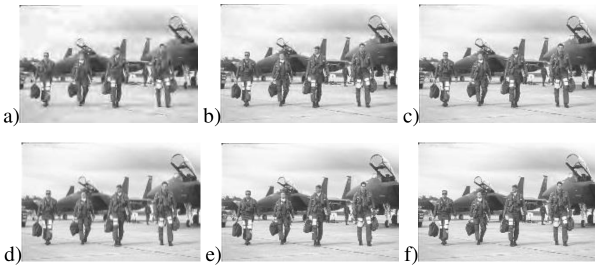


Figure 6
MD image with SPIHT compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

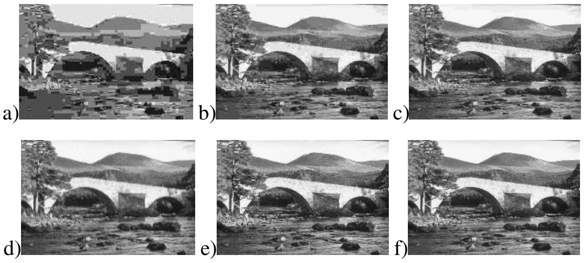


Figure 7
HD image with JPEG compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

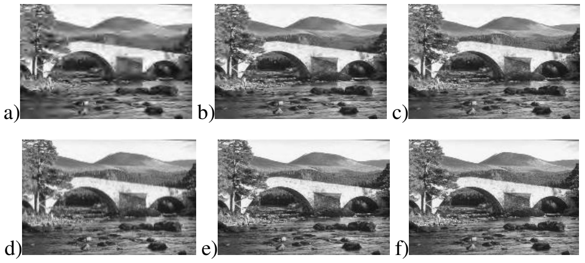


Figure 8
HD image with JPEG2000 compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

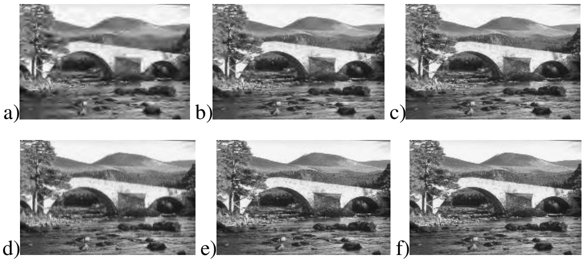


Figure 9

HD image with SPIHT compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

Thus, lower BPP, compression algorithm and number of details significantly affect image quality. However, the main aim of this paper is to examine the impact of edge detection over these images, i.e. how much all of this affects the quality of the detected edge. Table 3 shows the F, FOM and PR values obtained by applying five edge detection operators over images with different level of details. Based on these results, it can be seen that the best values are obtained when the number of details in the image is small. The Roberts operator obtained the best values and the LoG the worst when the number of details in the image is small and medium, while at a high number of details, Canny obtained higher values than the others. In order to present these results visually, Fig. 10, Fig. 11 and Fig. 12 show an image with small, medium and high number of details over which edge detection was performed using five operators.

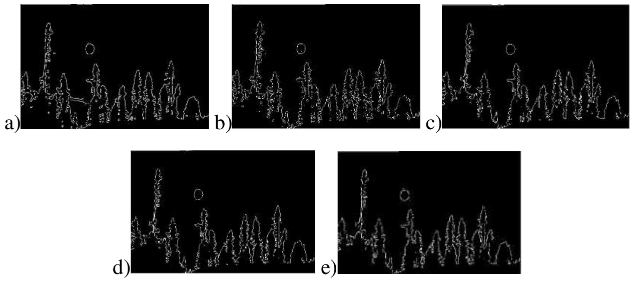
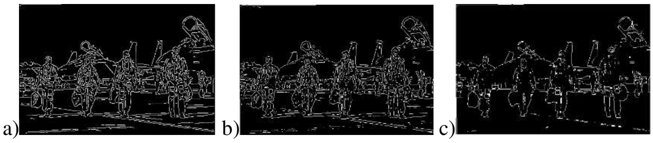


Figure 10

SD image: a) Canny, b) LoG, c) Prewitt, d) Sobel, e) Roberts



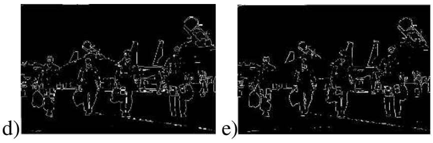


Figure 11
MD image: a) Canny, b) LoG, c) Prewitt, d) Sobel, e) Roberts

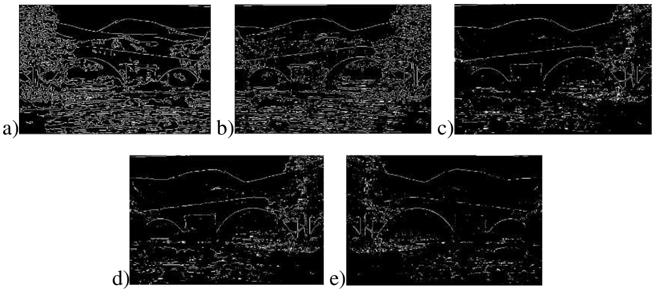


Figure 12
HD image: a) Canny, b) LoG, c) Prewitt, d) Sobel, e) Roberts

Table 3
F, FOM and PR values obtained by applying different edge detectors

	Operator	**F**	**FOM**	**PR**
SD	Canny	35.11	89.39	27.54
	LoG	32.40	90.07	24.38
	Prewitt	35.15	89.40	26.36
	Sobel	34.07	89.49	26.31
	Roberts	46.91	91.74	44.05
MD	Canny	20.99	46.59	13.28
	LoG	18.90	57.70	11.65
	Prewitt	23.73	80.98	15.56
	Sobel	23.80	81.01	15.62
	Roberts	35.24	80.17	27.21
HD	Canny	22.02	68.94	14.12
	LoG	19.05	80.83	11.76
	Prewitt	17.98	63.16	10.96
	Sobel	18.16	63.75	11.09
	Roberts	16.88	56.78	10.15

3 Results

The previous section showed how these results were obtained. In Section 3, the results are divided by edge detector, i.e. a sub-section is made for each detector to make the results more transparent. The results were obtained using the mathematical models defined in Section 2. The calculation is based on the theoretical models presented in [44-47]. The results are presented in tables and for each combination of parameters (compression and edge detector) can be found in the repository [53], as well as, full size images used code.

3.1 Canny Edge Detector

Table 4, Table 5 and Table 6 show the F, FOM and PR values, respectively, obtained by applying a Canny edge detector over images with different number of details compressed by different compression algorithms. Based on the obtained results, it can be seen that by increasing the number of bits per pixel, better values are obtained. The best values are obtained when the number of details in the image is small, while when the number of details in the image is medium and high obtained values are similar.

Table 4

F values obtained by using a Canny edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	25.55	32.91	35.47	35.52	35.49	35.49
	JPEG2000	28.19	34.42	35.32	35.55	35.55	35.55
	SPIHT	28.66	34.05	35.15	35.52	35.63	35.56
MD	**JPEG**	17.97	19.49	20.90	20.51	20.88	21.04
	JPEG2000	20.37	20.94	20.74	20.89	21.13	21.22
	SPIHT	20.63	20.94	21.18	21.19	21.36	21.96
HD	**JPEG**	20.45	20.44	21.94	21.97	21.90	22.00
	JPEG2000	19.24	21.26	21.29	22.08	22.09	22.98
	SPIHT	19.73	21.12	21.29	21.83	22.03	22.11

Table 5

FOM values obtained by using a Canny edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	75.03	89.19	89.39	89.39	89.36	89.36
	JPEG2000	86.75	89.30	89.39	89.36	89.36	89.36
	SPIHT	88.37	89.11	89.28	89.39	89.39	89.40
MD	**JPEG**	47.01	42.50	47.24	44.35	46.31	47.89
	JPEG2000	52.63	47.34	46.30	46.64	46.96	47.09
	SPIHT	52.21	47.51	49.91	49.97	51.11	51.79

HD	**JPEG**	58.16	65.58	67.46	67.08	66.95	69.95
	JPEG2000	64.25	73.27	73.83	78.63	78.55	78.32
	SPIHT	55.65	73.91	75.09	75.13	75.38	75.65

Table 6

PR values obtained by using a Canny edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	17.16	24.53	27.49	27.54	27.51	27.51
	JPEG2000	19.63	26.24	27.31	27.58	27.58	27.58
	SPIHT	20.08	25.82	27.10	27.55	27.68	27.59
MD	**JPEG**	10.95	12.10	13.21	12.90	13.20	13.32
	JPEG2000	12.79	13.24	13.08	13.21	13.39	13.71
	SPIHT	12.99	13.24	13.44	13.47	13.52	13.76
HD	**JPEG**	12.85	12.85	14.06	14.08	14.02	14.11
	JPEG2000	11.91	13.50	13.53	14.17	14.05	14.28
	SPIHT	12.29	13.38	13.52	13.96	14.13	14.21

3.2 LoG Edge Detector

Table 7, Table 8 and Table 9 show the F, FOM and PR values, respectively, obtained by applying LoG edge detectors over images with different number of details compressed by different compression algorithms. The LoG detector gave the best results when the number of details in the image is small. However, it can be seen that the values are lower when the number of details in the image is medium and high at BPP 0.1, but the values did not increase much by further increasing the BPP from 0.3 upwards.

Table 7

F values obtained by using a LoG edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	21.46	31.98	33.07	32.77	32.72	32.72
	JPEG2000	28.82	31.67	32.24	32.73	32.73	32.73
	SPIHT	29.10	31.89	32.50	32.77	32.77	32.84
MD	**JPEG**	16.22	17.99	18.45	18.55	18.83	18.89
	JPEG2000	18.61	18.38	18.54	18.61	18.84	18.98
	SPIHT	18.49	18.45	18.51	18.55	18.98	18.85
HD	**JPEG**	17.78	18.78	18.99	19.06	19.08	19.21
	JPEG2000	15.57	18.56	18.63	19.18	19.16	19.22
	SPIHT	16.48	18.26	18.70	18.86	18.96	19.04

Table 8

FOM values obtained by using a LoG edge detector

	BPP	0.1	0.3	0.5	1	1.5	3
SD	**JPEG**	74.29	89.92	90.11	90.07	90.72	90.06
	JPEG2000	89.49	90.02	90.09	90.09	90.09	90.09
	SPIHT	89.25	90.01	90.02	90.09	90.08	90.11
MD	**JPEG**	53.66	55.71	56.54	56.20	57.44	57.72
	JPEG2000	57.94	59.15	59.40	57.18	57.38	59.55
	SPIHT	69.19	59.70	59.13	57.55	59.71	60.69
HD	**JPEG**	78.42	79.42	80.88	81.50	81.54	81.86
	JPEG2000	64.75	77.07	79.16	80.79	80.80	80.88
	SPIHT	64.04	76.20	79.13	80.10	80.64	80.82

Table 9

PR values obtained by using a LoG edge detector

	BPP	0.1	0.3	0.5	1	1.5	3
SD	**JPEG**	13.66	25.51	24.71	24.37	24.31	24.31
	JPEG2000	20.25	23.18	23.79	24.33	24.33	24.33
	SPIHT	20.52	23.42	24.08	24.37	24.39	24.42
MD	**JPEG**	9.68	10.97	11.31	11.39	11.60	11.65
	JPEG2000	11.43	11.26	11.38	11.42	11.61	11.76
	SPIHT	11.34	11.31	11.36	11.36	11.71	11.91
HD	**JPEG**	10.81	11.56	11.72	11.77	11.78	11.92
	JPEG2000	9.22	11.40	11.45	11.87	11.77	11.94
	SPIHT	9.86	11.17	11.50	11.62	11.70	11.76

3.3 Sobel Edge Detector

Table 10, Table 11 and Table 12 give the F, FOM and PR values, respectively, obtained by applying a Sobel edge detector over images with different number of details compressed by different compression algorithms. Sobel detector gave good results when it comes to the small and medium number of details in the image. With high number of details, the results are poor at lower BPP.

Table 10

F values obtained by using a Sobel edge detector

	BPP	0.1	0.3	0.5	1	1.5	3
SD	**JPEG**	20.18	33.95	34.40	34.29	34.34	34.34
	JPEG2000	31.49	33.84	34.18	34.33	34.33	34.33
	SPIHT	30.99	33.83	34.42	34.34	34.41	34.56
MD	**JPEG**	17.81	22.61	23.31	23.75	23.89	23.77

	JPEG2000	20.33	22.84	23.32	23.47	23.49	23.76
	SPIHT	20.25	22.80	23.36	23.65	23.62	23.89
HD	**JPEG**	11.62	15.85	17.15	17.62	17.79	18.07
	JPEG2000	12.88	17.28	17.65	18.00	17.98	18.18
	SPIHT	13.44	17.44	17.82	18.18	18.11	18.14

Table 11

FOM values obtained by using a Sobel edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	68.69	88.86	89.32	89.47	89.50	89.50
	JPEG2000	85.54	88.94	89.36	89.38	89.37	89.38
	SPIHT	84.90	88.98	89.33	89.40	89.44	89.56
MD	**JPEG**	68.37	77.33	79.56	81.89	80.88	81.01
	JPEG2000	67.29	77.80	79.80	80.68	80.89	80.98
	SPIHT	66.28	77.92	79.28	80.57	80.80	81.08
HD	**JPEG**	48.64	57.85	60.82	62.73	64.66	63.77
	JPEG2000	46.09	59.73	61.93	63.32	63.61	63.97
	SPIHT	46.98	60.32	61.46	63.03	63.37	63.77

Table 12

PR values obtained by using a Sobel edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	12.64	25.70	26.22	26.10	26.15	26.15
	JPEG2000	22.98	25.58	25.96	26.13	26.13	26.13
	SPIHT	22.45	25.56	26.25	26.15	26.27	26.29
MD	**JPEG**	10.83	14.61	15.20	15.58	15.70	15.59
	JPEG2000	12.76	14.80	15.20	15.33	15.35	15.58
	SPIHT	12.69	14.77	15.24	15.49	15.57	15.69
HD	**JPEG**	6.58	9.42	10.35	10.70	10.82	11.03
	JPEG2000	7.39	10.45	10.72	10.98	10.96	11.11
	SPIHT	7.77	10.56	10.84	11.11	11.06	11.09

3.4 Prewitt Edge Detector

Table 13, Table 14 and Table 15 show the F, FOM and PR values, respectively, obtained by applying a Prewitt edge detector over images with different number of details compressed by different compression algorithms. The Prewitt operator obtained well values with JPEG2000 and SPIHT compression in SD images even when the number of bits per pixel is low.

Table 13

F values obtained by using a Prewitt edge detector

	BPP	0.1	0.3	0.5	1	1.5	3
SD	**JPEG**	18.52	33.93	34.54	34.49	34.62	34.62
	JPEG2000	31.71	34.10	31.40	34.50	34.50	34.52
	SPIHT	31.01	34.02	34.53	34.57	34.67	34.75
MD	**JPEG**	17.87	22.73	23.39	23.87	24.06	23.79
	JPEG2000	20.40	22.91	23.34	23.66	23.65	24.02
	SPIHT	20.37	22.91	23.34	23.64	23.67	23.74
HD	**JPEG**	11.72	15.87	17.15	17.78	17.98	18.05
	JPEG2000	12.92	17.39	17.97	18.05	18.04	18.15
	SPIHT	13.47	17.62	17.98	18.06	18.15	18.33

Table 14

FOM values obtained by using a Prewitt edge detector

	BPP	0.1	0.3	0.5	1	1.5	3
SD	**JPEG**	68.12	88.72	89.30	89.42	89.46	89.47
	JPEG2000	85.62	89.09	89.29	89.36	89.36	89.36
	SPIHT	85.02	88.82	89.22	89.31	89.40	89.44
MD	**JPEG**	68.08	77.44	79361	81.68	80.80	81.03
	JPEG2000	67.51	77.78	79.68	80.69	81.05	81.79
	SPIHT	66.56	77.94	79.18	80.66	80.72	80.86
HD	**JPEG**	48.76	57.13	60.49	62.61	64.00	64.13
	JPEG2000	46.25	59.53	61.77	63.14	62.97	63.23
	SPIHT	47.20	60.28	61.51	62.89	62.71	63.25

Table 15

PR values obtained by using a Prewitt edge detector

	BPP	0.1	0.3	0.5	1	1.5	3
SD	**JPEG**	11.36	25.68	26.38	26.32	26.48	26.50
	JPEG2000	23.21	25.87	26.22	26.33	26.33	26.33
	SPIHT	22.47	25.78	26.31	26.42	26.55	26.59
MD	**JPEG**	10.88	14.71	15.27	15.68	15.84	15.60
	JPEG2000	12.82	14.86	15.22	15.50	15.49	15.88
	SPIHT	12.79	14.86	15.30	15.48	15.51	15.57
HD	**JPEG**	6.63	9.43	10.35	10.81	10.96	11.04
	JPEG2000	7.42	10.53	10.95	11.01	11.00	11.19
	SPIHT	7.79	10.69	10.96	11.02	11.09	11.00

3.5 Roberts Edge Detector

Table 16, Table 17 and Table 18 show the F, FOM and PR values, respectively, obtained by applying Roberts edge detectors over images with different number of details compressed by different compression algorithms. When it comes to the small number of details in an image, also and mostly when the number of details in an image is medium, the Roberts operator obtained the best results using JPEG2000 and SPIHT compression.

Table 16

F values obtained by using a Roberts edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	20.64	43.98	46.85	46.64	46.69	46.69
	JPEG2000	38.21	45.22	46.86	46.55	46.55	46.61
	SPIHT	37.63	45.90	46.70	46.60	46.73	46.69
MD	**JPEG**	18.42	31.08	32.71	33.95	35.03	35.24
	JPEG2000	24.46	31.33	25.83	34.94	35.25	35.78
	SPIHT	24.51	31.36	34.32	35.01	35.25	35.42
HD	**JPEG**	12.78	14.06	14.81	15.11	14.55	17.05
	JPEG2000	11.03	14.16	15.77	17.48	17.74	17.98
	SPIHT	13.47	17.62	17.98	18.06	18.15	18.33

Table 17

FOM values obtained by using a Roberts edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	69.50	90.55	91.62	91.69	91.73	91.73
	JPEG2000	82.01	90.38	91.59	91.66	91.66	91.66
	SPIHT	80.93	90.79	91.50	91.72	91.72	91.66
MD	**JPEG**	66.86	75.36	76.53	80.48	80.44	81.11
	JPEG2000	58.17	72.75	79.08	80.21	80.14	81.01
	SPIHT	56.63	71.30	78.51	79.89	80.17	80.24
HD	**JPEG**	47.90	52.59	52.61	53.27	54.00	56.67
	JPEG2000	32.82	45.35	52.80	57.17	58.30	59.94
	SPIHT	47.20	60.28	61.51	62.89	62.71	63.25

Table 18

PR values obtained by using a Roberts edge detector

	BPP	**0.1**	**0.3**	**0.5**	**1**	**1.5**	**3**
SD	**JPEG**	13.01	39.55	44.07	43.70	43.79	43.79
	JPEG2000	30.92	41.27	44.09	43.54	43.54	43.54
	SPIHT	30.16	42.42	43.81	43.64	43.76	43.61
MD	**JPEG**	11.29	22.55	24.31	25.70	26.96	27.21

	JPEG2000	16.19	22.81	25.83	26.85	27.22	27.46
	SPIHT	16.23	22.84	26.13	26.94	27.22	27.29
HD	**JPEG**	7.33	8.18	8.69	8.90	8.51	10.27
	JPEG2000	6.20	8.25	9.36	10.59	10.79	10.88
	SPIHT	7.79	10.69	10.96	11.02	11.09	11.11

In Section 2 it could be seen visually and objectively how compression and different BPP values effect on quality of images. While in Section 3, edge detection was performed on these images, and based on the results presented in tables by all operators, the effect of compression on edge detection can be seen. BPP effect on edge detection as well as the number of details in an image. The Canny operator has proven to be a good solution but when the number of details in the image is small or medium, the Roberts operator finds its application. For this reason, Figure 13 shows the edge detection using the Roberts operator. Detection is shown for an image that is compressed with JPEG technique at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3.

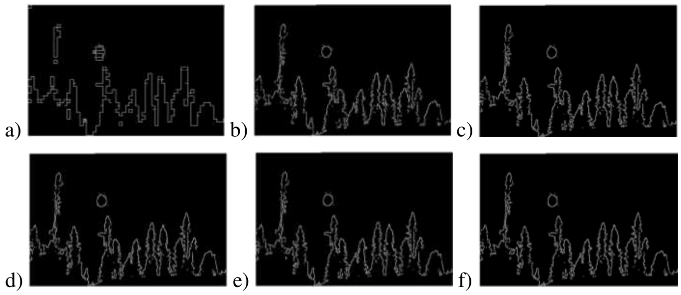


Figure 13
Roberts detection for SD image with JPEG compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

Since the Canny operator proved to be a very good solution when the number of details in the image is high even at lower BPP values, Figure 14 shows the edge detection using this operator. Detection is shown for an image that is compressed with JPEG technique at BPP a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3.

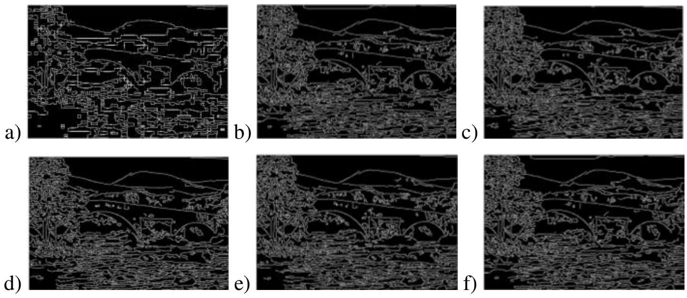


Figure 14
Canny detection for HD image with JPEG compression at BPP: a) 0.1, b) 0.3, c) 0.5, d) 1, e) 1.5, f) 3

Thus, Figure 13 and Figure 14 are only a visual representation of the results obtained in Section 3. All images for the results obtained can be found in the repository [53].

Conclusions

This paper analyzes gradient (Sobel, Prewitt, Robert), Laplacian of Gaussian and Canny operator. Operators were applied to images which consist different number of details (small, medium and high) and compressed by JPEG, JPEG2000 and SPIHT compression algorithms at different bits per pixel. Objective measures were used - F measure, FOM and PR and the results are tabulated.

Based on the obtained results, it can be concluded that when the number of details in the image is small and medium and when using JPEG compression, the best results are obtained using the Roberts operator, only at a BPP of 0.1 Canny achieved better values. Other operators have similar values. With the same amount of detail in image, JPEG2000 and SPIHT compression achieve better results than JPEG, which is reflected in edge detection. Also, the Roberts operator obtained the best results over the other operators, however, the values are similar to JPEG compression except when the BPP is low. Using high detail images and JPEG, JPEG2000 and SPIHT compression, the best edge detection was obtained with the Canny operator. Edge detection is better when JPEG2000 and SPIHT are used.

The results obtained in this paper contribute to the further development of image compression algorithms to be more suitable for use in systems where image processing such as segmentation and edge detection is used. However, it provides an incentive to optimize edge detectors for image compression at lower bits per pixel values, with consideration of the complexity of the image.

References

[1] A. J. Hussain, A. Al-Fayadh and N. Radi: Image compression techniques, A survey in lossless and lossy algorithms, Neurocomputing, 2018, Vol. 300, pp. 44-69

[2] D. Ravichandran, R. Nimmatoori and M. G. Ahamad: A Study on State-of-the-Art Image Compression Algorithms and Standards, International Journal On Advanced Computer Theory And Engineering (IJACTE), 2016, Vol. 5, Iss. 3, pp. 2319-2523

[3] K. V. Gomathi and R. Lotus: Digital image compression techniques, IJRET: International Journal of Research in Engineering and Technology, 2014, Vol. 03 Iss. 10, pp. 285-290

[4] G. K. Wallace: The JPEG Still Picture Compression Standard, Comm. of the ACM, 1991, Vol. 34, No. 4, pp. 30-44

[5] C. Christopoulos, A. Skodras and T. Ebrahimi: The JPEG2000 Still Image Coding System: An Overview, IEEE Trans. on Consumer Electronics, November 2000, Vol. 46, No. 4, pp. 1103-1127

[6] R. C. Gonzalez, R. E. Woods and S. L. Eddins: Digital Image Processing Using MATLAB. Pearson Prentice Hall, 2004

[7] D. Santa-Cruz, T. Ebrahimi, J. Askelof, M. Larsson and C. Christopoulos: JPEG 2000 Still Image Coding Versus other Standards, ISO/IEC JTC1/SC29/WG1 (ITU-T SG8), July 2000, http://www.jpeg.org/public/wg1n1816.pdf

[8] J. Kumar and M. Kumar: Comparison of image compression methods on various images, International Conference on Advances in Computer Engineering and Applications, Ghaziabad, 2015, pp. 114-118

[9] R. Singh, and V. K. Srivastava: JPEG2000: A review and its performance comparison with JPEG. Seconnd International Conference on Power, Control and Embedded Systems, Allahabad, India, 2012

[10] P. L. Dragotti, G. Poggi and A. Ragozini: Compression of multispectral images by three-dimensional SPIHT algorithm, IEEE Trans. on Geosciences and Remote Sensing, January 2000, Vol. 38, No. 1, pp. 416-428

[11] B. I. Kochi, and B. B. S. Kumar: EZW and SPIHT Algorithms for Image Compression and Denoising, ITSI Transactions on Electrical and Electronics Engineering (ITSI-TEEE), 2016, Vol. 4, No. 2, pp. 2320-8945

[12] C. Kaur and S. Budhiraja: Improvements of SPIHT in Image Compression-Survey, International Journal of Emerging Technology and Advanced Engineering, January 2013, Vol. 3, Iss. 1, pp. 652-656

[13] Sharon M. Perlmutter, Pamela C. Cosman, Chien-Wen Tseng, Richard A. Olshen, Robert M. Gray, King C. P. Li and Colleen J. Bergin: Medical Image Compression and Vector Quantization, Statistical Science, 1998, Vol. 13, No. 1, pp. 30-53

[14] D. A. Koff and H. Shulman: An Overview of Digital Compression of medical Images: Can We Use Lossy Image Compression in Radiology, Radiol J., Oct. 2006, Vol. 57, No. 4, pp. 211-217

[15] M. R. Ashwin Dhivakar, M. G. Ahamad and D. Ravichandran: Medical image compression using Embedded Zerotree Wavelet (EZW) coder, International Conference System Modeling & Advancement in Research Trends (SMART), Moradabad, Indina, 2016, pp. 17-23

[16] S. Suma and V. Sridhar: A Review of the Effective Techniques of Compression in Medical Image Processing, International Journal of Computer Applications, 2016, Vol. 97, pp. 23-30

[17] T. H. Oh and R. Besar: Medical image compression using jpeg-2000 and jpeg: a comparison study, Journal of Mechanics in Medicine and Biology, 2002, Vol. 2, No. 3 & 4, pp. 313-328

[18] Y. Li: Synthetic Aperture Radar (SAR) Image Compression Using the Wavelet Transform, Faculty of Engineering and Applied Sciences Mernorial University of Newfoundland August, 1997

[19] A. T. Chien, K. S. Miettinen, A. Lan amd M. A. Lepley: Applications of Digital Image Processing XXIII, 2000, doi:10.1117/12.411604

[20] U. Pestel-Schiller: Coding of SAR image data for data compression, in 10th European Conference on Synthetic Aperture Radar (EUSAR), Berlin, Germany, Jun. 2014

[21] K. Sahnoun and B. Noureddine: Satellite Image Compression Algorithm Based on the FFT, The International journal of Multimedia & Its Applications, 2016, Vol. 6, No. 1, pp. 77-83

[22] J. G. Kumar and A. Singh: Fractal Compression of Satellite Images, J. Indian Soc. Remote Sens. (December 2008) Vol. 36, pp. 299-311

[23] Z. Liang, T. Xinming, Z. Guo and W. Xiaoliang: Effects of jpeg2000 and spiht compression on image classification, In The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences Conference, Beijing, China, August 2008

[24] B. Emmanuel, M. Mu'Azu, S. Sani and S. Garba: A Review of Wavelet-Based Image Processing Methods for Fingerprint Compression in Biometric Application, British Journal of Mathematics & Computer Science, 2014, Vol. 4, pp. 2781-2798

[25] A. J. Hussain, D. Al-Jumeily, N. Radi and P. Lisboa: Hybrid neural network predictive-wavelet image compression system, Neurocomputing, 2014, Vol. 151, pp. 975-984

[26] R. Jumar, H. Maaß, V. Hagenmeyer: Comparison of lossless compression schemes for high rate electrical grid time series for smart grid monitoring and analysis, Computers & Electrical Engineering, Vol. 71, October 2018, pp. 465-476

[27] M. Trojanová, and A. Hošovský: Comparison of Different Neural Networks Models for Identification of Manipulator Arm Driven by Fluidic Muscles, Acta Polytechnica Hungarica, 2018, Vol. 15, No. 7, pp. 7-28

[28] R. Zall, 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

[29] J. C. Spall: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, IEEE Transactions on Automatic Control, 1992, Vol. 37, No. 3, pp. 332-341

[30] C. Pozna, R. E. Precup, J. Tar, I. Škrjanc, S. Preitl: New results in modelling derived from Bayesian filtering, Knowledge-Based Systems, 2010, Vol. 23, No. 2, pp. 182-194

[31] S. I. Jabbar, C. R. Day, N. Heinz and E. K. Chadwick: Using Convolutional Neural Network for edge detection in musculoskeletal ultrasound images, 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, Canada, 2016, pp. 4619-4626

[32] M. Hagara and P. Kubinec, About Edge Detection in Digital Images, Radioengineering, 2018, Vol. 27, No. 4, pp. 919-929

[33] E. Moya-Albor, H. Ponce and Jorge Brieva: An Edge Detection Method using a Fuzzy Ensem- ble Approach, Acta Polytechnica Hungarica, 2017, Vol. 14, No. 3, pp. 149-168

[34] K. B. Krishnan, S. P. Ranga and N. Guptha: Survey on Different Edge Detection Techniques for Image Segmentation, Indian Journal of Science and Technology, 2017, Vol. 10, No. 4

[35] N. A. Othman, M. U. Salur, M. Karakose and I. Aydin: An Embedded Real-Time Object Detection and Measurement of its Size, International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-4

[36] A. Rupani, P. Whig, G. Sujediya and P. Vyas: A robust technique for image processing based on interfacing of Raspberry-Pi and FPGA using IoT, International Conference on Computer, Communications and Electronics (Comptelix), Jaipur, India, 2017, pp. 350-353

[37] T. Mustafa: Object detection and tracking for real time field survelliance applications, January 2017, Ankara, Turkey

[38] M. S. Munna, B. K. Tarafder, M. G. Robbani and T. C. Mallick: Design and implementation of a drawbot using Matlab and Arduino Mega, International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 2017, pp. 769-773

[39] P. Arunkumar, S. Shantharajah, and M. Geetha: Improved canny detection algorithm for processing and segmenting text from the images, Cluster Computing, 2018, Vol. 22, pp. 7015-7021

[40] S. Biswas, D. Ghoshal and R. Hazra: A new algorithm of Image Segmentation using Curve Fitting Based Higher Order Polynomial Smoothing, Optik, 2016, Vol. 127, No. 20, pp. 8916-8925

[41] S. Biswas and R. Hazra: Robust edge detection based on Modified Moore-Neighbor, Optik, 2018, Vol. 168, pp. 931-943

[42] S. Sergyán: A New Approach of Face Detection-based Classification of Image Databases, Acta Polytechnica Hungarica, 2009, Vol. 6, No. 1, pp. 175-184

[43] B. Kuljić, J. Simon and T. Szakáll: Pathfinding Based on Edge Detection and Infrared Distance Measuring Sensor, Acta Polytechnica Hungarica, 2009, Vol. 6, No. 1, pp. 103-116

[44] V. Maksimovic, P. Lekic, M. Petrovic, B. Jaksic and P. Spalevic: Experimental analysis of wavelet decomposition on edge detection, Proceedings of the Estonian Academy of Sciences, 2019, Vol. 68, No. 3, pp. 284-298

[45] K. Wat and S.H. Srinivasan: Effect of Compression on Face Recognition, Proceedings of the fifth International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2004, 21-23 April 2004, Lisboa, Portugal

[46] H. Moon and P. J. Phillips: Computational and Performance Aspects of PCA-based Face recognition Algorithms, Perception, 2001, Vol. 30, pp. 303-321

[47] Delac K., Grgic M. and S. Grgic: Effects of JPEG and JPEG2000 Compression on Face Recognition, Pattern Recognition and Image Analysis, 2005, Vol. 3687, pp. 136-145

[48] The Berkeley Segmentation Dataset and Benchmark, https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/, Acessed: 14.07.2019

[49] S. Ilic, M. Petrovic, B. Jaksic, P. Spalevic, Lj. Lazic and M. Milosevic: Experimental analysis of picture quality after compression by different methods. Przegląd Elektrotechniczny, 2013, Vol. 89, No. 11, pp. 190-194

[50] H. Yu and S. Winkler: Image complexity and spatial information. Proceedings of fifth International Workshop on Quality of Multimedia Experience (QoMEX), Klagenfurt am Worthersee, Austria, 2013, pp. 12-17

[51] H. Wu, C. Mark, and K. Robert: A study of video motion and scene complexity. Tech. Rep. WPI-CS-TR-06-19, 2006, Worcester Polytechnic Institute

[52] V. Chikhman, V. Bondarko, M. Danilova, A. Goluzina and Y. Shelepin: Complexity of images: Experimental and computational estimates compared. Perception, 2012, Vol. 41, No. 6, pp. 631-47

[53] Dataset repository: https://drive.google.com/open?id= 1s6MkarMqT_au0JOsTUODCNOxgWfIYZut, Accesed: 28.01.2020

[54] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik: Contour Detection and Hierarchical Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, Vol. 33, No. 5, pp. 898-916

[55] W. K. Pratt: Digital Image Processing. 4th ed., Hoboken, New Jersey (USA): John Wiley & Sons, 2007, ISBN: 9780471767770

[56] P. A. Khaire, N. V. Thakur: A Fuzzy Set Approach for Edge Detection. International Journal of Image Processing, 2012, Vol. 6, No. 6, pp. 403-412

Sensorless Vector Control of Permanent Magnet Synchronous Machine Using High-Frequency Signal Injection

Gergely Szabó, Károly Veszprémi

Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics, Department of Electric Power Engineering,
Egry József u. 18, 1111 Budapest, Hungary
szabo.gergely@vet.bme.hu, veszpremi.karoly@vet.bme.hu

Abstract: The vector control theory of alternating current machines could provide high performance control during transient events, since these methods do not depend on the static equations of the selected machine, but on the space vector-based differential system of equations. These control methods have a very important common property; all of them require an angle, with which the system's input can be transformed into the common reference frame in which the space vector notation is construed. The sensorless control methods attempt to estimate the common coordinate system's angle, without using any information from the encoder, one of which is the high-frequency voltage injection method. This paper presents the mathematical model of the high-frequency synchronous voltage injection method on permanent magnet synchronous machines. The common coordinate system is estimated using a Phase-Locked-Loop (PLL). Based on the detailed mathematical model a new equivalent dynamic model for the PLL structure is proposed, with which the PLL's controller could be tuned, with the knowledge of the machine's parameters and injected voltage properties. Simulation results are provided for an off-the-shelf interior magnet synchronous machine.

Keywords: Vector Control; Permanent Magnet Synchronous Machine; Sensorless; High-Frequency Signal Injection

1 Introduction

The sensorless vector control methods try to eliminate the speed encoder from the controlled electric drives, but in most cases, the shaft angle encoder cannot be omitted, because of the safety level of the application. In such drives, a sensorless vector control method can be used as a backup algorithm, with which the drive system can be stopped without damaging the drive or the users. The sensorless term must be clarified depending on the machine to be controlled; in case of an induction machine, this term needs to be divided into at least two approaches because the shaft angular position and the common coordinate system's angle, which is used to be rotor flux vector's position, are not the same and depends on the state of the machine. The simple sensorless term can be used for an algorithm, with which the common coordinate system can be estimated without using angle feedback from the shaft, but in this case, the closed-loop speed control still depends on the encoder attached on the rotor. On the other hand, a speed-sensorless method is capable to control the induction machine's angular speed too, without any encoder built in the system. In the case of a synchronous machine, the sensorless and speed-sensorless terms can be merged since the pole-flux vector's angle is in direct relation with the shaft's mechanical angle.

From the algorithm point of view, the available methods could be categorized into two sections. The first one attempts to estimate the machine's signals based on its mathematical models and the combination of several filtering methods and controllers. Model Reference Adaptive System (MRAS) approaches are one of the well-known methods of this category. Authors in [1] detail a vector-controlled solution for permanent magnet synchronous machines, which is based on MRAS method. The control structure was constructed using the classical cascade PI speed and current control loops, the common coordinate system, which is required for the pole flux vector oriented vector control, was estimated using the MRAS method. Since angle estimation incorporates the entire machine equations, the precise knowledge of the parameters is required for a stable control over a wide range of synchronous frequency and temperature. Authors in [2] give a detailed case study of the parameter deviations' effect in an MRAS-based system, whilst [3] introduces an adaptive approach to overcome the uncertainty of the required parameters during operation. The computational effort of this method must be handled carefully, for which [4] proposes a reduced-order observer and neural network solution in terms of rotor flux estimation and compensation.

This category also incorporates the well-known estimator structures; such as Kalman filtering method, which also uses the machine's model, combined with model of the system's and the measurements' disturbances. Authors in [5] propose an Extended Kalman Filter-based (EKF) method, which does not require the knowledge of the mechanical parameters nor the initial rotor position. Since these algorithms are sample-based their implementation on pulse width modulated voltage-source inverter's microcontroller is straightforward. A major issue in these approaches is the selection of the covariance matrix, which is the result of a trial-and-error tuning in most of the cases. To overcome the difficulties of choosing the covariance matrix, [6] details an algorithm to select the appropriate parameters.

Sensorless algorithms in the second category try to exploit the machine's magnetic properties, which are dependent on the rotor's position or the magnitude of the flux.

A well-known method of this category is the INFORM, Indirect Flux detection by On-line Reactance Measurement [7], which tries to track the machine's impedance with which the actual rotor position can be estimated. This algorithm does not involve any test signals, it just calculates the machine's impedances based on the measured currents and voltages. Similar to the previous MRAS solutions, the addition of an Extended Kalman Filter to this system can provide a better estimation.

The high-frequency signal injection methods are also part of the second group, in which methods usually voltage signals are injected into the predefined points of the control structure. The latter one leads us two main approaches; the high-frequency stationary injection modifies the phase voltages of the machine by adding a symmetric voltage system to the motor's terminal voltages [8-9]. In the second approach, the test vectors are injected in the estimated common reference frame, hence these techniques are called synchronous injection methods. Authors in [10] give a comparative analysis of the two aforementioned solutions. Besides the point of the injections, the signal processing methods with which the common coordinate system can be estimated are different.

The stationary injections usually involve heterodyne filtering techniques, in which the measured high frequency stationary currents are transformed into several coordinates systems in which they are filtered to obtain the required angle information [11]. In the case of synchronous injection, the flux vector's position can be estimated with Phase-Locked-Loop (PLL), but [12] proposes a discrete algorithm-based solution, which enables higher bandwidth in the estimation and speed control.

Section 2 presents the detailed mathematical model of high-frequency synchronous injection method on a permanent magnet synchronous machine. This is followed by the description of the common coordinate system's estimation, which is performed using a PLL. Based on the mathematical description a new dynamic model is proposed for the PLL-based estimator structure, which is required for most of the controller structures and their proper tuning.

The control structure assumed to be a widely used cascade control loop, containing d-direction current control loop, a q-direction current control loop which setpoint is provided by the outer speed controller loop. With the help of the new dynamic model the PLL's and the cascade control loop's PI controllers be tuned with one of the published methods to achieve a stable control. In Section 3 simulation results are presented using an off-the-shelf interior permanent magnet synchronous motor's parameters.

2 High-Frequency Syncronous Voltage Injection

2.1 Mathematical Model of Permanent Magnet Synchronous Machine

The sinusoidal field, generated by an ideal permanent magnet can be modeled with a constant amplitude pole flux vector, which is bound to the point, where the rotor magnets' spatial flux density distribution reaches its maximum value. The common coordinate system is fixed to this vector because its cross product with the stator current defines the magnitude of the motor's torque. Equations (1)-(4) show the system of equations, which models the machine in the $d - q$ frame, whilst Fig. 1 shows the assumptions of the dynamic model, including the space vector-based equivalent circuit. In Fig. 1(b) $\bar{\bar{L}}$ denotes the 2×2 diagonal stator inductance matrix, which contains the d- and q-direction inductances.

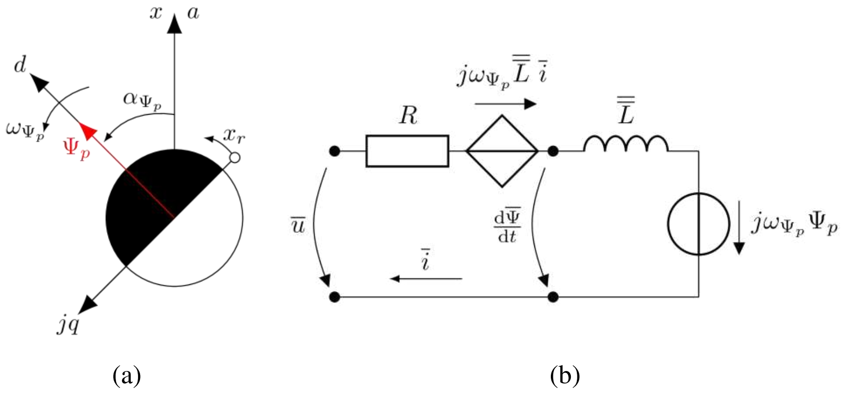


Figure 1

(a) Permanent magnet rotor and the definition of the common coordinate system's angle,

(b) Equivalent circuit of permanent magnet synchronous machine

$$u_d = Ri_d + L_d\frac{di_d}{dt} - \omega_{\Psi_p}L_q i_q \ , \tag{1}$$

$$u_q = Ri_q + L_q\frac{di_q}{dt} + \omega_{\Psi_p}L_d i_d + \omega_{\Psi_p}\Psi_p \ , \tag{2}$$

$$m = \frac{3}{2}p\left(\left(L_d - L_q\right)i_d i_q + \Psi_p i_q\right), \tag{3}$$

$$\Theta\frac{d\omega}{dt} = m - m_l - F\omega \ , \tag{4}$$

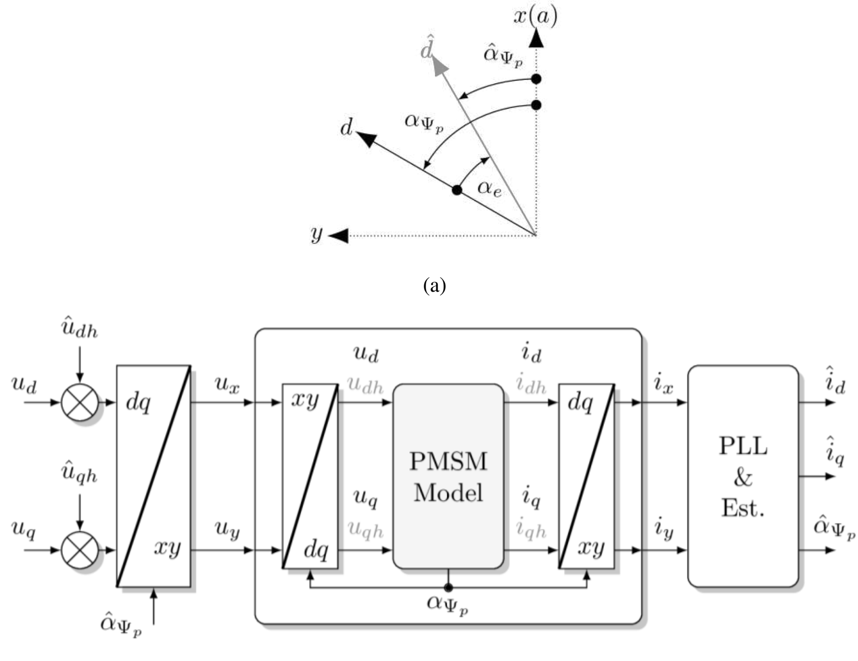
where u_d is the d component of the stator voltage, R is the stator resistance, i_d is the d component of stator current, L_d is the inductance in the d-direction, L_q is the inductance in the q-direction, i_q is the q component of stator current, u_q is the q component of the stator voltage, Ψ_p is the pole-flux vector's amplitude, m is the machine's electromagnetic torque, m_l is the load torque on the shaft, Θ is the rotor's moment of inertia, F is the friction loss factor, p is the number of pole pairs, ω is the shaft angular speed and ω_{Ψ_p} is the pole-flux vector's angular speed where $\omega_{\Psi_p} = p\omega$.

2.2 High-Frequency Synchronous Injection

In case of high-frequency synchronous injection, the test vectors are injected in the estimated $\hat{d} - \hat{q}$ frame shown in Fig. 2(a). This figure also explains the angle relations, which were used during the modeling process; based on Eq. (5), the angle displacement between the real and estimated coordinate systems is considered to be positive, if the estimated angle lags behind the real one.

$$\alpha_e = \alpha_{\Psi_p} - \hat{\alpha}_{\Psi_p} \, , \tag{5}$$

where α_{Ψ_p} is the pole flux vector's angle, $\hat{\alpha}_{\Psi_p}$ is its estimated value. In the following, the hat symbol (^) will denote the estimated values, h subscripts will refer to high-frequency components. Fig. 2(b) illustrates the block diagram of the injection method, where the signals, which will be detailed, are indicated.



(a)

(b)

Figure 2

(a) Angle relation between the real $d - q$ and estimated $\hat{d} - \hat{q}$ coordinate systems (b) Block diagram of synchronous injection

The injected voltages are not DC like quantities as they are used to be in the $d - q$ frame, but high-frequency sinusoidal signals. During the mathematical modeling, these time signals were handled using complex phasors, where the complex rotating vector was bound to the sine wave. According to this, the injected voltages are the following in both time and complex frequency domain,

$$\begin{bmatrix} \hat{u}_{dh} \\ \hat{u}_{qh} \end{bmatrix} = \begin{bmatrix} u_h \sin(\omega_h t) \\ -u_h \cos(\omega_h t) \end{bmatrix} = \begin{bmatrix} u_h \\ -j u_h \end{bmatrix} , \tag{6}$$

where $\hat{u}_{dh}$ and $\hat{u}_{qh}$ are the injected voltages in the estimated $\hat{d}$-and $\hat{q}$-directions, u_h is the amplitude of the injected voltages, $\omega_h = 2\pi f_h$, where f_h is the injection frequency, j is the imaginary unit.

With these definitions of the angle displacement and the injected voltages, the projections of the high-frequency signals on the real $d - q$ axes can be calculated using the rotation operator, as

$$\begin{bmatrix} u_{dh} \\ u_{qh} \end{bmatrix} = \bar{\bar{R}}(\alpha_e) \begin{bmatrix} \hat{u}_{dh} \\ \hat{u}_{qh} \end{bmatrix} , \tag{7}$$

where $\bar{\bar{R}}(\alpha_e)$ is the rotational operator and

$$\bar{\bar{R}}(\alpha_e) = \begin{bmatrix} \cos(\alpha_e) & \sin(\alpha_e) \\ -sin(\alpha_e) & \cos(\alpha_e) \end{bmatrix} , \tag{8}$$

and its inverse can be calculated as follows,

$$\bar{\bar{R}}^{-1}(\alpha_e) = \begin{bmatrix} \cos(\alpha_e) & -\sin(\alpha_e) \\ sin(\alpha_e) & \cos(\alpha_e) \end{bmatrix} . \tag{9}$$

Equations (8)-(9) are not only valid for α_e and Eq. (7), but any given α angle. This method uses test voltages; therefore, the current response of the system could be calculated using Ohm’s law as shown in Eq. (10).

$$\begin{bmatrix} u_{dh} \\ u_{qh} \end{bmatrix} = \bar{\bar{Z}}_h \begin{bmatrix} i_{dh} \\ i_{qh} \end{bmatrix} = \bar{\bar{Z}}_h \bar{\bar{R}}\left(\alpha_{\Psi_p}\right) \begin{bmatrix} i_{xh} \\ i_{yh} \end{bmatrix} , \tag{10}$$

where i_{dh} and i_{qh} are the high-frequency currents in the $d - q$ frame, i_{xh} and i_{yh} are the current is the stationary frame. This equation contains $\bar{\bar{Z}}_h$, the high-frequency impedance matrix of the machine, which can be derived from Eqs. (1)-(4) on the injection frequency as

$$\bar{\bar{Z}}_h = \begin{bmatrix} R + j\omega_h L_d & -\omega_{\Psi_p} L_q \\ \omega_{\Psi_p}\left(L_d + \frac{\Psi_p}{i_d}\right) & R + j\omega_h L_q \end{bmatrix} \approx \begin{bmatrix} R + j\omega_h L_d & 0 \\ 0 & R + j\omega_h L_q \end{bmatrix} = \begin{bmatrix} \bar{Z}_{h11} & 0 \\ 0 & \bar{Z}_{h22} \end{bmatrix} . \tag{11}$$

Equation (11) shows that the impedance matrix’s main diagonal contains the d- and q-direction impedances, which are only the function of the motor parameters and the injection frequency. On the other hand, the anti-diagonal elements are the function of the actual state of the system, since they depend on the common coordinate system's angular speed and the actual d-direction current. These elements can be neglected comparing to the other elements because the injection frequency expected to be much higher, than the highest possible ω_{Ψ_p}, removing the nonlinearity from Eq. (10) [13]. The measurements can only be performed in the stationary coordinate system, therefore Eq. (10) combined with Eq. (7) are also needed to be organized to express i_{xh} and i_{yh}, which results in the following,

$$\begin{bmatrix} i_{xh} \\ i_{yh} \end{bmatrix} = \bar{\bar{R}}^{-1}\left(\alpha_{\Psi_p}\right)\bar{\bar{Z}}^{-1}{}_{h}\bar{\bar{R}}(\alpha_e)\begin{bmatrix} \hat{u}_{dh} \\ \hat{u}_{qh} \end{bmatrix}, \tag{12}$$

where the current components are

$$i_{xh} = \left(\cos(\alpha_e)\cos\left(\alpha_{\Psi_p}\right)\frac{1}{\bar{Z}_{h11}} + \sin(\alpha_e)\sin\left(\alpha_{\Psi_p}\right)\frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{dh}$$

$$+\left(\sin(\alpha_e)\cos\left(\alpha_{\Psi_p}\right)\frac{1}{\bar{Z}_{h11}} - \cos(\alpha_e)\sin\left(\alpha_{\Psi_p}\right)\frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{qh}\ , \tag{13}$$

$$i_{yh} = \left(\cos(\alpha_e)\sin\left(\alpha_{\Psi_p}\right)\frac{1}{\bar{Z}_{h11}} - \sin(\alpha_e)\cos\left(\alpha_{\Psi_p}\right)\frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{dh}$$

$$+\left(\sin(\alpha_e)\sin\left(\alpha_{\Psi_p}\right)\frac{1}{\bar{Z}_{h11}} - \cos(\alpha_e)\cos\left(\alpha_{\Psi_p}\right)\frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{qh}\ . \tag{14}$$

2.3 Signal Processing

Equations (13)-(14) clearly show, that the high-frequency current response is in relation to the angle displacement between the real and estimated coordinate systems, so they can be used as an input for an angle estimator algorithm. In most cases, a Phase-Locked-Loop (PLL) structure is used, which block diagram is illustrated in Fig. 3.

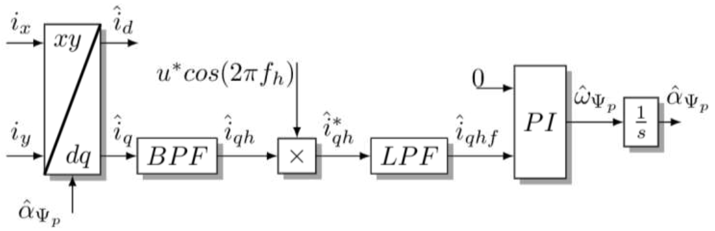


Figure 3

PLL structure used for angle estimation

In the first step of the estimation, the measurable stationary coordinate system currents are transformed into the estimated reference frame, as shown in Eq. (15). This equation combines the rotational operator and BPF block of Fig. 3, where BPF is the abbreviation for band-pass-filter, which lets through the current components around the injection frequency region.

$$\begin{bmatrix}\hat{\imath}_{dh}\\ \hat{\imath}_{qh}\end{bmatrix} = \bar{\bar{R}}\left(\hat{\alpha}_{\Psi_p}\right)\begin{bmatrix}i_{xh}\\ i_{yh}\end{bmatrix}, \tag{15}$$

where

$$\hat{\imath}_{dh} = \left(cos^2(\alpha_e)\frac{1}{\bar{Z}_{h11}} + sin^2(\alpha_e)\frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{dh}$$

$$+\ \sin(\alpha_e)\cos(\alpha_e)\left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{qh}\ , \tag{16}$$

$$\hat{\imath}_{qh} =\ \sin(\alpha_e)\cos(\alpha_e)\left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{dh}$$

$$+\ \left(sin^2(\alpha_e)\frac{1}{\bar{Z}_{h11}} + cos^2(\alpha_e)\frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{qh}\ . \tag{17}$$

Equation (17) contains both $\hat{u}_{dh}$ and $\hat{u}_{qh}$ which allows three possible solutions. In the first one, both $\hat{u}_{dh}$ and $\hat{u}_{qh}$ are used, as shown in Eq. (6). This approach is called synchronous rotating injection, for which authors in [14] proposed a demodulation algorithm. The other two solutions, which are simply called synchronous injection, use only one of the available voltages. Authors in [15] discussed the case, when the high-frequency voltage is applied in the estimated $\hat{q}$ axis, but in most of the cases only $\hat{u}_{dh}$ is injected, so Eq. (17) becomes as follows,

$$\hat{\imath}_{qh} =\ \sin(\alpha_e)\cos(\alpha_e)\left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)\hat{u}_{dh}\ , \tag{18}$$

which is a complex phasor and its equivalent time signal is shown in Eq. (19).

$$\hat{\imath}_{qh}(t) =\ u_h\sin(\alpha_e)\cos(\alpha_e)\left|\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right| sin\left(\omega_h t + arc\left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)\right) \tag{19}$$

This signal is fed into the phase detector, where it is multiplied with a cosine function having the same frequency as the injected voltage and amplitude u^*. The phase detector's output can be split into two components as shown in Eq. (20).

$$\hat{\imath}_{qh}{}^*(t) = \hat{\imath}_{qh}(t)u^*\cos(\omega_h t) =$$

$$\frac{1}{2}u_h u^*\sin(\alpha_e)\cos(\alpha_e)\left|\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right| sin\left(arc\left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)\right) +$$

$$\frac{1}{2}u_h u^*\sin(\alpha_e)\cos(\alpha_e)\left|\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right| sin\left(2\omega_h t + arc\left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)\right)\ . \tag{20}$$

One of these components is a DC-like quantity, whilst the other has twice frequency as the injection one. The latter can be filtered off using a Low-Pass-Filter, which is referred as LPF in Fig. 3, resulting $\hat{\imath}_{qhf}(t)$, as shown in Eq. (21).

$$\hat{\imath}_{qhf}(t) \approx\ \frac{1}{2}u_h u^*\sin(\alpha_e)\cos(\alpha_e)\left|\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right| sin\left(arc\left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)\right). \tag{21}$$

$\hat{\imath}_{qhf}(t)$ is fed into the PLL's PI controller as a feedback signal, and the controller's reference is set to zero. Observing Eq. (21) the feedback signal can be zero in four possible ways:

1. u_h or u^* equals zero,
2. $\left|\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right|$ becomes zero. In this case the machine is fully symmetrical in magnetic point of view, so no high-frequency signal injection methods can be applied, because Eqs. (13-17) will not contain any information from the angle error,
3. $arc\left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)$ becomes zero, which also means that the high-frequency impedances are equal,
4. $\sin(\alpha_e)\cos(\alpha_e)$ becomes zero, which means that the angle displacement between the real and estimated reference frames disappears.

This list clearly shows, that beside the trivial case when u_h or u^* is set zero, this method cannot deliver any angle information in case of machines, where the d- and q-direction high-frequency impedances are equal. Such machine could be a surface mounted permanent magnet synchronous machine, where L_d and L_q are the same. On the other hand, the zero setpoint of the controller is the result of the fourth point of this list; the angular displacement is zero between the real and estimated coordinate systems if the PI controller reaches its zero setpoint.

2.4 Dynamic Model for the PLL Structure

The tuning of PLL's controller requires a dynamic model, which can be created with the respect of the estimator structure. Equations (6)-(21) are valid in steady state, since phasors were involved in the calculations, but these results can be used to obtain the dynamic model. The proposed closed loop structure is illustrated in Fig. 4.

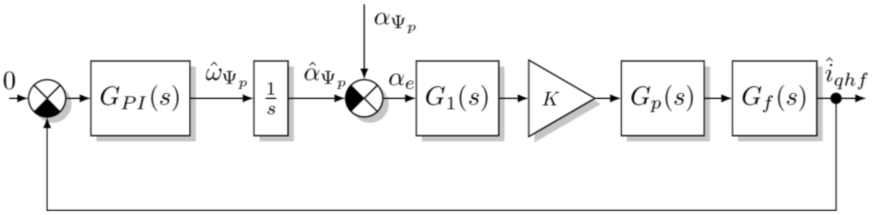


Figure 4

Dynamic model of the PLL

$G_{PI}(s)$ denotes a PI type controller, which was taken into account with the ideal form, so

$$G_{PI}(s) = A_p\left(1 + \frac{1}{sT_i}\right), \tag{22}$$

where A_p is the proportional gain, T_i is the integral time. The controller's output, which is the estimated pole-flux vector angular speed, is fed into an integrator, resulting in the estimated value of the rotor flux vector's angle. The dynamic model's upcoming parts are can be constructed using the steady state current response in Eq. (21). According to this, the real and estimated angle's difference will be the input of $G_1(s)$, which is the operator domain transfer function of the $y = \sin(u)\cos(u)$ expression. The angle error is assumed relatively small during normal operation, so the transfer function $G_1(s)$ can be modeled with unity gain, therefore

$$G_1(s) = 1. \tag{23}$$

$G_p(s)$ denotes the plant's transfer function, which is created using both d- and q-direction R-L circuits, which is depicted in Fig. 5. In order to obtain the s domain transfer function, the inductances are assumed to be non-energized in the $t = 0$ step time, and the input voltage to be a sine wave step function, so

$$u(t) = \sin(\omega_h t)\,\varepsilon(t)\,, \tag{24}$$

where $\varepsilon(t)$ is the Heaviside step function. The current response can be calculated using test functions, where the steady state current is assumed to be as follows

$$\mathrm{i}_{\mathrm{stac}}(t) = \mathrm{i}_{\mathrm{stac}}\sin(\omega_h t + \varphi)\,, \tag{24}$$

where $\mathrm{i}_{\mathrm{stac,d}} = \left|\frac{1}{\mathrm{R+j\omega_h L_d}}\right|$, $\varphi_d = arc\left(\frac{1}{\mathrm{R+j\omega_h L_d}}\right)$ in case of the d-direction impedance, and $\mathrm{i}_{\mathrm{stac,q}} = \left|\frac{1}{\mathrm{R+j\omega_h L_q}}\right|$, $\varphi_q = arc\left(\frac{1}{\mathrm{R+j\omega_h L_q}}\right)$ in case of the q-direction impedance.

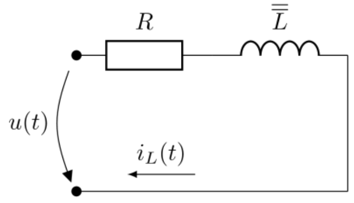


Figure 5
d- and q-direction equivalent circuit

The first order system's response is sought as shown in Eq. (25).

$$i_L(t) = \mathrm{M}e^{\lambda t} + \mathrm{i}_{\mathrm{stac}}(t), \tag{25}$$

where $\lambda_\mathrm{d} = -\frac{\mathrm{R}}{\mathrm{L_d}}$, $\lambda_\mathrm{q} = -\frac{\mathrm{R}}{\mathrm{L_q}}$ for both directions, M is a constant. With the respect of the boundary value problem, the current response time signal will be the following:

$$i_L(-0) = i_L(+0) = \mathrm{M} + \mathrm{i}_{\mathrm{stac}}\sin(\varphi) = 0, \tag{26}$$

$$M = -\mathrm{i}_{\mathrm{stac}}\sin(\varphi)\,, \tag{27}$$

$$i_L(t) = \left(-\mathrm{i}_{\mathrm{stac}}\sin(\varphi)\,\mathrm{e}^{\lambda t} + \mathrm{i}_{\mathrm{stac}}(t)\right)\varepsilon(t). \tag{28}$$

According to Eq. (18) the current response fed into the phase detector is the difference of the d- and q-direction responses, so

$$\hat{\imath}_{qh}(t) = -\mathrm{i}_{\mathrm{stac,d}}\ \sin(\varphi_d)\,\mathrm{e}^{\lambda_\mathrm{d}\mathrm{t}} + \mathrm{i}_{\mathrm{stac,q}}\ \sin(\varphi_q)\,\mathrm{e}^{\lambda_\mathrm{q}\mathrm{t}}$$

$$+\ \mathrm{i}_{\mathrm{stac,d}}\sin(\omega_h t + \varphi_d) - \mathrm{i}_{\mathrm{stac,q}}\sin(\omega_h t + \varphi_q) =$$

$$-\mathrm{i}_{\mathrm{stac,d}}\ \sin(\varphi_d)\,\mathrm{e}^{\lambda_\mathrm{d}\mathrm{t}} + \mathrm{i}_{\mathrm{stac,q}}\ \sin(\varphi_q)\,\mathrm{e}^{\lambda_\mathrm{q}\mathrm{t}}$$

$$+\ \mathrm{c}\sin(\omega_h t + \varphi_d + \varphi^*)\,, \tag{29}$$

where

$$\mathrm{c} = \sqrt{\mathrm{i}_{\mathrm{stac,d}}{}^2 + \mathrm{i}_{\mathrm{stac,q}}{}^2 - 2\mathrm{i}_{\mathrm{stac,d}}\mathrm{i}_{\mathrm{stac,q}}\cos(\varphi_q - \varphi_d)}\,, \tag{30}$$

$$\varphi^* = \tan^{-1}\left(\frac{-\mathrm{i}_{\mathrm{stac,q}}\sin(\varphi_q-\varphi_d)}{\mathrm{i}_{\mathrm{stac,d}}+\mathrm{i}_{\mathrm{stac,q}}\cos(\varphi_q-\varphi_d)}\right). \tag{31}$$

In the PLL structure we are interested in only the envelope of the current response because the PI controller tries to force it to zero. Based on Eq. (29)-(31) this can be described as follows,

$$e(\hat{\imath}_{qh}(t)) = k\left(-\mathrm{i}_{\mathrm{stac,d}}\sin(\varphi_d)\,\mathrm{e}^{\lambda_\mathrm{d}\mathrm{t}} + \mathrm{i}_{\mathrm{stac,q}}\sin(\varphi_q)\,\mathrm{e}^{\lambda_\mathrm{q}\mathrm{t}} + \mathrm{c}\sin(\varphi_d + \varphi^*)\right), \tag{32}$$

where $e(\,)$ denotes the envelope function and which equation needs to fulfill the boundary value problem and needs to have the steady state amplitude c. This will result $k = \frac{1}{\sin(\varphi_d+\varphi^*)}$, so

$$e(\hat{\imath}_{qh}(t)) = -\frac{\mathrm{i}_{\mathrm{stac,d}}\ \sin(\varphi_d)}{\sin(\varphi_d+\varphi^*)}\mathrm{e}^{\lambda_\mathrm{d}\mathrm{t}} + \frac{\mathrm{i}_{\mathrm{stac,q}}\ \sin(\varphi_q)}{\sin(\varphi_d+\varphi^*)}\mathrm{e}^{\lambda_\mathrm{q}\mathrm{t}} +\ \mathrm{c}\,. \tag{33}$$

Equation (33) is the step response of the system, so its transfer function can be calculated as $s\mathcal{L}\left(e(\hat{\imath}_{qh}(t))\right)$, which results in

$$G_p(s) = \frac{p_1 s^2 + p_2 s + p_3}{(s-\lambda_\mathrm{d})(s-\lambda_\mathrm{q})}\,, \tag{34}$$

where

$$p_1 = -\frac{\mathrm{i}_{\mathrm{stac,d}}\sin(\varphi_d)}{\sin(\varphi_d+\varphi^*)} + \frac{\mathrm{i}_{\mathrm{stac,q}}\ \sin(\varphi_q)}{\sin(\varphi_d+\varphi^*)} + \mathrm{c} = 0\,, \tag{35}$$

$$p_2 = \frac{\mathrm{i}_{\mathrm{stac,d}}\sin(\varphi_d)}{\sin(\varphi_d+\varphi^*)}\lambda_\mathrm{q} - \frac{\mathrm{i}_{\mathrm{stac,q}}\ \sin(\varphi_q)}{\sin(\varphi_d+\varphi^*)}\ \lambda_\mathrm{d} - \mathrm{c}\lambda_\mathrm{d} - \mathrm{c}\lambda_\mathrm{q}, \tag{36}$$

$$p_3 = \mathrm{c}\lambda_\mathrm{d}\lambda_\mathrm{q}. \tag{37}$$

In the control structure shown in Fig. 4 K remained as a yet unknown gain. This gain can be defined based on the PLL structure, especially its phase detector part and Eq. (21). The $\left|\frac{1}{\bar{Z}_{h11}}-\frac{1}{\bar{Z}_{h22}}\right|$ part of this equation equals c, whilst sin $(\alpha_e)\cos(\alpha_e)$ part of it was dealt with $G_1(s)$ above. The rest of this equation will be the K constant, so

$$K=\frac{1}{2}u_h\, u^{*}sin\left(arc\left(\frac{1}{\bar{Z}_{h11}}-\frac{1}{\bar{Z}_{h22}}\right)\right)=\frac{1}{2}u_h\, u^{*}sin(\varphi_d+\varphi^{*}) \ . \tag{38}$$

$G_f(s)$ denotes the loop filter of the PLL, which is used to a be a simple low pass filter, so the open loop transfer function of the dynamic model can be modeled as shown in Eq. (39).

$$G_o(s)=-KA_p\left(1+\frac{1}{sT_i}\right)\frac{1}{s}\,\frac{p_1s^2+p_2s+p_3}{(s-\lambda_\mathrm{d})(s-\lambda_\mathrm{q})}\frac{1}{sT_f+1}\ , \tag{39}$$

where $G_o(s)$ is the open loop transfer function, T_f is the low-pass filter's time constant.

3 Simulation Results

The simulations were carried out on a permanent magnet synchronous machine and its parameters are listed in Table 1.

Table 1

Permanent Magnet Synchronous Machine Parameters

Parameter	Value
P_n	$1.1kW$
U_n	$400V$
i_n	$2.53A$
R	6.2Ω
L_d	$20.025mH$
L_q	$40.17mH$
Ψ_p	$0.305Vs$
p	3
F	$0.0011Nms$
m_n	$3.5Nm$

Fig. 6 illustrates the cascade PI controller loops, where the ref subscript refers to the reference values. The tuning of the cascade control loop's PI controllers, which are taken into account using Eq. (22), are tuned based on the predefined cut off frequency and the damping constant methodology [16].

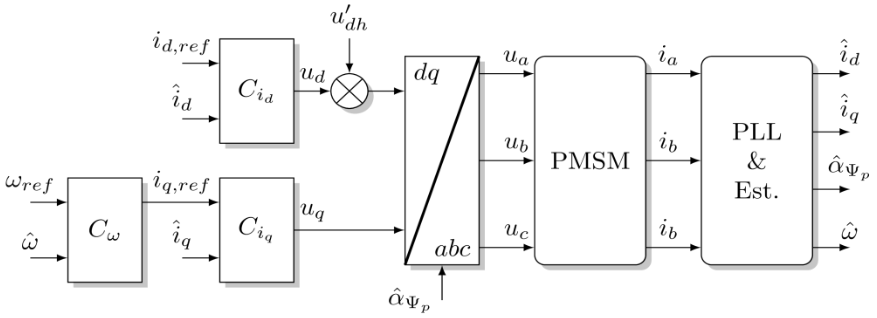


Figure 6

Cascade PI control loop

The overall performance of the whole cascade control loop depends on the estimator's PI controller, which can be tuned using the proposed dynamic model. In the first step of the simulation and tuning process, the injection frequency f_h and the voltage amplitude u_h must be defined with the respect of the highest possible ω_{Ψ_p}. Figure 7 gives an overview of the effect of the selected frequency. The higher the injection frequency, the higher bandwidth can be achieved, but the signal amplitudes with which the PLL structure operates becomes smaller.

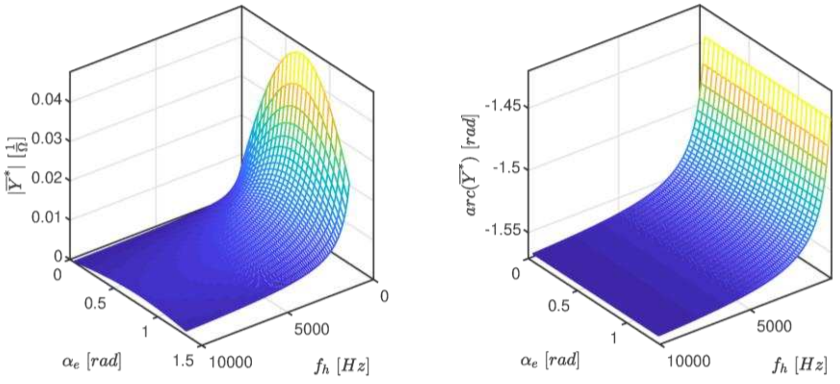


Figure 7

The amplitude and the phase of the response, where $\bar{Y}^* = \frac{1}{2} u_h u^* \sin(\alpha_e) \cos(\alpha_e) \left(\frac{1}{\bar{Z}_{h11}} - \frac{1}{\bar{Z}_{h22}}\right)$

The selected machine had 3 pole-pairs and nominal mechanical angular frequency of 50 Hz, therefore the injection frequency was chosen to be 1500 Hz, one decade higher than the expected highest possible ω_{Ψ_p} in normal operation. The high frequency voltage amplitude was 40 V and the phase detector's u^* was set to 1. During the simulations, no field weakening was examined.

Based on Table 1, the process' Bode diagram including the K gain is illustrated in Fig. 8. To obtain a stable control of the estimator, its PI controller's parameters were chosen to achieve 60 degrees phase margin [17] in the open loop Bode diagram, which is shown in Fig. 9.

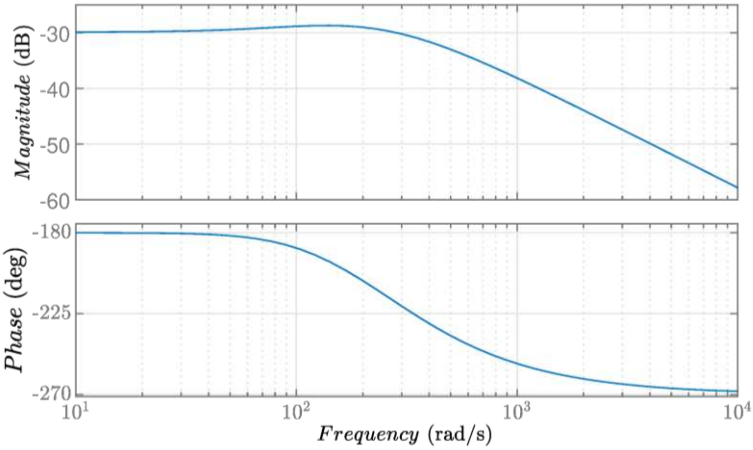


Figure 8

The process' Bode diagram

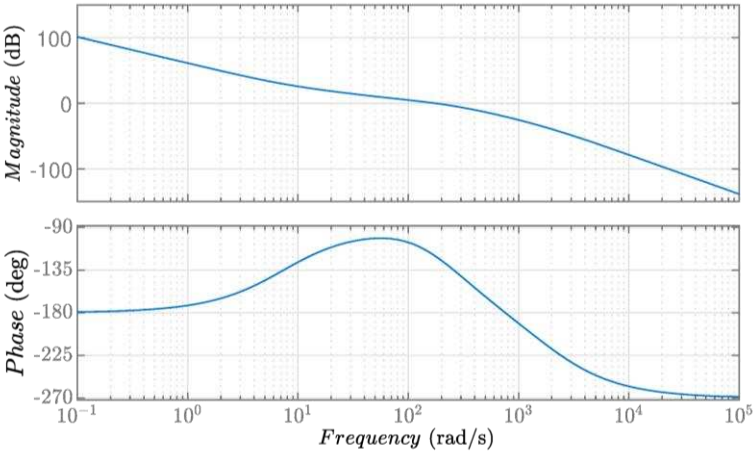


Figure 9

Open loop Bode diagram of the $G_o(s)$ PLL structure

The predefined phase margin belongs to 175 rad/s cut-off frequency in Fig. 9, so the speed control loop, which uses the estimator's angular velocity, must be tuned to be slower than this. Table 2 summarizes the PI controller and filter parameters, which were used in the cascade PI control loops and the estimator.

Table 2

Controller parameters and filter time constants

Parameter	Value	Description
$A_{p,PLL}$	4853	PLL's proportional gain
$T_{i,PLL}$	$136ms$	PLL's integral time
$A_{p,d}$	11.49	d-direction current controller proportional gain
$T_{i,d}$	$1.8ms$	d-direction current controller integral time
$A_{p,q}$	23.92	q-direction current controller proportional gain
$T_{i,q}$	$4.2ms$	q-direction current controller integral time
$A_{p,\omega}$	0.287	angular speed controller proportional gain
$T_{i,\omega}$	$36ms$	angular speed controller integral time
T_f	$0.53ms$	PLL's low-pass filter time constant

With these setting of the controllers and filters the following simulation results were obtained.

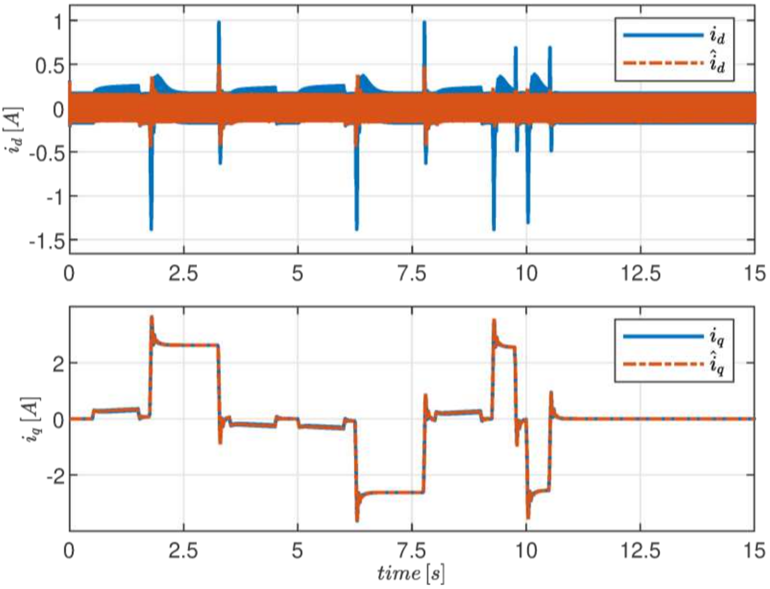


Figure 10

Simulated d- and q-direction currents

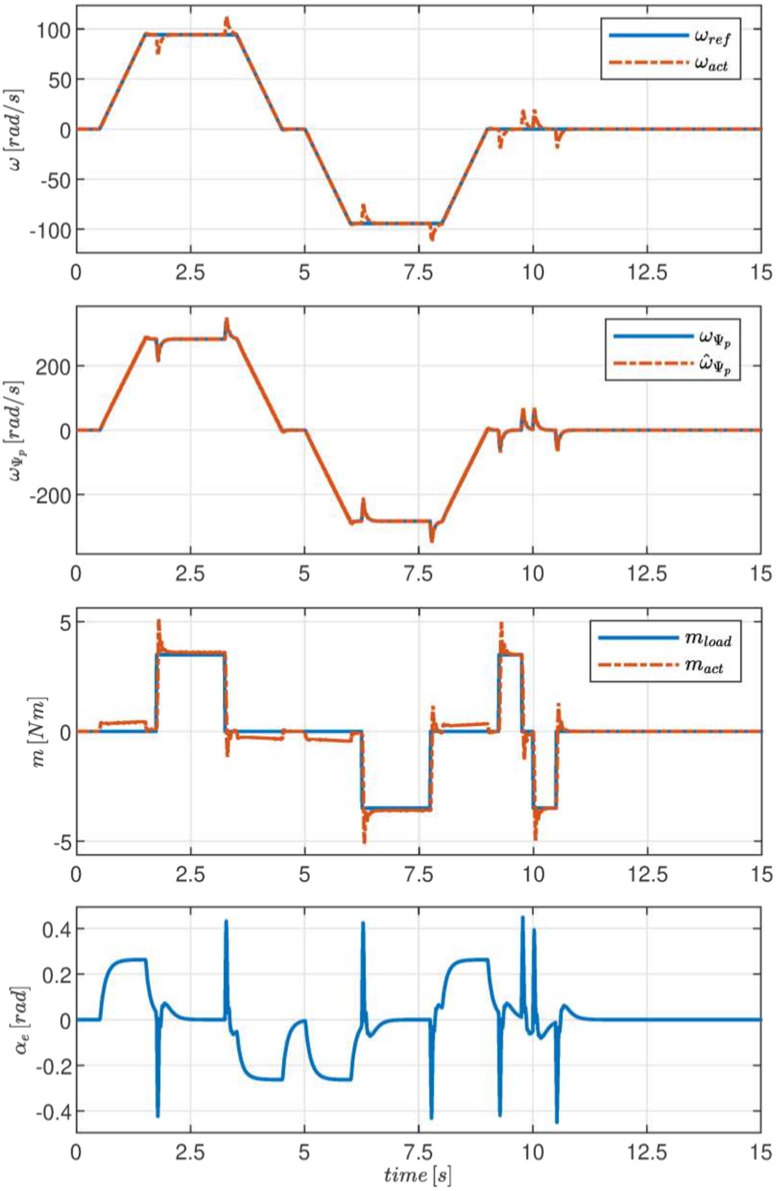


Figure 11

Simulated angular speed, rotor flux vector's angular speed, torque and angular displacement between the real and estimated coordinate systems

Figures 10 and 11 summarize the simulation results, including the actual, referred with act subscripts, and estimated signals of the machine. The speed-load profile was chosen to have positive and negative direction loads and angular speeds and zero-frequency cases with active load. The comparison of the field currents shows that the injection made its influence felt mainly in the flux branch of the system, which confirms that the injection voltage had only d-direction voltage component.

The implemented speed-sensorless algorithm was able to follow the speed and the torque requests, where the nominal torque was applied. The angle error also correlated with the control structure, since during dynamic changes, when the angular speed reference changes or torque is applied, the displacement between the real and estimated coordinate systems must occur, because it feeds the estimator's PI controller.

Conclusions

This paper presented the high-frequency synchronous voltage injection method on a permanent magnet synchronous machine. The mathematical model of this method was described, where it was shown that the method is unable to provide any angle information if the d- and q- direction impedances are equal. The angle estimation depended on a PLL structure, for which a new dynamic model was proposed. With the help of this model, the PLL's PI type controller was tuned, which purpose is to force down its feedback signal to zero, in which condition the estimated and real coordinate systems' angle are the same. After tuning the estimator's controller, the widely used cascade control loop could be tuned and its performance could be analyzed.

The proposed dynamic model can be adapted to other types of machines, such as to squirrel-cage induction machines, synchronous reluctance machines. The ongoing research focuses on the application of high-frequency synchronous injection on different machines. High-frequency stationary injection is also investigated, where the Authors would like to report a comparison. Measurement results will be reported soon using a PWM IGBT inverter [18].

References

[1] M. Eskola and H. Tuusa: Comparison of MRAS and novel simple method for position estimation in PMSM drives, IEEE 34th Annual Conference on Power Electronics Specialist, 2003, PESC '03, Acapulco, Mexico, 2003, pp. 550-555 Vol. 2

[2] T. Orlowska-Kowalska and M. Dybkowski, Stator-current-based mrasestimator for a wide range speed-sensorless induction-motor drive, IEEE Transactions on industrial electronics, Vol. 57, No. 4, pp. 1296-1308, 2010

[3] H. M. Kojabadi and M. Ghribi, MRAS-based adaptive speed estimator in PMSM drives, 9th IEEE International Workshop on Advanced Motion Control, 2006, Istanbul, 2006, pp. 569-572

[4] K. Wiedmann and A. Mertens, Self-sensing control of PM synchronous machines including online system identification based on a novel MRAS

approach, 3rd IEEE International Symposium on Sensorless Control for Electrical Drives (SLED 2012), Milwaukee, WI, 2012, pp. 1-8

[5] S. Bolognani, R. Oboe and M. Zigliotto, Sensorless full-digital PMSM drive with EKF estimation of speed and rotor position, IEEE Transactions on Industrial Electronics, Vol. 46, No. 1, pp. 184-191, Feb. 1999

[6] S. Bolognani, L. Tubiana and M. Zigliotto, Extended Kalman filter tuning in sensorless PMSM drives, IEEE Transactions on Industry Applications, Vol. 39, No. 6, pp. 1741-1747, Nov.-Dec. 2003

[7] M. Schroedl, Sensorless control of AC machines at low speed and standstill based on the "INFORM" method, IAS '96. Conference Record of the 1996 IEEE Industry Applications Conference Thirty-First IAS Annual Meeting, San Diego, CA, USA, 1996, pp. 270-277, Vol. 1

[8] J. M. Liu and Z. Q. Zhu, Sensorless Control Strategy by Square-Waveform High-Frequency Pulsating Signal Injection Into Stationary Reference Frame, IEEE Journal of Emerging and Selected Topics in Power Electronics, Vol. 2, No. 2, pp. 171-180, June 2014

[9] S. Kim, J. Im, E. Song and R. Kim, A New Rotor Position Estimation Method of IPMSM Using All-Pass Filter on High-Frequency Rotating Voltage Signal Injection, IEEE Transactions on Industrial Electronics, Vol. 63, No. 10, pp. 6499-6509, Oct. 2016

[10] D. Raca, P. Garcia, D. Reigosa, F. Briz and R. Lorenz, A comparative analysis of pulsating vs. rotating vector carrier signal injection-based sensorless control, 2008 Twenty-Third Annual IEEE Applied Power Electronics Conference and Exposition, Austin, TX, 2008, pp. 879-885

[11] Hyunbae Kim and R. D. Lorenz, Carrier signal injection based sensorless control methods for IPM synchronous machine drives, Conference Record of the 2004 IEEE Industry Applications Conference, 2004. *39th IAS Annual Meeting.*, Seattle, WA, USA, 2004, pp. 977-984, Vol. 2

[12] Y. Yoon, S. Sul, S. Morimoto and K. Ide, High-Bandwidth Sensorless Algorithm for AC Machines Based on Square-Wave-Type Voltage Injection, IEEE Transactions on Industry Applications, Vol. 47, No. 3, pp. 1361-1370, May-June 2011

[13] Pozna, Claudiu, and Radu-Emil Precup. An Approach to the Design of Nonlinear State-Space Control Systems. Studies in Informatics and Control 27.1 (2018): 5-14

[14] C. Caruana, G. M. Asher, K. J. Bradley and M. Woolfson, Flux position estimation in cage induction machines using synchronous HF injection and Kalman filtering, IEEE Transactions on Industry Applications, Vol. 39, No. 5, pp. 1372-1378, Sept.-Oct. 2003

[15] M. J. Corley and R. D. Lorenz, Rotor position and velocity estimation for a salient-pole permanent magnet synchronous machine at standstill and high speeds, IEEE Transactions on Industry Applications, Vol. 34, No. 4, pp. 784-789, July-Aug. 1998

[16] Wang, Liuping, et al. PID and predictive control of electrical drives and power converters using MATLAB/Simulink. John Wiley & Sons, 2015

[17] Kiam Heong Ang, G. Chong and Yun Li, PID control system analysis, design, and technology, IEEE Transactions on Control Systems Technology, Vol. 13, No. 4, pp. 559-576, July 2005

[18] Vukosavic, Slobodan N. Digital control of electrical drives. Springer Science & Business Media, 2007

Generalized Quasi-Orthogonal Polynomials Applied in Sliding Mode-based Minimum Variance Control of ABS

Staniša Perić, Dragan Antić, Darko Mitić, Saša Nikolić, Marko Milojković

University of Niš, Faculty of Electronic Engineering, Department of Control Systems, Aleksandra Medvedeva 14, 18000 Niš, Republic of Serbia
stanisa.peric@elfak.ni.ac.rs, dragan.antic@elfak.ni.ac.rs,
darko.mitic@elfak.ni.ac.rs, sasa.s.nikolic@elfak.ni.ac.rs,
marko.milojkovic@elfak.ni.ac.rs

Abstract: This paper deals with the design of a digital sliding mode based minimum variance control on the basis of a discrete-time representation of the ABS model derived from a new type of generalized quasi-orthogonal filter. In the proposed control, the minimum variance enables the design of digital sliding mode control only on the basis of ABS outputs measurements, while sliding mode increases the ABS robustness under certain conditions. On the other hand, it is shown that orthogonal functions can be successfully used to obtain a model of a dynamical system with high accuracy. The proposed control scheme has been applied in the laboratory experimental setup and obtained experimental results show significant improvement in ABS performances.

Keywords: sliding mode control; minimum variance control; orthogonal polynomials; anti-lock braking system

1 Introduction

Nowadays, we are witnessing the tremendous growth of the automotive industry in the world. Unfortunately, an increasing number of traffic accidents occur due to improper vehicle speed, sudden braking, bad road conditions, etc. The anti-lock braking system (ABS) is just one of the modern electronic systems found in vehicles, which contributes to the reduction of these accidents. It prevents the loss of control over the vehicle during sudden braking by disabling the vehicle wheels blocking in different road conditions (ice, snow, water, sand, etc.).

It has been shown that the ABS control problem can be solved by using different control approaches, starting from classical PID, through fuzzy logic and sliding mode controllers up to advanced control techniques based on the use of artificial neural networks, machine learning, etc. Herein, an overview of the most important contributions in sliding mode control (SMC) of ABS is given since this paper only deals with this type of control. In [1], SMC based on the exponential reaching law for ABS is developed to maintain the optimal slip value. The authors developed a two-wheel vehicle model in [2] and proposed an SMC algorithm to regulate ABS. The objective of [3] is to modify an optimal SMC method for hydraulic ABS in order to achieve both robustness and optimal control performances. SMC using a grey system theory approach for the ABS control is considered in [4, 5]. The wheel slip control of the traction control system using a moving sliding surface is presented in [6]. One more approach of the moving sliding surface implementation for vehicle slip ratio control is shown in [7]. In [8], SMC on the basis of a two-axle vehicle model is discussed. In that paper, the authors also introduced the integral switching surface to cope with the chattering phenomenon. The similar approach of using the integral sliding surface is given in [9], but this time for the hybrid electric brake system. The traditional approach of the SMC design is applied in the control of the magnetorheological brake system in [10]. One more application of the traditional SMC is proposed in [11]. The second order sliding mode using the super-twisting technique to manipulate the braking torque is introduced in [12]. The quasi-continuous control for an automobile ABS is proposed in [13]. Therein, two controllers are developed, one to realize slip control and others for pressure tracking control. In [14], the same authors gave a more detailed analysis of the previous approach and compared the obtained results with the traditional SMC approach. Unlike the previous papers, where the wheel slip is used as a controlled variable, the authors of [15] considered a different approach based on the sliding mode by using the slip velocity. The latter approach is simulated on a “Magic formula” tire model.

From the previous analysis, it can be concluded that the further improvement of ABS performances can be realized by new control algorithms. Therefore, in this paper, an attempt will be made to achieve the optimum slip value by applying the novel control law that results from the combination of the sliding mode based minimum variance (MV) control and ABS modelling with orthogonal functions. It that way, the maximal value for the road adhesion coefficient is provided leading to better vehicle steering characteristics.

In the past several years the authors of this paper developed the new types of orthogonal filters [16], almost orthogonal filters [17, 18], improved almost orthogonal filters [19], quasi-orthogonal filters [20], orthogonal filters with complex zeros and poles [21] and generalized quasi-orthogonal filters [22]. These filters have proven to be a very powerful mathematical tool for modeling [18] and control of dynamical systems [16, 23, 24], as well as, for the approximation of real signals generated by industrial systems [19, 21]. These filters can be also used to analyze the sensitivity of models of complex dynamical systems [25]. On the other hand, the signals generated by the generalized quasi-orthogonal filters of *k*-th order can be successfully applied as the activation functions of neural networks [22, 24]. Moreover, it has been shown in [26] that these functions can replace the functions inside the layer that imitates Sugeno style defuzzification in the traditional ANFIS network. In [27], it has been already proved that orthogonal models, obtained by almost orthogonal filters, can be very effective for the design of SMC in the continuous-time domain. Herein, a similar concept for the new type of quasi-orthogonal filters, specially designed for this purpose, will be implemented, but this time in the discrete-time domain. The advantage of the proposed generalized quasi-orthogonal filters of shifted Müntz-Legendre type comes from the fact that they have the general values for poles in a transfer function, which significantly expands the possibility of their applications in comparison to the other, previously developed filters. This filter is used for obtaining the model of a plant, which will be employed then as a reference model in the design of SMC on the basis of MV control. The similar concepts of obtaining several linear and nonlinear models with successful applications in various fields including control are presented in [28, 29, 30].

The main goal of the combination of the sliding mode and MV control techniques is to improve the individual good characteristics of two control methods and to suppress their main drawbacks. The MV control can provide the realization of the digital sliding mode control (DSMC) only on the basis of the sensed system outputs. On the other hand, DSMC has been introduced to improve the robustness of MV control under the influence of external disturbances and parameter perturbations. It has been shown in [31] that digital sliding mode based MV control with accuracy $O(T^2)$ can be obtained by introducing the relay component into the control law, previously filtered through a digital integrator. The presence of a digital integrator significantly reduces the undesirable chattering phenomenon [32] providing a relatively smooth control signal in that way.

The laboratory setup of ABS is used in this paper [33]. The experimental results show the effectiveness of the proposed type of filters and control method in the field of modeling and control of ABS, respectively. The chosen laboratory framework proves to be very suitable for testing of different control algorithms [34, 35, 36]. By implementing the same control law with the previously derived model of ABS [27] and with the model derived in this paper, the obtained experimental results favour the latter control algorithm. In other words, the obtained orthogonal model describes the considered plant in a more efficient way due to the introduced parameter of imperfections in the very definition of the filter. From the control point of view, both experiments confirm the effectiveness of the proposed robust control method. The vehicle stopping time is further shortened with the preserved steering control, leading to the increased safety of the passengers.

The paper is organized as follows. Section 2 describes a new type of quasi-orthogonal filter. In Section 3, the basic mathematical background of the proposed digital sliding mode based MV control is presented. The orthogonal model of ABS and modified control law in the case of ABS, are given in Section 4. The sensed outputs are presented and discussed in Section 5. Section 6 gives the most important concluding remarks.

2 Generalized Quasi-Orthogonal Filters of Müntz-Legendre Type

In this section, the generalized quasi-orthogonal filters of shifted Müntz-Legendre type (GQOFMLT), which contain an imperfections measure δ in their definition, are derived. This parameter actually describes imperfections of all the elements the system consists of, imperfections in the model, impact of the noise on the system output etc. [17, 22].

Now, let us consider a transfer function, $W_n^{(k,\delta)}(s)$, in the form suitable for the practical design of the proposed *k*-th order filter:

$$W_n^{(k,\delta)}(s) = K\frac{\prod_{i=1}^{n-k}(s-p_i\delta)}{\prod_{i=0}^{n}(s+p_i)} = K\frac{(s-p_1\delta)(s-p_2\delta)\cdots(s-p_{n-k}\delta)}{s(s+p_1)(s+p_2)\cdots(s+p_n)}, \tag{1}$$

where p_i represent poles of the transfer function, and K is a gain of filter. The value of the constant δ is determined by performing several experiments so that it reflects the rate of parameters modification due to the changes in working temperature, humidity, etc. [19, 22]. The main idea, herein, is to use this free parameter for description of non-modelled dynamics of a plant, obtaining in that way more faithful model representation. On the other hand, the effectiveness of control method heavily depends on model accuracy.

Development of (1) in partial fractions results in:

$$W_n^{(k,\delta)}(s) = K\sum_{i=0}^{n}\frac{A_{n,i}^{(k,\delta)}}{(s+p_i)}, \tag{2}$$

where the coefficients $A_{n,i}^{(k,\delta)}$ are calculated as:

$$A_{n,i}^{(k,\delta)} = K\frac{\prod_{j=1}^{n-k}(s_j-p_i\delta)}{\prod_{\substack{j=0\\ j\neq i}}^{n}(s_j+p_i)}. \tag{3}$$

By using the transformation mapping $f(s) = \overline{s} = -s$ [16], the poles p_i are being mapped into the zeros located in the right semi plane. After applying the inverse Laplace transform to (2), the sequence of orthogonal rational functions in time domain is derived as:

$$\varphi_i^{(k,\delta)}(t) = \sum_{i=0}^{n} A_{n,i}^{(k,\delta)} e^{-p_i t} \ . \tag{4}$$

By taking e^{-t} as a member x of the polynomial, the latter relation can be rewritten as:

$$L_n^{(k,\delta)}(x) = \sum_{i=0}^{n} A_{n,i}^{(k,\delta)} x^{p_i} \ , \tag{5}$$

where $L_n^{(k,\delta)}(x) = L^{-1}\left\{\varphi_i^{(k,\delta)}(t)\right\}$ represent the k-th order generalized quasi-orthogonal polynomials of the shifted Müntz-Legendre type.

Note that if the poles p_i in (2) have integer values, i.e., $p_i \in Z, \ i = 0,1,\ldots,n$, someone can get the generalized quasi-orthogonal k-th order polynomials of Legendre type $P_n^{(k,\delta)}(x)$ [22], defined with the following expression:

$$P_n^{(k,\delta)}(x) = \sum_{i=0}^{n} A_{n,i}^{k,\delta} x^i \ , \tag{6}$$

where:

$$A_{n,i}^{k,\delta} = (-1)^{n+i+k} \frac{\prod_{j=1}^{n-k}(i + j\delta)}{i!(n-i)!} \ . \tag{7}$$

The main difference between these two filter types lies in the fact that poles in the transfer function which correspond to (7) are fixed and a priori known. On the other hand, any real constant values for the poles can be selected in (1) causing in that way greater possibility of applications for the proposed type of filters.

On the basis of (1), it is easy to obtain the structure of the first order GQOFMLT ($k = 1$) which is very suitable for practical realization [18, 19, 20]. The general structure is shown in Figure 1, where Heaviside function is used as an input signal.

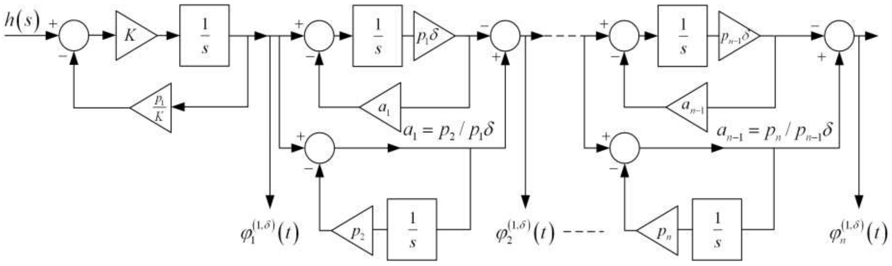

$h(s)$
K
$\frac{1}{s}$
$\frac{P_1}{K}$
$p_1\delta$
a_1
$a_1 = p_2 / p_1\delta$
p_2
$\varphi_1^{(1,\delta)}(t)$
$\varphi_2^{(1,\delta)}(t)$
$p_{n-1}\delta$
a_{n-1}
$a_{n-1} = p_n / p_{n-1}\delta$
p_n
$\varphi_n^{(1,\delta)}(t)$

Figure 1
Schematic representation of the proposed filter for the first order (k=1)

It has been already shown that orthogonal signals $\varphi_i^{(1,\delta)}(t)$, generated using the orthogonal filters, are appropriate for the analysis and synthesis of different technical systems [21, 22, 25]. In this case, the parameters of filter (K, p_i, and c_i) are adjusted to minimize the value of the mean squared error:

$$J = \frac{1}{T}\int_0^T \left(y_S(t) - y_M(t)\right)^2 dt\,, \tag{8}$$

where $y_S(t)$ is a real system output and $y_M(t)$ represents an output of the orthogonal model described by the following expression:

$$y_M(t) \approx \sum_{i=0}^{n} c_i \varphi_i^{(1,\delta)}(t)\,. \tag{9}$$

On the basis of these plant models and their representation in the discrete-time domain, a digital sliding mode controller is designed in the next section. The model parameters are obtained by identification based on the measured ABS responses, using some of the optimization techniques. The procedure itself is briefly described in Section 4, while the entire modelling process, using genetic algorithms as an optimization method, is thoroughly described in [18, 19].

3 Digital Sliding Mode-based Minimum Variance Control

Let us consider a continuous-time SISO plant model in the following form:

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) + \mathbf{d}f(t), \\ y(t) &= \mathrm{c}\mathbf{x}(t), \end{aligned} \tag{10}$$

where $\mathbf{x}(t)=\left[x_1(t) \quad x_2(t) \quad \dots \quad x_n(t)\right]^T \in R^n$ is a vector of state coordinates, $u(t)\in R$ represents a plant input, $f(t)\in R$ is an external disturbance, $y(t)\in R$ denotes a plant output, n determines plant order, and matrix $\mathbf{A}$ and vectors, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{d}$ are with the following dimensions: $\mathbf{A}=\left[a_{ij}\right]_{n\times n}$, $\mathbf{b}=\left[b_i\right]_{n\times 1}$, $\mathbf{c}=\left[c_j\right]_{1\times n}$, $\mathbf{d}=\left[d_i\right]_{n\times 1}$. This model can be directly derived from (9) taking into account (1)-(4).

The discrete-time model of (10) is given by:

$$\begin{aligned}\dot{\mathbf{x}}_{k+1} &= \boldsymbol{\phi}\mathbf{x}_k + \boldsymbol{\gamma}u_k + \mathbf{h}_k,\\ y_k &= \mathbf{c}\mathbf{x}_k,\end{aligned} \tag{11}$$

whereas:

$$\begin{aligned}\boldsymbol{\phi} &= e^{\mathbf{A}T},\\ \boldsymbol{\gamma} &= \int_0^T e^{\mathbf{A}\tau}\mathbf{b}d\tau,\\ \mathbf{h}_k &= \int_0^T e^{\mathbf{A}\tau}\mathbf{d}f\left((k+1)T-\tau\right)d\tau.\end{aligned} \tag{12}$$

To enhance further writing, the following notation $\bullet_k = \bullet(kT)$ is accepted, where T denotes the sampling period. The external disturbance $f(t)$ is supposed to be a bounded function, i.e., there exists a constant $F<\infty$ such that $\left|f(t)\right|<F$. On the basis of (11), the plant model in the z-domain can be obtained as:

$$y_k = \frac{z^{-1}B\left(z^{-1}\right)}{A\left(z^{-1}\right)}u_k + \frac{z^{-1}\mathbf{D}\left(z^{-1}\right)}{A\left(z^{-1}\right)}\mathbf{h}_k, \tag{13}$$

where z^{-1} represents a delay operator, i.e., $z=e^{pT}$ (p is a complex variable), and:

$$A\left(z^{-1}\right) = z^{-n}\det\left(z\mathbf{I}-\boldsymbol{\phi}\right), \tag{14}$$

$$B\left(z^{-1}\right) = z^{-n+1}\mathbf{c}\left[\operatorname{adj}\left(z\mathbf{I}-\boldsymbol{\phi}\right)\boldsymbol{\gamma}\right], \tag{15}$$

$$\mathbf{D}\left(z^{-1}\right) = z^{-n+1}\mathbf{c}\left[\operatorname{adj}\left(z\mathbf{I}-\boldsymbol{\phi}\right)\right]. \tag{16}$$

The main objective of the designed control is to ensure minimum variance of the variable:

$$s_k = M\left(z^{-1}\right)\left(y_k - r_k\right), \tag{17}$$

i.e., in an ideal case $s_k = 0$. In addition, the polynomial $M\left(z^{-1}\right)$ is a Jury's polynomial, and r_k is a reference input signal in the *k*-th time period. The plant output in the steady state can be defined with:

$$y_\infty = r_\infty + \frac{s_\infty}{M(1)}. \tag{18}$$

From the last equation it can be concluded that the accuracy of the system output will depend only on the accuracy of the variable s_k. Therefore, by keeping the smallest value of s_k, the smallest possible tracking error will be achieved.

To accomplish the above-mentioned control goal, the digital sliding mode based MV control is proposed in the following form [31]:

$$u_k = -\frac{1}{E\left(z^{-1}\right)B\left(z^{-1}\right)}\left(F\left(z^{-1}\right)y_k - M\left(z^{-1}\right)r_{k+1} + \frac{\alpha T}{1-z^{-1}}\operatorname{sgn}\left(s_k\right)\right), \tag{19}$$

where $E\left(z^{-1}\right)$ and $F\left(z^{-1}\right)$ are the solutions of Diophantine equation:

$$E\left(z^{-1}\right)A\left(z^{-1}\right) + z^{-1}F\left(z^{-1}\right) = M\left(z^{-1}\right), \tag{20}$$

with assumption that r_k is known in advance. The digital integrator (in the front of sgn (s_k)) in (19) should alleviate the chattering phenomenon [32, 37]. By substituting (19) in (13), and taking into account (17) and (20), the switching function dynamics can be obtained as:

$$s_{k+1} = s_k - \alpha T \operatorname{sgn}\left(s_k\right) + E\left(z^{-1}\right)\mathbf{D}\left(z^{-1}\right)\left(\mathbf{h}_k - \mathbf{h}_{k-1}\right), \tag{21}$$

where parameter α should provide stable quasi-sliding motion and it is chosen in accordance with the following theorem.

Theorem: Let us consider the system described by (13) and (19), where the switching function and its dynamic is given by (17) and (21), respectively. If the parameter α is chosen to satisfy:

$$\alpha T > \max\left|E\left(z^{-1}\right)\mathbf{D}\left(z^{-1}\right)\left(\mathbf{h}_k - \mathbf{h}_{k-1}\right)\right|, \tag{22}$$

then the control (19) forces a system phase trajectory to reach the quasi-sliding manifold *S* determined by:

$$S = \left\{s_k : \left|s_k\right| < \alpha T + \max\left|E\left(z^{-1}\right)\mathbf{D}\left(z^{-1}\right)\left(\mathbf{h}_k - \mathbf{h}_{k-1}\right)\right|\right\}, \tag{23}$$

in finite time and keeps it on it for every $k > K_0$, where $K_0 = K_0\left(s_0\right)$ is a positive number.

The detailed proof of this Theorem could be found in [38].

4 A Case Study: Anti-Lock Braking System

As a case study, a laboratory ABS, presented in Figure 2 has been chosen, because it is suitable for the practical verification of the proposed modelling and control approaches due to its strong nonlinear nature [39, 40, 41].

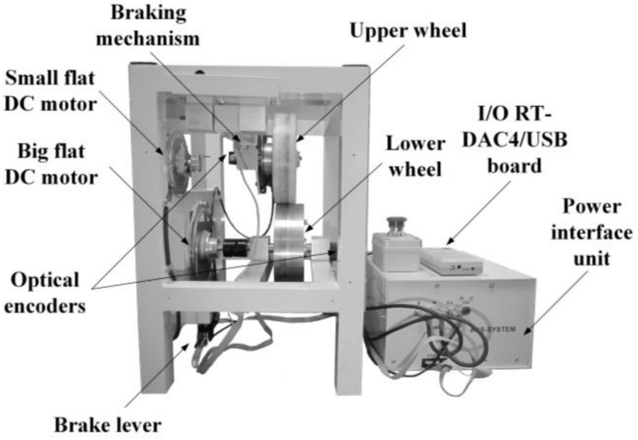


Figure 2
Laboratory test setup of ABS

The complete mathematical and physical description of this system can be found in our earlier papers [24, 42] and it is derived on the basis of graphical representation of ABS setup presented in Figure 3.

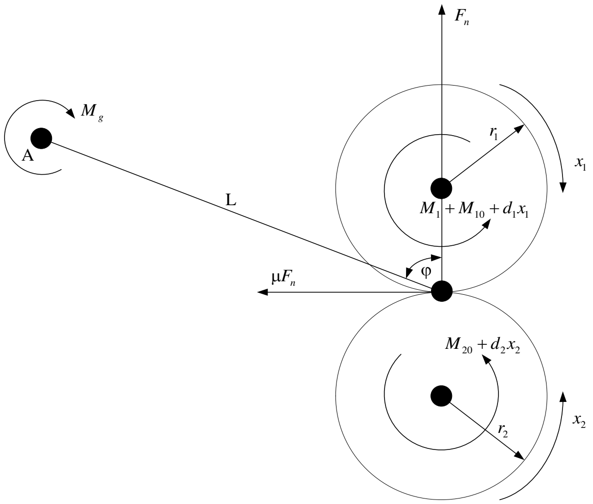


Figure 3
Schematic diagram of the experimental setup

Based on this first-order nonlinear model, the DSMC has been designed in [42].

In [43], it has been shown that the dynamics of ABS can be represented well enough by the second-order transfer function with finite zero:

$$W(s) = \frac{\overline{b}_1 s + \overline{b}_0}{s^2 + \overline{a}_1 s + \overline{a}_0} \tag{27}$$

where wheel slip and braking moment are used as system output and input, respectively. In order to control such minimum-phase plant, the starting model (27) is transformed into the controllable canonical form first and then divided into two subspaces [44]:

$$\begin{aligned} \dot{x}_1 &= d_{11}x_1 + d_{12}x_2 + ku, \\ \dot{x}_2 &= d_{21}x_1 + d_{22}x_2, \\ y &= x_1, \end{aligned} \tag{28}$$

On the basis of the latter model, SMC is designed for the upper subsystem (28) of the first order in [27]. Notice that the control approaches presented so far use the first-order model of ABS for design purposes. It should be expected that the use of the second-order model of ABS in controller design would give better results.

In this paper, several real-time experiments have been performed and wheel slip for different values of braking moment has been recorded. After that, the parameters of the second-order model (27) have been identified by using the classical approach as $\overline{a}_0$=0.7708, $\overline{a}_1$=74.3301, $\overline{b}_0$=0.0059, and $\overline{b}_1$=0.6840. This model will be used further for the validation purposes of the proposed control approach. The same experimental results are used to obtain the suggested orthogonal model parameters. Based on the plant response, it has been concluded that the ABS model can be described by proposed GQOFMLT with two sections according to (1), (4), and (8). As mentioned earlier, for parameters adjustment of filter (K, p_1, p_2, c_0, c_1, and c_2) the genetic algorithm [19, 21, 22] can be used. Herein, the genetic algorithm has a chromosome consisting of 6 parameters coded by real numbers (filter parameters) and fitness function (8). After the parameters adaptation procedure, K=0.0065, p_1=76.9230, p_2=0.0103, c_0=0.00000047, c_1=109.44429, c_2=0.0000120548, J_{min} =2.312654·10^{-8} are obtained. Simulation time is 20s. The value of parameter δ has been chosen to be 1.002837. The parameter c_0 can be neglected because its value is very close to zero.

Now, the state-space model of ABS can be obtained in the form (10), where:

$$\mathbf{A}=\begin{bmatrix}0 & 1\\ -p_1p_2 & -\left(p_1+p_2\right)\end{bmatrix},$$
$$\mathbf{b}=\begin{bmatrix}0\\1\end{bmatrix}, \tag{29}$$
$$\mathbf{c}=\left[c_1+c_2 \quad K\left(c_1p_2-c_2p_1\delta\right)\right],$$
and $\mathbf{d}=[0]$.

Using (12) the discrete-time ABS model can be derived in the form (11) with:

$$\boldsymbol{\phi}=\begin{bmatrix}-0.7e^{-64T}+1.1e^{-4.2T} & -0.01e^{-64T}+0.02e^{-4.2T}\\ -268.331\left(-0.01e^{-64T}+0.02e^{-4.2T}\right) & 1.07e^{-64T}-0.07e^{-4.2T}\end{bmatrix},$$
$$\boldsymbol{\gamma}=\begin{bmatrix}0.003+0.0003e^{-64T}-0.004e^{-4.2T}\\ 3.46\times10^{-18}-0.02e^{-64T}+0.02e^{-4.2T}\end{bmatrix}, \tag{30}$$

and, finally, the ABS model in the form of discrete-time transfer function is:

$$\lambda_k=\frac{z^{-1}B\left(z^{-1}\right)}{A\left(z^{-1}\right)}M_{1k}, \tag{31}$$

where the coefficients of the polynomials $A\left(z^{-1}\right)$ and $B\left(z^{-1}\right)$ are $a_0=1$, $a_1=-1.4861$, $a_2=0.5055$, $b_0=0.008$ and $b_1=-0.0015$ for T=10 ms.

In the case of ABS, the digital sliding mode based MV control (19) is given by:

$$M_{1k}=-\frac{1}{E\left(z^{-1}\right)B\left(z^{-1}\right)}\left(F\left(z^{-1}\right)\lambda_k-M\left(z^{-1}\right)\lambda_{k+1}^{\ ref}+\frac{\alpha T}{1-z^{-1}}\operatorname{sgn}\left(s_k\right)\right), \tag{32}$$

The proposed control algorithm should provide zero value of the switching function:

$$s_k=M\left(z^{-1}\right)\left(\lambda_k-\lambda_k^{\ ref}\right), \tag{33}$$

with $M\left(z^{-1}\right)=m_o+m_1z^{-1}+m_2z^{-2}$ where $m_0=1$, $m_1=-1.0670$, $m_2=0.2846$. The coefficients of the polynomial $M\left(z^{-1}\right)$ are calculated by using $M\left(z^{-1}\right)=\left(z-\exp(-2\pi f_cT)\right)^2$ where $f_c=10\text{Hz}$ is a cut-off frequency. According to (20), the polynomials $E\left(z^{-1}\right)$ and $F\left(z^{-1}\right)$ are defined in this case as:

$$E\left(z^{-1}\right)=e_0=\frac{m_0}{a_0}, \tag{34}$$

$$
\begin{aligned}
&F\left(z^{-1}\right)=f_0+f_1z^{-1},\\
&f_0=m_1-e_0a_1,\\
&f_1=m_2-e_0a_2.
\end{aligned}
\tag{35}
$$

In order to compare the results obtained by using the proposed control approach based on the ABS orthogonal model, the same control law (32)-(35) has been implemented by using the discrete-time representation of (27) in the form of (31).

5 Experimental Results

To verify the effectiveness of the proposed type of filters and control algorithm designed on the basis of sliding mode and MV, two experiments have been done on ABS experimental setup [33]. First, one has been performed by using the digital sliding mode based MV control (32) designed on the basis of the discrete-time representation (13) of ABS model (27) and its results are shown in Figure 4. The second experiment has been realized by using the proposed control law (32) developed by using the discrete-time representation (31) of ABS orthogonal model (29) described with GQOFMLT, and the obtained results are presented in Figure 5. In this way, by implementing the same control law and different referent models in their design, it can be concluded whose model better describes the ABS dynamics. In both cases, the run time is 2.7 s, the sampling period T=10 ms, the reference wheel slip $\lambda_{k+1}^{ref}=0.2$ [45], and the controller parameter α is chosen to be 1.

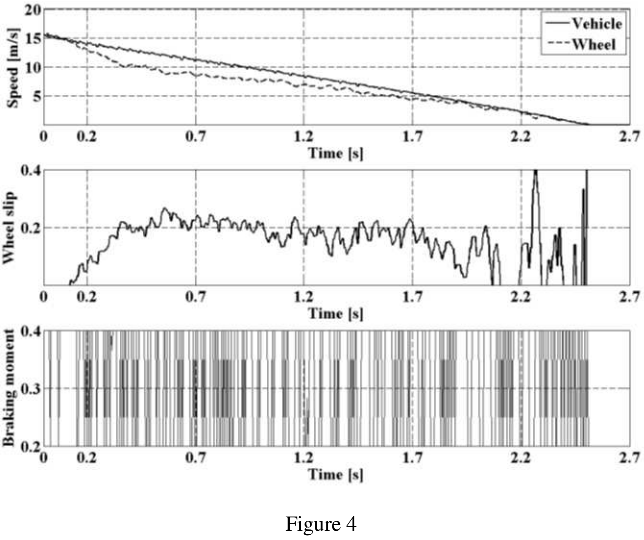


Figure 4

ABS responses with the control law (32) derived from the ABS model (27)

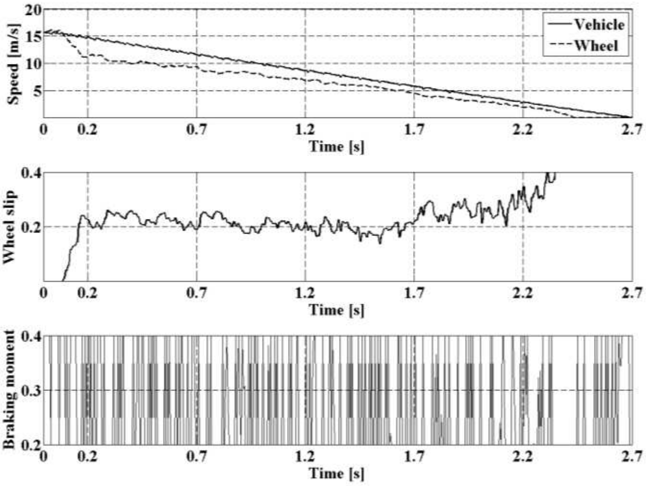


Figure 5

ABS responses with the control law (32) derived from ABS orthogonal model

In both Figures, there are three subplots representing the vehicle and wheel speeds, wheel slip and braking moment, respectively. These three system responses are enough for the valuation of control algorithm. As it can be seen from Figures, the proposed control law suppresses the chattering phenomenon due to the presence of a discrete-time filter in the front of the relay component. In this way, the control law does not lead to excitation of non-modeled high-frequency system dynamics and does not cause deterioration of the mechanical parts of the system.

To make the analysis of the obtained results much easier to explain, the comprehensive index I is introduced as:

$$I = 2k_1 T_{\text{stop}} + k_2 N + k_3 E\,, \tag{36}$$

where k_1, k_2, k_3 represent the real constant coefficients, T_{stop} is a stopping time, N is a total number of changes in the control law, and E is an error calculated as $E = \frac{1}{N_s}\sum_{i=0}^{N_s}\left(\lambda_k - \lambda_k^{ref}\right)^2$, where N_s is a number of samples. It is obvious that all of these values should be as small as possible. The obtained results for the coefficients $k_1 = 0.1984$, $k_2 = 0.0038$, $k_3 = 0.0421$ are presented in Table 1. These values of coefficients provide the normalization of the considered parameters.

Table 1

Analysis of the obtained experimental results

Model reference	T_{stop}	N	E	I
Discrete-time model	2.59	147	14.56	2.20
Orthogonal model	2.45	113	9.19	1.79

As it can be seen, the stopping time was slightly shortened when the orthogonal model is used in the controller design. From the tracking point of view, better experimental results are obtained using the orthogonal model, due to a smaller deviation between the current and referent values of wheel slip. In this way, the maximum value for the friction force is constantly ensured leading to the better steering characteristics of the vehicle. A more faithful representation of the plant dynamics comes from the introduced parameter *δ,* which successfully describes the imperfections in the form of the noise presence, parameter variations, and additional limitation of the control signal to [0.2, 0.4] (to avoid the saturation).

Conclusions

In this paper, a new sliding mode based minimum variance control algorithm designed on the basis of a model obtained by using a new type of orthogonal filters is presented. After giving the necessary background of the used theory, the proposed control law is applied in the anti-lock braking system control. Experimental results have shown that the stopping time, the number in changes in the control law and the tracking error are lesser than in the case where the proposed control is designed on the basis of previously developed ABS model. Having that in mind, it can be concluded that the orthogonal model represents much better system dynamics due to the introduced parameter *δ,* which characterizes all the imperfections in the system. From the control point of view, both experiments confirm the effectiveness of the proposed robust control method.

Acknowledgement

This paper was realized as a part of projects TR 35005, III 43007 and III 44006, financed by the Ministry of Education and Science of the Republic of Serbia for the period 2011-2019.

References

[1] Jingang, G., Xiaoping, J. and Guangyu, L.: Performance evaluation of an anti-lock braking system for electric vehicles with a fuzzy sliding mode controller, Energies, 2014, Vol. 7, No. 10, pp. 6459-6476

[2] Shuwen, Z., Siqi, Z. and Qingming, C.: Vehicle ABS equipped with an EMB system based on the slip ratio control, Transactions of FAMENA, 2019, Vol. 43, No. SI-1, pp. 1-12

[3] Jun-Cheng, W. and Ren, H.: Hydraulic anti-lock braking control strategy of a vehicle based on a modified optimal sliding mode control method, Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2018, Vol. 233, No. 12, pp. 3185-3198

[4] Oniz, Y., Kayacan, E. and Kaynak, O.: Simulated and experimental study of antilock braking system using grey sliding mode control, Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Montreal, Que., Canada, 7-10 October 2007, pp. 90-95

[5] Kayacan, E., Oniz, Y. and Kaynak, O.: A grey system modeling approach for sliding-mode control of antilock braking systems, IEEE Transactions on Industrial Electronics, 2009, Vol. 56, No. 8, pp. 3244-3252

[6] Chun, K. and Sunwoo, M.: Wheel slip control with moving sliding surface for traction control system, International Journal of Automotive Technology, 2004, Vol. 5, No. 2, pp. 123-133

[7] Jing, Y., Mao, Y., Dimirovski, G. M., Zheng, Y. and Zhang, S.: Adaptive global sliding mode control strategy for the vehicle antilock braking systems, Proceedings of American Control Conference, St. Louis, MO, USA, 10-12 June 2009, pp. 769-773

[8] Harifi, A., Aghagolzadeh, A., Alizadeh, G. and Sadeghi, M.: Designing a sliding mode controller for slip control of antilock brake systems, Transportation Research Part C: Emerging Technologies, 2008, Vol. 16, No. 6, pp. 731-741

[9] Song, J.: Performance evaluation of a hybrid electric brake system with a sliding mode controller, Mechatronics, 2005, Vol. 15, No. 3, pp. 339-358

[10] Park, E. J., Stoikov, D., Falcao da Luz, L. and Suleman, A.: A performance evaluation of an automotive magnetorheological brake design with a sliding mode controller, Mechatronics, 2006, Vol. 16, No. 7, pp. 405-416

[11] Zheng, S., Tang, H., Han, Z. and Zhang, Y.: Controller design for vehicle stability enhancement, Control Engineering Practice, 2006, Vol. 14, No. 12, pp. 1413-1421

[12] Norhazimi, H., Khairi, A., Yahaya, S., Hazlina, S. and Rozaimi, G.: Second order sliding mode controller for longitudinal wheel slip control, Proceedings of IEEE 8th International Colloquium on Signal Processing and its Applications, Melaka, Malaysia, 23-25 March 2012

[13] Wu, M. and Shih, M.: Using the sliding-mode pwm method in an anti-lock braking system, Asian Journal of Control, 2001, Vol. 3, No. 3, pp. 2553-261

[14] Wu, M. and Shih, M.: Simulated and experimental study of hydraulic anti-lock braking system using sliding-mode PWM control, Mechatronics, 2003, Vol. 13, No. 4, pp. 331-351

[15] El Hadri, A., Cadiou, J. C. and M'sirdi, N. K.: Adaptive sliding mode control of vehicle traction, Proceedings of 15th Triennial World Congress, Barcelona, Spain, 21-26 July, 2002, Vol. 35, No. 1, pp. 391-396

[16] Nikolić, S., Antić, D., Danković, B., Milojković, M., Jovanović, Z. and Perić, S.: Orthogonal functions applied in antenna positioning, Advances in Electrical and Computer Engineering, 2010, Vol. 10, No. 4, pp. 35-42

[17] Danković, B., Nikolić, S., Milojković, M. and Jovanović, Z.: A class of quasi-orthogonal filters, Journal of Circuits, Systems, and Computers, 2009, Vol. 18, No. 5, pp. 923-931

[18] Milojković, M., Nikolić, S., Danković, B., Antić, D. and Jovanović, Z.: Modeling of dynamical systems based on quasi-orthogonal polynomials, Mathematical and Computer Modeling of Dynamical Systems, 2010, Vol. 16, No. 2, pp. 133-144

[19] Antić, D., Danković, B., Nikolić, S., Milojković, M. and Jovanović, Z.: Approximation based on orthogonal and almost-orthogonal functions, Journal of the Franklin Institute, 2012, Vol. 349, No. 1, pp. 323-336

[20] Milojković, M., Antić, D., Nikolić, S., Jovanović, Z. and Perić, S.: On a new class of quasi-orthogonal filters, International Journal of Electronics, 2013, Vol. 100, No. 10, pp. 1361-1372

[21] Nikolić, S., Antić, D., Perić, S., Danković, N. and Milojković, M.: Design of generalised orthogonal filters: application to the modelling of dynamical systems, International Journal of Electronics, 2016, Vol. 103, No. 2, pp. 269-280

[22] Nikolić, S., Antić, D., Milojković, M., Milovanović, M., Perić, S. and Mitić, D.: Application of neural networks with orthogonal activation functions in control of dynamical systems, International Journal of Electronics, 2016, Vol. 103, No. 4, pp. 667-685

[23] Spasić, M., Mitić, D., Hovd, M. and Antić, D.: Predictive sliding mode control based on Laguerre functions, Journal of Control Engineering and Applied Informatics, 2019, Vol. 21, No. 1, pp. 12-20

[24] Perić, S., Antić, D., Milovanović, M., Mitić, D., Milojković, M. and Nikolić, S.: Quasi-sliding mode control with orthogonal endocrine neural network-based estimator applied in anti-lock braking system, IEEE/ASME Transactions on Mechatronics, 2016, Vol. 21, No. 2, pp. 754-764

[25] Antić, D., Nikolić, S., Milojković, M., Danković, N., Jovanović, Z. and Perić, S.: Sensitivity analysis of imperfect systems using almost orthogonal filters, Acta Polytechnica Hungarica, 2011, Vol. 8, No. 6, pp. 79-94

[26] Milojković, M., Antić, D., Milovanović, M., Nikolić, S., Perić, S. and Almawlawe, M.: Modeling of dynamic systems using orthogonal endocrine

adaptive neuro-fuzzy inference systems, Journal of Dynamic Systems Measurement and Control, 2015, Vol. 137, No. 9, pp. DS-15-1098

[27] Perić, S., Antić, D., Nikolić, V., Mitić, D., Milojković, M. and Nikolić, S.: A new approach to the sliding mode control design: anti-lock braking system as a case study, Journal of Electrical Engineering, 2014, Vol. 65, No. 1, pp. 37-43

[28] Amitava, C., Ranajit, C., Fumitoshi, M. and Takahiro, E.: Augmented stable fuzzy control for flexible robotic arm using LMI approach and neuro-fuzzy state space modeling, IEEE Transactions on Industrial Electronics, 2008, Vol. 55, No. 3, pp. 1256-1270

[29] Haidegger, T., Kovacs, L., Preitl, S., Precup, R. E., Benyo, B. and Benyo, Z.: Controller design solutions for long distance telesurgical applications, International Journal of Artificial Intelligence, 2011, Vol. 6, No. S11, pp. 48-71

[30] Németh, B. and Péter, G.: LPV design for the control of heterogeneous traffic flow with autonomous vehicles, Acta Polytechnica Hungarica, 2019, Vol. 16, No. 7, pp. 233-246

[31] Mitić, D. and Milosavljević, Č.: Sliding mode-based minimum variance and generalized minimum variance controls with $O(T^2\ )$ and $O(T^3)$ accuracy, Electrical Engineering, 2004, Vol. 86, No. 4, pp. 229-237

[32] Thangavelusamy, D. and Ponnusamy, L.: Elimination of chattering using fuzzy sliding mode controller for drum boiler turbine system, Journal of Control Engineering and Applied Informatics, 2013, Vol. 15, No. 2, pp. 78-85

[33] Inteco, The laboratory anti-lock braking system controlled from PC-User's Manual, 2008, available at www.inteco.com.pl

[34] Topalov, A., Oniz, Y., Kayacan, E. and Kaynak, O.: Neuro-fuzzy control of antilock braking system using sliding mode incremental learning algorithm, Neurocomputing, 2011, Vol. 74, No. 11, pp. 1883-1893

[35] Wei, Z. and Xuexun, G.: An ABS control strategy for commercial vehicle, IEEE/ASME Transactions on Mechatronics, 2015, Vol. 20, No. 1, pp. 384-392

[36] Lin, C. M. and Hsu, C. F.: Self-learning fuzzy sliding-mode control for antilock braking systems, IEEE Transactions on Control Systems Technology, 2003, Vol. 11, No. 2, pp. 273-278

[37] Mitić, D., Milosavljević, Č. and Veselić, B.: One approach to I/O based design of digital sliding mode control for nonlinear plants, Electronics, 2004, Vol. 8, No. 2, pp. 64-67

[38] Mitić, D.: Digital variable structure systems based on input-output model, University of Niš, Faculty of Electronic Engineering in Niš, 2006

[39] Mirzaeinejad, H. and Mirzaei, M.: A novel method for non-linear control of wheel slip in anti-lock braking systems, Control Engineering Practice, 2010, Vol. 18, No. 8, pp. 918-926

[40] Martinez-Gardea, M., Mares Guzman, I., Acosta Lua, C., Di Gennaro, S. and Vazquez Alvarez, I.: Design of a nonlinear observer for a laboratory antilock braking system, Journal of Control Engineering and Applied Informatics, 2015, Vol. 17, No. 3, pp. 105-112

[41] Stan, M., Precup, R. E. and Paul, A. S.: Analysis of fuzzy control solutions for anti-lock braking systems, Journal of Control Engineering and Applied Informatics, 2007, Vol. 9, No. 2, pp. 11-22

[42] Mitić, D., Perić, S., Antić, D., Jovanović, Z., Milojković, M. and Nikolić, S.: Digital sliding mode control of anti-lock braking system, Advances in Electrical and Computer Engineering, 2013, Vol. 13, No. 1, pp. 33-40

[43] Precup, R.-E., Preitl, S., Rădac, B. M., Petriu, E. M., Dragoş, C. A. and Tar, J. K.: Experiment-based teaching in advanced control engineering, IEEE Transactions on Education, 2011, Vol. 54, No. 3, pp. 345-355

[44] Utkin, V. I.: Sliding modes in optimization and control, Springer-Verlag, New York, 1992

[45] Zanten, A., Erhardt, R. and Lutz, A. Measurement and simulation of transients in longitudinal and lateral tire forces, SAE Technical Paper 900210, 1990

Student Employment as a Possible Factor of Dropout

Zsófia Kocsis and Gabriella Pusztai

University of Debrecen, Faculty of Arts Doctoral School of Human Sciences,
Egyetem tér 1, 4032 Debrecen, Hungary
kocsis.zsofia@arts.unideb.hu, pusztai.gabriella@arts.unideb.hu

Abstract: One of the possible reasons for student dropout, is the attraction of the labor market. Nowadays, the date of employment does not coincide with the date of graduation, sometimes the income of those without a degree are higher than those with a degree. In addition, it may also lead to the interruption of university studies that the students judge negatively the marketability of their studies, in which case, the appeal of the labor market is even more prevalent. During our research, we tried to identify the process of dropout using quantitative and qualitative methods. As a first step, we interviewed dropped out and 'at risk' students, but in our current analysis, we only processed those interviews where student work played a significant role in the life of the interviewee and this affected their dropouts. In addition, during our quantitative research, we were looking for individuals who had left their higher education studies without graduation in the last 10 years, and finally we worked with a database of 605 people. Both our qualitative and quantitative results show that financial reasons dominate during student employment that make them fall into a vicious circle. The results draw attention to the fact that working during, or instead of the university is an inevitable point of analysis for dropping out.

Keywords: dropout; student employment; financial burdens; student work

1 Theoretical Overview

Student employment is a double-edged sword, as on the one hand, it can reduce the academic performance of the students and on the other hand, it can have positive effects on the long run. In addition of earning money, the main advantage of working during studies is gaining work experience, enriching the CV with that work experience and gaining a better understanding of the structure of the labor market, which may be an advantage for students after graduation [1, 2, 3]. It also has a positive impact on oral communication skills, teamwork, improves time management capabilities and students can expand their social connections through work [4, 5]. Nowadays, employment is playing an increasingly important role in the lives of the students. They devote a lot of time and energy to integrate work into their daily routine. According to the latest, Eurostudent VI, a little over half of the students worked during their university studies, and some of the students interrupted their studies for work-related reasons, in their case the attraction of the labor market appears strongly [6]. Several factors have already been identified in the 2012 material of OECD which enhance the chance of student dropouts. Among the top reasons was the attractiveness of the labor market, due to which the students are willing to leave the higher education institution without qualifications to ensure and improve their financial situation. There is no single definition to determine dropout of students. According to Lukács & Sebő [7], there are three possible ways for student output: graduated, exited and dropped out. In the case of drop-out students, we mean cases where the student was excluded from the training due to passive semesters, non-acceptance of the cost reimbursement, the student was expelled due to disciplinary procedures or the student dropped their major or did not enroll. According to Fenyves et al. [8], the forms of dropout can be distinguished: the students themselves request dismissal, the institution asks for dismissal (due to study reasons, exam conditions), or the dismissal has health or financial reasons (study costs, the payment of the tuition fee was not realized on time). However, the studies focus on students who have finished their higher education without obtaining a degree and have thus become dropout students [7, 8]. Student dropout rates in Hungary are high, 36-38% in undergraduate trainings and 14-17% in graduate trainings [9]. There is a suspicion that intensive student employment contributes to poorer academic achievements, postponement of the fulfillment of requirements, and eventually even leads to dropout. As a result, student employment can be interpreted as a risk factor that increases the risk of dropping out of school by keeping students away from university culture and embedding into the community [1, 10, 11, 13]. According to McCoy and Smyth [13] the drop-out rate is higher for students who take permanent, long-term jobs. According to Eurostudent VI survey data, 39% of students in Hungary regularly work during the semester, while 14% work periodically during the study period [6]. Earlier research has confirmed that the attractiveness of the labor market is much stronger for students who do not receive any financial support, they are more likely to work during their university studies with higher intensity [2]. The results of the data collection of Eurostudent VI. show that the primary motivation for Hungarian students to work is to cover their living expenses. Most of them then said that without paid work, they would not be able to afford to be university students and the same amount of them work for experience. The fewest number of students work because they need to support someone else financially. In Hungary, working students spend more than 30 hours a week with work. Previous research has also confirmed that the tendency to drop out is higher among students who have long-term, intensive employment [6, 13]. 7% of Hungarian University Students interrupted their studies for at least a year and during the process, a quarter of students referred to workplace related reasons [6]. Work carried out during studies, especially during the passive semester, has a negative impact on graduating. Mostly, master students interrupted their studies and those who work more intensively than other student groups. A group of students (25%) who interrupted their studies indicated that they requested to be passive due to workplace reasons. We can see that working during the university studies is an inevitable point of analysis in terms of dropout. In the process of dropout, it is worth mentioning the combination of factors that may increase the negative impact of employment, such as the family and financial background of the students, the motivation to work and their relationship to their studies. Previous research [3] also showed that students who have low, unfavorable status indicators and work because of their financial difficulties, and students who find their studies non-marketable are more vulnerable to dropping out. Working not only reduced the time spent on studying, but also the students' commitment to study, as the attractiveness of the labor market was even more pronounced among dropout students, they believed that their workplace provided a greater financial security and work was a 'promise' of their future employment, so they have chosen to interrupt their studies instead of completing them [3]. In our research, we focused on the role of employment in the process of dropout and we have examined how often and for what reason the dropout students worked during their studies.

2 The Method of Research

The Center for Higher Education Research and Development (CHERD) of the University of Debrecen in the framework of 123847 NKIFH (National Research, Development and Innovation Office) research project titled Social and institutional factors of student dropout in higher education conducts research focusing on factors affecting dropout since September 2017. During the course of our research, we tried to identify the process of dropout using quantitative and qualitative methods. As a first step, we interviewed dropout and 'at risk' students. Students were at risk for higher than the average number of passive semesters and procrastination, moreover, there were unfavorable changes in terms of the form of study and financing. In the qualitative phase of the research, we tried to get a clearer picture of the details of student employment and the reasons behind dropouts. At this stage of the study, we conducted a structured personal interview with several students, the staff of the Center for Higher Education Research and Development (CHERD) of the University of Debrecen was also involved in the research. We did not seek representativeness in the selection of interviewees, so we cannot draw clear conclusions about which social groups are more at risk of dropout due to work. We focus on the relationship between employment and dropout. We highlighted 7 interviews in which the students interrupted their studies due to the attractive impact of the labor market. With the help the method of Nagy [14] we recorded the interviews and we split the texts into units. We used an open code system and marked the topic of the units of the text. We analyzed the interviews on the basis of the characteristics of employment. With the help of the qualitative methods, we were able to get to know the target group's way of thinking and their motivation to work. Our goal is to continue to explore the underlying meanings and patterns of the connections. The lives and university careers of the interviewees are different. In addition to socio-cultural, financial and personal reasons, regular work played a significant role in the dropout of the surveyed students. We present to the university career of seven students, as they were those interviewees whose work was continuously present in their lives, in different ways. Basic information about the interviewees can be found in the list below, a numbering was used to ensure anonymity.

Student 1: 27-year-old male, completing his thesis and a successful final exam is needed to get his degree. He is currently spending his 14th semester in higher education.

Student 2: 26-year-old male, who started their studies in 2011, first studying computer science, then chemical engineering, his expected graduation date was December 2017.

Student 3: 24-year-old female law student, who became a correspondent student for financial reasons and accepted a legal assistant job in Siófok.

Student 4: 20-year-old female studying pedagogy, who is in multiple-semester procrastination due to student work and the difficulties in studies.

Student 5: 23-year-old male student, currently studying chemical engineering, had multiple passive semesters.

Student 6: 37-year-old male, graduated from the same training program as a correspondent student after multiple years, as they realized that obtaining the degree was essential to change jobs.

Student 7: 37-year-old male, from a small town near the county seat, currently living with his wife and his child in the county seat.

Based on the results of the qualitative research, our research team has formulated several statements in the questionnaire that may be the causes of dropouts identified by students. As a next step in the research, we contacted students who discontinued their higher education studies without graduating in the past 10 years. In the case of discontinued studies, we tried to cover a wide range of courses from natural sciences, engineering, and pedagogy to arts, medicine, healthcare, military, art, etc. We contacted students from 32 Hungarian higher education institutions (mainly from the Northern Great Plain) and five cross-border higher education institutions using the snowball method, a total of 605 people filled out the questionnaire. In our research, we used the DEPART 2018 database. Our research team formulated several statements in the questionnaire, which may be the causes of dropout identified by the students, which are listed in the appendix. These statements (listed in the appendix) had to be evaluated on a scale of 1 to 4, and the following four clusters were created based on the statements in the quantitative phase:

1) Dropouts due to financial reasons and employment
2) Dropouts due to educational and institutional reasons
3) Multiple causes identified
4) Disappointed in their major and further education [15]

We used descriptive and bivariate analysis. In addition to the characteristics of the work, social backgrounds and institutional (funding of training) variables were used. In addition, motivation appeared as an individual variable.

3 Results of the Qualitative Research

3.1 How has Work Played a Role in the Dropout Process?

In analyzing the interviews, the main guideline was working during the university. Almost all of the interviewees took up working during their university studies for financial reasons. Some of them were motivated by the extra income or the desire to be independent from their parents, while some students were saving for their tuition fee. There was only one person who worked and discontinued their university studies to gain experience. We present the dropout process of the interviewees based on their attitude towards work.

In the first round, we interviewed those students who are typical examples of working students. Previous research has also shown that most students working during university seek employment due to financial difficulties. The first three students started working during their university years due to financial constraints. This was due to financial constraints, lack of parental support or the obligation to pay tuition fees.

> ’As a student worker, I worked several times in factories, shops, horticulture and construction sites. For a time, there was social support, I had a scholarship once, I had one better semester…’ (Interviewee 1)
>
> ’At first, I did not work, later I needed to do so and so to speak, this continues to this day. It was annoying at first that almost everyone around me came with much more pocket money than I did. I decided to take a job. There were jobs where you had to go 2-3 times a month, obviously it was not a big help financially.’ (Interviewee 2)

In their view, work had a major negative impact on their studies. All three students were transferred to fee-paying trainings, 'perhaps this was the first domino that fell down' (Interviewee 2). A vicious circle has started, they had to work even more. The combination of work and studies was not easy, and as financial constraints were stronger, studying was clearly pushed into the background and they could not conciliate work with studying.

> 'I went to work whenever I could, and yes there were examples that this happened at the expense of studying. During the last semesters, I barely had any subjects and I tried to work as much as possible and study possibly less...' (Interviewee 1)

> 'However, when you went almost every day, at nights, you did not really have time to live. Getting up and going to class after a night shift was a problem many times, and we have not even talked about the time spent on studying, which was almost nonexistent then.' (Interviewee 2)

Work was constantly present in the life of the third interviewee as well, she regularly took typical student jobs during her university years to be able to pay for daily expenses and tuition fees. In her case, the attractive effect of the labor market has clearly prevailed, she decided to leave her full-time training and became a correspondent student, so she had the opportunity to work even more. The university student from Debrecen moved to Siófok for personal reasons, where first she worked as a receptionist, then as a legal assistant.

> 'I am currently working in Siófok at a notary's office, but I have been continuously working during the university years. Now I am a 4th year law student. For three years, I was a full-time student and took student jobs. I became a correspondent student last September, as I wanted to work more to support myself and pay my tuition. During the summer, I was a receptionist at Lake Balaton, but after the season, I started working here at the office…' (Interviewee 3)

Work makes it easier for her to save money for tuition. What she sees positive in her decision is that after her several student jobs, she is finally doing work that is related to her studies. Although work, being a correspondent student and travelling a lot makes her life difficult, she still feels that she made a good decision.

> 'I found it important to gain new experiences, my student jobs were completely different… Hostessing, bartending, call centers, I have worked for almost all student job agencies. School took away a lot of time, and now I get to make my own schedule, 40 hours a week must be fulfilled, sometimes it is difficult to bring it together, but not impossible.' (Interviewee 3)

The reason for the dropout of the first three students was clearly working during university. From the beginning of their studies, they did some kind of gainful activity, which they could hardly make work with their university studies over time. As a result, they have become fee-paying students, further increasing their financial burdens, resulting in more intensive work. In spite of the work and the changes taking place, they continue their studies: 'If I have started, I will finish it!' (Interviewee 2).

The following two interviewees belong to the group of young people who are motivated to work by the desire to be independent from their parents. Even though they were not financially reliant on the extra income, they wanted to work, but their employment greatly compromised their university career. For the interviewees, working became a part of their lives during their university years. Initially, time management was difficult for them, to be able to study, work and rest.

> 'My week is almost spent with working. Usually 5 days a week, but they adapt to my schedule, so I let them know when I am free and they schedule me to work, so I can attend my classes. There are days when I work four hours, but generally nine. […] Then I will tell you how my day went today, because, there was everything. I had a class from 8am until 10am, then I went to work at 11am, I was there until 3pm, then I had a class from 4pm, which is until 6pm. Then I have workout from 7pm until 9pm. But actually I work more days, and I have a rest day when I try to catch up with myself.' (Interviewee 4)

In addition to working, the interviewees name the problems arising from their chosen major as the other reason for dropout. On the one hand, the problems were due to the negative perception of the major and the lack of supporting institutional environment. These factors increase the attractiveness of the labor market, which is why the interviewees chose their work instead of the university.

> 'I feel that I cannot do anything with this degree. And because of this, I am negative. Now I slip a year and I am here for four more years… and actually for nothing. However, my work is not related to my university studies, but it is positive that I have learned to work in a team, how to manage my time and the financial independence. I was able to hold my ground at least financially if I could not do the same at the university.' (Interviewee 4)

My fifth interviewee initially started to work to become independent from the parents. The reason for dropping out of the first major was not due to their employment, but rather institutional reasons. However, this reason was enough to consider student jobs more useful than continuous failing, even though studying and graduation were always among their goals.

> 'My high school class teacher, who was also my chemistry and math teacher pulled me down so much that I did not want to learn chemistry anymore. So I applied for the university to study computer science, I was accepted and I studied here for two years. Then there was my first break. Then I had a passive semester and went to work as a student, then there was a semester when I tried to informatics, but I did not like it, that subject and that

> teacher… and I quit permanently. Then I became an individual entrepreneur. I was a financial advisor, but I realized it is not something I really want to do, so I came back to study chemical engineering…' (Interviewee 5)

The interviewed student, as a result of negative work experience, realized that obtaining a degree is still an advantage for them in the labor market, so they applied for another training. But soon after beginning their chemical engineering studies, the attractive effect of the labor market prevailed again for the student. They decided to apply to work for TEVA to earn income and gain professional experience. According to them, they felt that they could not lose anything with this work, but could only gain additional knowledge, something they had no opportunity for at the university.

> 'The second time I became passive, within chemical engineering, it was a more conscious decision to see if it would be better to work... and then I went to TEVA for almost a full year, it was connected to my major. I believe you can learn something from every job. That is why they call back from many places. I also had a lot of benefit from TEVA, as I got to know people and I know what it takes to get in there. I also met the boss. To 60-40% I needed the work to make money, that was stronger. And I also got a little insight where I should get in the future.' (Interviewee 5)

Among the interviewees, there was one who was less far-sighted, and only realized with the passage of time that having a degree is essential in today's world. They have been working in student jobs since high school, most of which were physical work. Similarly, to other interviewees, they called it positive that they expanded their network of contacts through work and made acquaintances which could be beneficial. However, their employment resulted in dropout. They have been working on weekends for months during their studies, which significantly reduced the time spent with studying. Moreover, they thought they do not need a degree:

> 'There will be no big difference in the salary of a new graduate and that of someone with vocational education in the future.' (Interviewee 6)

The warning signs have already appeared during the first exam period; this was when they decided to discontinue their university studies. They were able to find a company where they were chosen on the basis of their expertise, not the documents they obtained. There were no differences in salary, which further strengthened the belief that they do not need to complete their higher education. After several years, they were convinced that at least one degree is required to change jobs and advance in their careers.

The last interviewee returned to the labor market time and time again. He dropped out of two faculties and two majors, and work was constantly present in his life. He typically spent a period on the labor market after pausing his studies or before starting a new major. He said he chose to work, because the workplace was still more secure than completing certain subjects and then finding a job with a degree.

> ’There were certain subjects that were simply impossible to complete. Among the students, there were some who were in the three-year training for 7-8 years already. I failed the exam so many times for one of my subjects that I had to pause the training for two semesters. After that, I started another training, as a geographer, but I did not see the labor market outcome of the training.’ (Interviewee 7)

The third major, which he finished as a correspondent student, was the andragogy training, which at the time was connected to his job. He did not find completing the training burdensome, but by the time he finished, he also changed jobs. At his current job, he is writing project tenders. In his free time, he is currently attending a vocational training to become a carpenter. His whole life and his relationship to work was accompanied by a positive perception of manual work, he is studying the carpentry profession because he says that they are doing ’nothing’ at the office but producing papers.

4 Results of the Quantitative Research

In our research, we focused on the role of employment in the dropout process, examining how often dropout students did paid work during their studies. We defined the work frequency values as follows: never, yearly, monthly and weekly work. 28% of the respondents performed paid work weekly during their studies, 11% monthly, 12% yearly, but the highest proportion (48%) were those students who did not work at all during their years of higher education. Previous studies [6, 16] had similarly low numbers of students whose work was related to their studies, 23% of them said that they had this kind of job.

Motivation has a particularly important role to play in why students start working. Usually the job of those students were related to their studies, who started working specifically to gain experience. However, in addition to gaining experience, financial reasons are dominant. Earlier research shows that the majority of university students work to cover their living expenses and to a lesser extent, finance their entertainment [3, 6]. According to Masevičiūtė et al. [6] in the Balkan countries, the main reason for students to work is to pay tuition fees, finance living expenses, gain experience and support others. The research of Fónai [17] also confirmed that there are three times as many dropouts in fee-paying trainings as in state-funded trainings. The financial burdens resulting from tuition fees and the employment can often be the causes of dropout. For this reason, we have examined what percent of dropout students were affected by fee-paying trainings and why they began to work during their studies. 33% of the respondents continued their last training as fee-paying students. However, the payment of the tuition fees as a motivational factor affected 24% of the respondents, they were fully motivated by it during employment. 36% of the respondents were motivated by raising the financial resources needed for sustenance, while 35% were motivated by the desire to be independent from their parents.

It is clear from the answers that the majority of the dropout students were motivated to work due to financial reasons. Work for gaining experience was typical for only 16% of the respondents, as well as work for financing leisure activities. Getting new acquaintances and building relationships did not motivate the respondents almost at all.

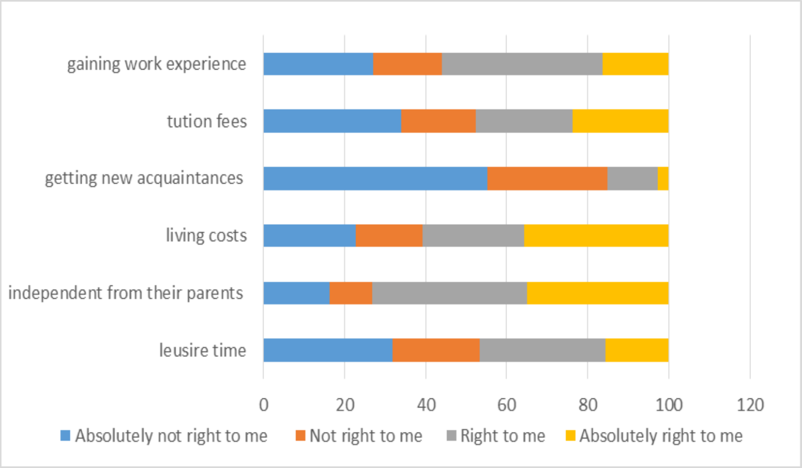


Figure 1

Motivations for employment (N=min. 384). Source: DEPART 2018

The special type of employment is working abroad, which can have positive effects on higher education studies such as learning languages, increased commitment, gaining professional experience, but also negative effects such as the general draining effects of the labor market and the strengthening of the migration bubble. 46 of the respondents (8.6%) worked abroad during their academic years, but there was no significant correlation with dropout clusters. We may assume that the demonstration of the potential positive or negative effects and connections of foreign employment would require a larger sample.

In the remainder of the analysis, we have examined the connections between the above variables and the clusters of dropouts. First, we analyzed the frequency of employment in the clusters developed along the reasons for dropping out, where we found significant connections. 39% of students dropping out due to financial reasons and work were affected by regular employment on a weekly basis, while fifth of them on a monthly basis. The negative effects of employment have also been proven by previous research, based on which the dropout rate is higher among those students who are in permanent, long-term employment, as working prevents them from attending courses to a large extent [10, 18]. A similar rate of employment characterized those respondents who dropped out due to educational reasons, 27.5% of them worked weekly during their university studies, and 28.3% of those who were disappointed with their studies started working. If the students judge the marketability of their degree negatively, the attractiveness of the labor market becomes even more pronounced and the presence of these two factors may lead to dropout. 26.3% of the respondents who were unsure of the reasons and dropped out were also doing some kind of money-making activity on a weekly basis.

Table 1

Frequency of employment in the clusters formed along the reasons for dropping out (p= 0,000) (N=541)

	Dropouts due to financial reasons and employment	**Dropouts due to educational and institutional reasons**	**Multiple causes identified**	**Disappointed in their major and further education**
weekly	39.0%	27.5%	26.3%	28.3%
monthly	21.2%	10.1%	8.6%	11.5%
yearly	6.8%	13.4%	17.1%	12.4%
never	33.1%	49.0%	48.0%	47.9%

Source: DEPART 2018

As a next step in our research, we investigated whether there is a difference between the motives of employment for the clusters formed. Respondents who dropped out due to financial reasons and the cluster of those who were disappointed in the training show some kind of correlation with employment motivations. It is clear that those students who dropped out due to financial reasons and employment were most likely to work due to their financial burdens. 41% of respondents in this cluster needed a job to cover their living expenses, while 32% due to their tuition fees. We assume that these students ended up in kind of a vicious circle, as they were working due to the increasing living expenses and tuition fees, however, the positive effect of employment was only significant financially, as they earned extra income, while working had a negative impact on their academic performance and thus they dropped out.

Presumably, these students were also characterized by what Masevičiūtė et al. [6] observed, namely that the majority of the students could not afford to study in higher education without paid work. In contrast, it was the most typical of those who were disappointed in their major and further education to work for reasons like gaining experience and new acquaintances. 22% of those who were disappointed in further education worked specifically to gain work experience, while 19% to get to know new people during their work.

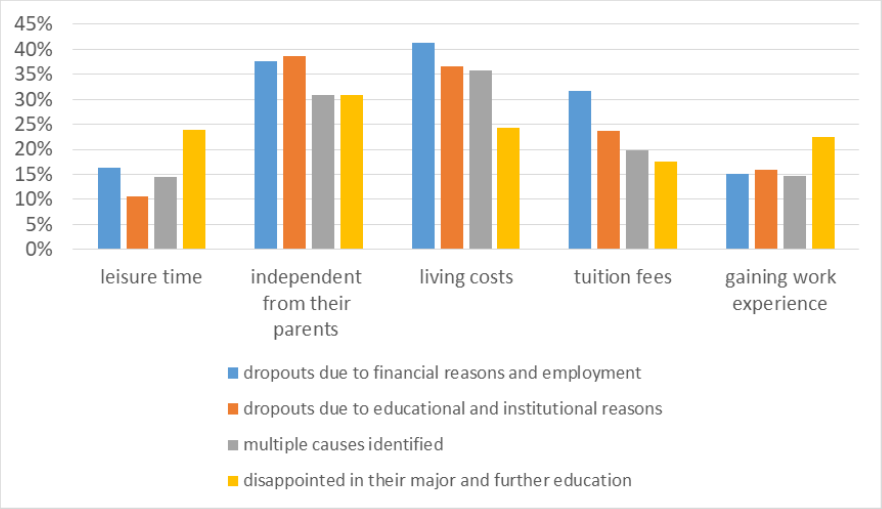


Figure 2

The connection of employment reasons with dropout reasons (p>0,05) (N=605)

Source: DEPART 2018

The results clearly outline that the negative perception of the university studies or the disappointment in the training increases the attractiveness of the labor market, so it is not surprising that the students favored their employment responsibilities instead of their university studies. The data clearly indicates that at least a quarter of the students in each of the clusters worked during their university studies, either weekly or monthly, so to some extent, the negative impact of employment was shown as well as the attractive effect of the labor market. However, it should be noted that the correlation was not significant (Figure 2).

The draining effect of the labor market was not only present during the university years of the respondents but also after the interruption of their studies. 52% of the dropout students said that they had taken a job domestically after completing their studies, the proportion of those were very low who applied for another training, educated themselves or took a job abroad.

Based on our previous interviews and the results of our current quantitative research, we have identified four dimensions that could help reducing dropout rates and the impact of the draining effect of the labor market. In our opinion, the positive effect of student employment could be enhanced by practice-oriented university courses and the promotion of dual training. As work related to the studies can increase student success, but this cannot happen without the active involvement of higher education. Strengthening the relationship between the university and the places of practice, employment agencies and businesses is an essential element in ensuring that students do such work during their studies that are related to their university training.

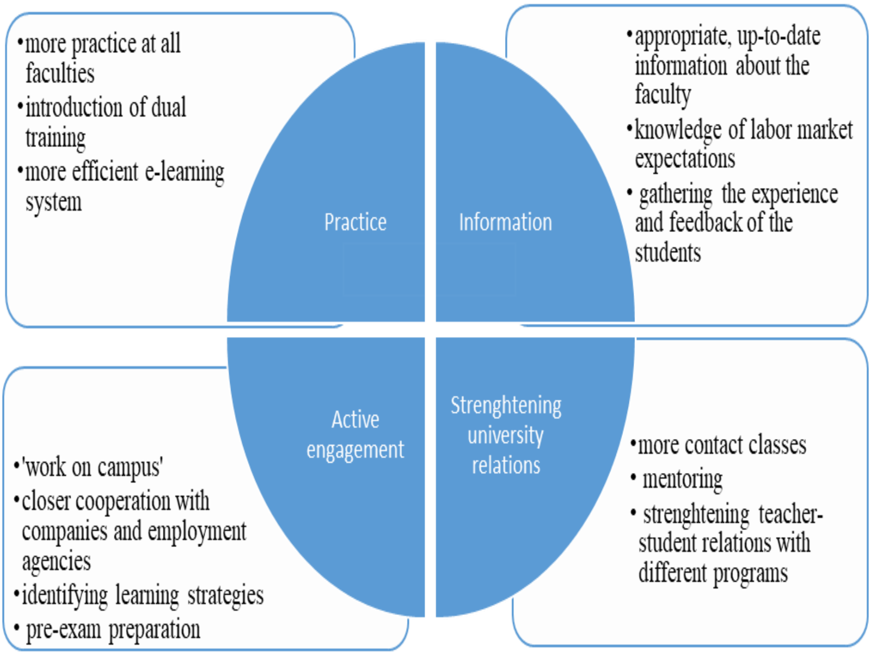

•more practice at all faculties
•introduction of dual training
•more efficient e-learning system
Practice
Information
•appropriate, up-to-date information about the faculty
•knowledge of labor market expectations
• gathering the experience and feedback of the students
Active engagement
Strenghtening university relations
•'work on campus'
•closer cooperation with companies and employment agencies
•identifying learning strategies
• pre-exam preparation
•more contact classes
• mentoring
• strenghtening teacher-student relations with different programs

Figure 3
Suggestions for reducing dropout
Source: [19]

The engagement of the institutions lies not only in liaising with the partner institutions, but from the aspect of dropouts strengthening the relationship between teachers and students is also important, which could be improved through mentoring and other programs. The relationship between the two parties would enable teachers to recognize the learning strategies of the students, or to provide the students with preparatory or additional classes. These options could help working students to 'bring in' their fallback, which is due to their responsibilities at the workplace. In addition, it would be important to take into account the feedback and experiences of the students. Moreover, obtaining and transmitting information can be crucial for students at risk due to employment reasons. On the one hand, applicants should be aware of the outcomes of their training on the labor market, thereby reducing the risk of being disappointed in university education, and possibly less likely to discontinue their studies for a more attractive student job or job opportunity. The knowledge and tracking of labor market expectations can help students gain the knowledge and experience they need during their university years that would support their employment after graduation.

Conclusion

As seen in the results of Eurostudent VI, we realized that that more than half of the students work, due to financial difficulties. It is clear that financial reasons have a significant impact on the frequency of student employment [6]. According to international research results, the clear negative impact of employment is reflected in the deterioration of study achievement, and it may be a possible factor of drop-out. The negative impact of employment on academic performance has been supported by numerous studies [1, 6, 10, 11, 14]. During the analysis of the relationship between employment and dropout, we found that financial burdens, and consequently, work played a significant role in the dropouts of the students. The employment motivation for the majority of the respondents can be connected to finances, independence of their parents, covering daily expenses and the payment of the tuition fees. Motivational factors such as work experience and making new acquaintances motivated students to work to a lesser extent, however, these factors would be the most important for future return. Regular employment affected a quarter of respondents on average, with the highest rate of dropouts due to financial reasons and employment, and those who were disappointed in the major and further education. From the point of view of motivation, the cluster of those who were disappointed in the major and further education was outstanding as they were motivated mostly by work experience and making new acquaintances. We can conclude from the results that the negative perception of university education further enhances the already significant draining effect of the labor market. Therefore, it is not surprising that the attractive effect of the labor market was not only present during the university years, but also after the discontinuation of the studies; 52% of dropout students said that they took a job domestically after finishing their studies and the proportion of people who applied for another training or started to work abroad was very low. 8.6% of the respondents worked abroad during their academic years, but the rate of employment abroad was not significant even after the interruption of their studies. 16% of those who were disappointed in their major and further education decided to work abroad, while to rest of the clusters have a lower rate of the same. Bocsi et al. [3] found that students' unfavorable socio-cultural background exacerbates the negative effects of employment. Furthermore, endangered those working students who were forced to work due to their financial difficulties. The interviewees' reports showed that employment plays and important role in dropout.

Interviewees identified employment as one of the reasons for their dropout. Some of them had no choice due to financial constraints, however more of them deliberately chose to work. They believe, as a career starter, experience is a prerequisite for a successful career, but it matters whether the fresh graduate has real professional experience or not. Some students believe it would be worthwhile to pay more attention to application and scholarship opportunities, as this would help their situation both professionally and financially, making student work less desirable. In addition, there would be a need for information that may be useful in their future employment, to have a clear picture of where and how they can work with their chosen profession. They emphasized the importance of professional internships and organization. According to them, regardless of their faculty and major, they would like to participate in professional internships, where they could gain the skills they have no opportunity for within the university walls. For this, it is essential to transform the teacher-student relationships.

Based on the unanimous opinion of the students, work and connections related to their profession would be the most useful, for which the contribution of the higher education institutions would be needed. At the moment, it is not easy to judge how strong a risk factor working during university is, as people deliberately chose employment instead of the university.

This decision raises another question: How could Higher Education Institutions benefit from the comments and experiences of these students?

Acknowledgement

This work was supported by the The Ministry of Human Capacities under Grant ÚNKP-18-3. Supported through the New National Excellence Program of the Ministry of Human Capacities.

Project no. 123847 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K_17 funding scheme.

References

[1] Riggert, S. C., Boyle, M., Petrosko, M. J., Ash, D. & Rude-Parkins, C. (2006) Student Employment and Higher Education: Empiricism and Contradiction. Review of Educational Research, 76(1), 63-92

[2] Roshchin, S. & Rudakov, V. (2015) Russian University student and the combination of study and work: is it all about earning, learning or job market signaling? DOI:10.2139/ssrn.2566775

[3] Bocsi, V.; Ceglédi, T.; Kocsis, Zs.; Kovács, K. E.; Kovács, K.; Müller, A. É.; Pallay, K.; Szabó, B. É.; Szigeti, F.; Tóth, D. A. (2018) The discovery of the possible reasons for delayed graduation and dropout in the light of a qualitative research study. Journal of Adult Learning Knowledge and Innovation, 3(1), 27-38

[4] Beerkens, M., Mägi, E. & Lill, L. (2011) University studies as a side job. Causes and consequences of massive student employment in Estonia. *High Education*, 61(6), 679-692, DOI: 10.1007/s10734-0109356-0

[5] Sanchez-Gelabert, A., Figueroa, M. & Elias, M. (2017) Working whilst studying in higher education. The impact of the economic crisis on academic and labour market success. *European Journal of Education,* 52(2), 232-245, DOI: 10.1111/ejed.12212

[6] Masevičiūtė, K., Šaukeckienė, V., & Ozolinčiūtė, E. (2018) Eurostudent VI Combining Studies and Paid Jobs. ISBN 978-609-468-169-1

[7] Lukács, F. & Sebő, T. (2015) Az egyetemi lemorzsolódás kérdőíves vizsgálata (Questionnaire survey on university drop-out) *Iskolakultúra*, 10, 78-86

[8] Fenyves, V., Bácsné Baba, É., Szabóné Szőke, R., Kocsis, I., Juhász, Cs., Maté, E. & Pusztai, G. (2017) Kísérlet a lemorzsolódás mértékének és okainak megragadására a Debreceni Egyetem Gazdaságtudományi Kar példáján (Attempt to understanding the extent and causes of dropouts at the University of Debrecen, Faculty of Economics). *Neveléstudomány,* 3, 5-14

[9] Derényi, A. (2015) Bizonyítékokra alapozott kormányzás és a kommunikáció képzés (Evidance-based Governance and Communication Training) Jelkép, 1-21

[10] Darmody, M. & Smyth, E. (2008) Full-time students? Term-time employment among higher education students in Ireland. *Journal of Education and Work,* 21(4), 349-362

[11] Perna, L. (2010) *Understanding the Working College Student New Research and Its Implications for Policy and Practice*. Sterling: Stylus Publishers

[12] Pusztai, G. (2014) The Effects of Institutional Social Capital on Students' Success in Higher Education. *Hungarian Educational Research Journal,* 4(3) DOI: 10.14413/HERJ2014.03.06

[13] McCoy, S. & Smyth, E. (2004) *At work in school.* Dublin: ESRI/Liffey Press

[14] Nagy, Mária. 2006. *A tanárok "hangja", osztálytermi viselkedésük.* Budapest: Országos Közoktatási Intézmény

[15] Kovács, K. et al. (2019) *Lemorzsolódott hallgatók (Dropout students).* Debrecen: Egyetemi Kiadó

[16] Pusztai, G. & Kocsis, Zs. (2019) Combining and Balancing Work and Study on the Eastern Border of Europe. *Social Sciences*, 8(6)

[17] Fónai, M. (2018) Hallgatói lemorzsolódás a Debreceni Egyetemen (Student dropout at the University of Debrecen) In: Pusztai, G. & Szigeti, F. (eds.). *Lemorzsolódás és perzisztencia a felsőoktatásban* (pp. 239-250) Debrecen: Debreceni Egyetemi Kiadó

[18] Curtis, S. & Shani, N. (2002) The effect of taking paid employment during term-time on students' academic studies. *Journal of Further and Higher Education,* 26(2), 129-138

[19] Kocsis, Zs. (2019) How to support working students during their studies? In Erdei G., Erika Juhász, Salih Sahin & Adnan Kan (eds), Ways of Promoting Excellence in Higher Education

Appendix

	N	Average	Standard deviation
I was often short on time.	539	2,3766	1,1247
I found a better opportunity to succeed.	541	2,2773	1,15823
The teachers were incorrect.	542	2,1384	1,07794
Exams and papers always had worse results than expected.	540	2,1074	0,99327
I took too much work.	541	2,0647	1,07904
After admission, I realized I was not interested in the major.	554	2,0271	1,07964
After the failures, I no longer trusted myself.	543	1,9982	1,13401
The administration was not supportive.	537	1,9907	1,06588
I did not even know what to do, I always lacked information.	541	1,9834	1,02094
I could not bear the costs.	550	1,9818	1,14187
I did not care, studying was not important.	546	1,8095	0,97697
I could hardly process the textbooks and notes.	543	1,7348	0,9049
I ran out of exam opportunities.	537	1,7095	0,9875
I was going out too much.	543	1,698	0,93168
I was transferred to fee-paying training.	545	1,6202	1,05752
I could not pay attention during classes.	536	1,5746	0,826
I did not even want to study that major.	548	1,5712	0,93307
I missed my friends and/or family.	538	1,5576	0,842
The other students did not help me.	545	1,5486	0,794
I ran out of state-funded semesters.	541	1,5065	0,90598
Due to health reasons.	537	1,3203	0,7665

Hierarchical Agglomerative Clustering of Selected Hungarian Medium Voltage Distribution Networks

Attila Sandor Kazsoki[1,2], Balint Hartmann[2]

[1] Department of Electric Power Engineering, Budapest University of Technology and Economics, Egry József u. 18, 1111 Budapest, Hungary

[2] Department of Environmental Physics, Centre for Energy Research, KFKI Campus, Konkoly-Thege Miklós u. 29-33, 1121 Budapest, Hungary;

kazsoki.attila@vet.bme.hu, hartmann.balint@energia.mta.hu

Abstract: Nowadays the increase of photovoltaic penetration and simultaneously, the decentralization of electricity system, poses a number of challenges for distribution system developers and operators. The spread of high output power photovoltaic power plant connections demands the development of a network infrastructure. The analysis of development directions can be done with software simulation, for which network models are needed, which can characterize real networks well. To create such reference networks, knowing existing topologies, hierarchical agglomerative clustering can be a solution. When the parameters of the clusters are specified well, their software implementation can be done. In this study, a possible clustering process of selected Hungarian medium voltage overhead networks (including the determination of the optimal cluster number too), and the formulated network clusters are presented. The clustering of twenty selected 22 kV medium voltage networks was done using hierarchical agglomerative clustering. Then the optimal cluster number was determined. Based on Davis-Bouldin and Silhouette criterions, this cluster number was four. Two of the four generated clusters are single clusters, containing only one feeder. The size and looping of the characterized sample networks are well observable. In this paper a method has been created to generate medium voltage distribution network models, which can be used to simulate the effects of the growing photovoltaic penetration in the Hungarian distribution network.

Keywords: distribution network; network clustering; hierarchical agglomerative clustering

1 Introduction

Nowadays, in both domestic and international energy market trends, photovoltaic penetration quickly increases. Photovoltaic systems, considering their output power, covering the entire power plant range (from household size small to size small and over that). A large amount of this (approx. 333 MWp, which is 56%) is household size small power plants. Due to a change of the renewable support system at the end of 2016, the number of "applications for licenses for the installation" of photovoltaic power plants with 500 kVA output power increased significantly. Based on the photovoltaic market predictions, in the next decade, the number of small power plants is going to increase, which will increasingly decentralize the structure of the electricity infrastructure (firstly at the distribution voltage level) [1] [2]. In Table 1, the increasing tendency of the built-in photovoltaic capacity is shown.

Table 1
Cumulative photovoltaic capacity development in Hungary [1] [2]

	2015	**2016**	**2017**	**2018**
Built-in household size small power plant capacity [MWp]	129	164	241	324
Built-in small power plant capacity [MWp]	16	27	61	approx. 240
Total built in photovoltaic capacity [MWp]	172	235	344	665

In order for the low, and medium voltage networks to approximate the smart grid structure, electricity infrastructure development is necessary. Modeling of these distribution networks (here on medium voltage level) is essential to determine development directions and to answer emerging questions. For these simulations, the software implementation of medium voltage networks is recommended. Because in the examined area, there is a significant number of varied topology medium voltage networks, their software implementation and running simulations is a powerful time and resource consuming task. It is recommended to formulate reference networks with which the real feeders can well be described. Such reference networks can be created by clustering real networks. These distribution network models can be approximated more precisely than the mathematical models described in the literature [5]. Thus, real decision situations can be handled by the generated reference networks. In this paper a method has been created to generate Hungarian medium voltage distribution network models.

In Section 2 the used hierarchical agglomerative clustering method is presented, with which the numerous medium voltage networks are decreased to a manageable number of clusters. In Section 3.1, the examined distribution feeders are presented. In Section 3.2, the clustering method and principal component analysis are used, in Section 3.3 the determination process of the optimal cluster number are presented. Section 4 d the results of clustering and the generated clusters are presented.

2 Data Analysis Techniques

Data mining techniques can be used to get significant information from the examined database. [2].By reviewing a number of studies (Table 2) in which some kin fog these techniques (typically clustering methods) were used for grouping low or medium voltage electricity network, it can be said that classification, K-means (and K-medoids) clustering and hierarchical clustering are the most frequently used methods. [3]

Table 2
Methods of network analyzing techniques

Name of method	**Referenced in literature**	**Number of groups**	**Number of analyzed networks**
Classification	[3] [4] [5] [6] [7] [8] [9] [10] [11] [30]	Small number (approx. 3–5)	some 100– some 1000
Dimension reduction (SOM)	[3] [4] [12] [13]	Medium number (approx. 8–9)	some 100
Agglomerative clustering	[3] [4] [10] [14] [16] [26]	Large number (approx. 10–25)	some 100–10000+
Partitional clustering	[3] [4] [10] [17] [26]	–	some 100
K-means clustering	[3] [4] [14] [17] [19] [20] [22] [23] [24] [25] [26] [27] [29]	Variable number (approx. 2–12)	some 10–10000+
K-medoids clustering	[3] [4] [17] [23]	Medium number (approx. 8–9)	some 1000

The description of the mentioned methods are not presented in this paper, the clustering processes, the advantages and disadvantages of them can be found in another review paper of the authors [3].

The number of the examined feeders is relatively small (20), hierarchical agglomerative clustering can be used.

2.1 Hierarchical Agglomerative Clustering

“In hierarchical clustering, clusters are determined with the relative distance (Euclidean distance) between the examined data points. The main concept is that a selected item is more tied to a closer data point that to a farther one.” [3] [15]

At the beginning of this process, all the data points (*n*) are considered as a single cluster. At each step of the algorithm, all data points are moved to a larger cluster. The clustering algorithms stop when all the *n* points are in the same cluster. As the graphical representation of the clustering, a tree-structure (dendrogram) can be used, which can be cut off at any level. At this level, the leave elements of the tree represent the clusters [2] [3] [4] [14].

The advantage of the algorithm is that "it corrects the distance errors between the local minimum and the center of the clusters" [3] [4]. Besides the positive attributions, there are many negative ones too. The greatest one is the irrevocability of cluster merges. In one step if two clusters are combined, they cannot be divided again later, since the new cluster is used in the future steps of the algorithm. These steps are critical because incomplete mergers give incorrect results (clusters) [3] [4] [14]

3 Clustering Method

3.1 Input Network Data

In this publication, 20 selected Hungarian medium voltage, 22 kV overhead distribution feeders which can be found in the same distribution system operator area, but at four different locations were examined. At the selection of the feeders, the most important criteria were to be able to physically accommodate (approx. 500 kVA) photovoltaic small power plants (output power approx. at least 500 kVA, area is at least 1 ha). Half of the examined networks are located in rural areas and the other half of them are located in suburban settings.

Here, the examined networks are handled as graphs. These graphs can be characterized by specific mathematical variables, such as:

- Total node number
- Average node degree
- Clustering coefficient (CC)
- Characteristic path length (CPL)

The average node degree can be defined with Eq. 1. [4]

$$average\ node\ degree = \frac{2 * E}{N} \tag{1}$$

"where E is the number of edges, N is the number of nodes of the graph". [4]

The clustering coefficient can be defined with Eq. 2. [4]

$$CC = \frac{1}{n} * \sum_{i=1}^{n} \frac{2|\{e_{j,k}: v_j, v_k \in N_i, e_{j,k} \in E\}}{k_i * (k_i - 1)} \tag{2}$$

"where $e_{j,k}$ is edge between vertex v_j with v_k; N_i is the set of immediately connected neighboring vertices for a vertex v_i; k_i is the element number of N_i and n is the size of the graph". [4]

The CPL is interpreted as the impedance values of the lines (feeders). It can be defined with Eq. 3. [4]

$$CPL = \frac{1}{n * (n-1)} * \sum_{i \neq j}^{k} d(v_i, v_j) \qquad (3)$$

“where n is the size of the graph, and d is the distance between any two nodes of the graph”. [4]

The values of these parameters can be found in Table 3. To calculation they, the built-in functions of MATLAB R2018b were used.

Table 3
The parameters of the examined feeders

Network identifier	Node number	Average node degree	CC	CPL
N1	350	2.0057	0.7230	27.0648
N2	100	2.0000	0.7273	12.4315
N3	153	1.9869	0.7349	18.0999
N4	193	1.9896	0.7375	25.2386
N5	32	1.9375	0.7646	6.1734
N6	835	2.0216	0.7225	48.4478
N7	82	2.0000	0.7224	11.5414
N8	228	2.0175	0.7294	31.2079
N9	244	2.0164	0.7279	22.4480
N10	243	1.9918	0.7373	22.2443
N11	125	1.9840	0.7347	20.6679
N12	180	1.9889	0.7378	21.1089
N13	59	1.9661	0.7480	10.1473
N14	153	1.9869	0.7383	19.4555
N15	140	1.9857	0.7438	16.7857
N16	491	2.0000	0.7238	27.9195
N17	175	1.9886	0.7290	18.8393
N18	89	1.9775	0.7367	15.4949
N19	166	1.9880	0.7301	21.4499
N20	64	1.9688	0.7370	10.8889

Based on the parameters, it can be said that the feeders have a varied size (node number) and topology.

According to the confidentiality agreement signed between the Centre for Energy Research and the Distribution System Operator (DSO), the authors are not allowed to publish raw data.

3.2 Principal Component Analysis

In this article, hierarchical agglomerative clustering is used.

The most frequently used variables to describe clusters are the size of the network (number of nodes), the degree distribution of feeders (average node degree), the clustering coefficient and the characteristic path length of the feeders. The values of the parameters can be seen in Table 3.

The network analysis is a procedure, in which often more than two variables are taken into account. The handling of a large dataset of multiple variables as a compact unit is a tough assignment. It is recommended to decrease the number of variables, without losing information. A solution can be for this reduction is the principal component analysis (PCA). Using PCA the nature of the array can be written with fewer mathematical parameters (factors) that contain most of the original information. Another task is to describe the nature of correlation between the original variables with the principal components [15] [18].

In this case, the feeders are characterized by four variables. Treating them as a unit is not easy, it is recommended to complete the principal component analysis. To get the values of the principal components, MINITAB 18.0, a statistical software was used. [4] The description of the algorithm used in the MINITAB software can be found at [36].

Based on the scree plot of the main components (seen in Figure 1) and using the “Elbow” criterion, the optimal number of principal components can be determined, which is equal to 2. This means that the examined feeders can be described with the first and second principal components [4].

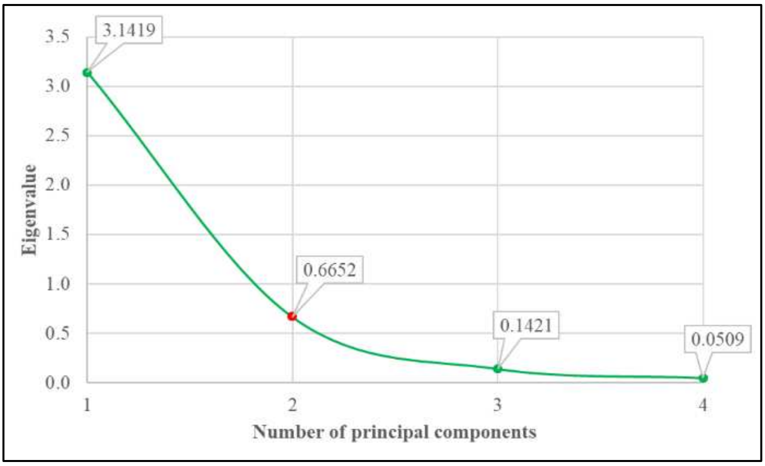


Figure 1
The scree plot for the eigenvalue of principal components for the feeders
(the Elbow point is marked with a red dot)

The numeric values of the principal components for the original parameters and for the examined networks can be seen in Tables 4 and 5, respectively [4]. These values can be used in the clustering process, using a hierarchical agglomerative clustering algorithm.

Table 4

Values of the principal components for the variables

	PCA 1st component	**PCA 2nd component**
Node number	0.49287	-0.53713
Average node degree	0.51853	0.37171
CC	-0.47021	-0.62420
CPL (impedance)	0.51682	-0.42860

Table 5

Values of the principal components for the feeders

Network identifier	**The eigenvalue of the 1st principal component**	**The eigenvalue of the 2nd principal component**	**The value of the 1st principal component**	**The value of the 2nd principal component**
N1	1.711	0.262	187.192	-199.301
N2	-0.136	1.301	56.407	-58.752
N3	-0.379	0.164	85.448	-89.659
N4	-0.079	-0.395	108.853	-114.204
N5	-4.092	-1.722	19.608	-19.591
N6	4.669	-1.817	437.294	-468.968
N7	-0.003	1.703	47.078	-48.699
N8	1.637	0.266	129.207	-135.547
N9	1.230	0.700	132.565	-140.386
N10	0.112	-0.351	131.950	-139.777
N11	-0.381	0.082	72.974	-75.721
N12	-0.221	-0.198	100.311	-105.452
N13	-2.253	-0.408	34.991	-35.776
N14	-0.466	-0.115	86.147	-90.242
N15	-0.938	-0.314	78.357	-82.119
N16	1.949	-0.351	257.125	-275.406
N17	0.041	0.462	96.677	-101.788
N18	-1.037	0.177	52.553	-54.171
N19	0.091	0.283	93.590	-98.074
N20	-1.612	0.277	37.846	-38.772

3.3 Optimal Cluster Number

At the first step of the agglomerative clustering, the cluster number is decided. The determination of the optimal cluster number is based on the simultaneous application of Davies-Bouldin (*DB*) validity criteria and Silhouette (*Si*) validity criteria [4] [21] [28]. For the determination of *DB* and *Si* values the built-in functions of MATLAB R2018b academic version is used.

3.3.1 Determination of the Optimal Cluster Number

Empirically, in the case of a small number of data sets (around some 10 to 100), the optimal cluster number is between 2 and 5. It coincides with what was described in [28]. Therefore, the minimum number of the clusters can be determined with Eq. 4, and the maximum number of clusters can be determined with Eq. 5.

$$M_{min} = 1 + 1 = 2 \tag{4}$$

"where M_{min} is the minimal number of the clusters" [28].

$$M_{max} = \lceil \sqrt{N/2} \rceil + 1 = \lceil \sqrt{20/2} \rceil + 1 = 5 \tag{5}$$

"where M_{max} is the maximal number of the clusters, N is the number of examined data points" [28].

$$M_{opt} = [M_{min}; M_{max}] \tag{6}$$

In this paper, the optimal cluster number has been investigated in the range, defined in Eq. 6 (cluster number is 2, 3, 4 or 5), their values are calculated with the simultaneous application of Davies-Bouldin and Silhouette criterions.

3.3.2 Davies-Bouldin Criterion

"The Davies-Bouldin evaluation is an object consisting of sample data, clustering data, and Davies-Bouldin criterion values used to evaluate the optimal number of clusters. This criterion is based on a ratio of within- and between-cluster distances." [4] [31] The Davies-Bouldin index can be defined with Eq. 7 [4] [28] [31] [32].

$$DB = \frac{1}{k} * \sum_{i=1}^{k} max_{j \neq i}\{D_{i,j}\} \tag{7}$$

"where $D_{i,j}$ is the within-to-between cluster distance ratio for the i^{th} and j^{th} clusters" [28]. The mathematical description of this distance can be seen in Eq. 8 [24] [25] [28].

$$D_{i,j} = \frac{(\overline{d}_i + \overline{d}_j)}{d_{i,j}} \tag{8}$$

“where $\overline{d_i}$ is the average distance between each point i and the centroid of the i^{th} cluster, $\overline{d_j}$ is the average distance between each point and the centroid of the j^{th} cluster, $d_{i,j}$ is the Euclidean distance between the centroids of the i^{th} and j^{th} clusters” [24] [25] [28].

There is the worst-case for cluster i when $D_{i,\mathrm{j}}$ has a global maximum at within-to-between cluster ratio. The optimal cluster number can be identified when the Davies-Bouldin index has a global minimum [4] [24] [25] [28].

The objective function of the optimization problem based on Davies-Bouldin validity index is defined with Eq. 9.

$$M_{opt} = \min_{m\in[M_{min};M_{max}]} DB_m \tag{9}$$

“where M_{opt} is the optimal number of the clusters, m is the number of clusters” [28].

3.3.3 Silhouette Criterion

“The value of the Silhouette criterion is a metric of how similar is the examined point to the other points in the same cluster, compared to points in other clusters.” [4] [33] The Silhouette value (Si) for the point i, can be defined with Eq. 10 [4] [28] [33].

$$S_i = \frac{(b_i + a_i)}{\max\{a_j, b_i\}} \tag{10}$$

“where a_i is the average distance from point i to the other points of the cluster, b_i is the minimum average distance from point i to the points in another cluster” [4] [28] [33].

The value of the Si can be in the range from -1 to $+1$. If it is closer to $+1$, point i is well-matched to its own, and poorly-matched to the other clusters. The optimal cluster number is then when the Silhouette index has a global maximum [4] [28] [33].

The objective function of the optimization problem based on Davies-Bouldin validity index is defined with Eq. 11.

$$M_{opt} = \max_{i=m\in[M_{min};M_{max}]} S_i \tag{11}$$

“where M_{opt} is the optimal number of the clusters, m is the number of clusters”[28].

The results of the two methods described above can be seen in Figure 2. The values of the validity indexes for each cluster number are depicted in Figure 2. Based on these, the optimal cluster number is 4 [4] [28] [33].

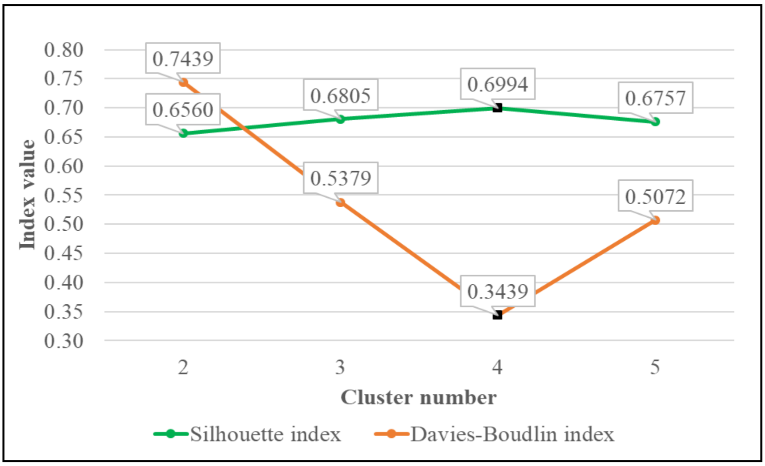


Figure 2
The values of the Davies-Bouldin and Silhouette evaluations in the case of 2, 3, 4, and 5 clusters calculated with MATLAB R2018b built-in functions

The clustering algorithm was run 20 times to avoid local minima. The result of clustering was always the same. The clustering algorithm was convergent.

4 Results

In this publication, from the previously presented 20 medium voltage networks described above, by using principal component analysis and hierarchical agglomerative clustering algorithm, 4 network clusters were created. Clustering was done using the tutorial version of MINITAB 18.0 statistical software.

In MINITAB 18.0, the agglomerative clustering method is based on the complete linkage method (also called furthest neighbor method), in which “the distance between two clusters is the maximum distance between an observation (feeder or data point) in one cluster and an observation (feeder or data point) in the other cluster” [37]. The complete distance is calculated with Eq. 10 [37].

$$d_{m,j} = max\{d_{k,j}; d_{l,j}\} \tag{10}$$

where $d_{m,j}$ is the distance between clusters m and j; m is a merged cluster that consists of clusters k and l, with $m = (k,j)$; $d_{k,j}$ distance between clusters k and j; $d_{l,j}$ distance between clusters l and j [37].

The graphical representation of the clustering (dendrogram), can be seen in Fig. 3. In this figure, the clusters are colored respectively.

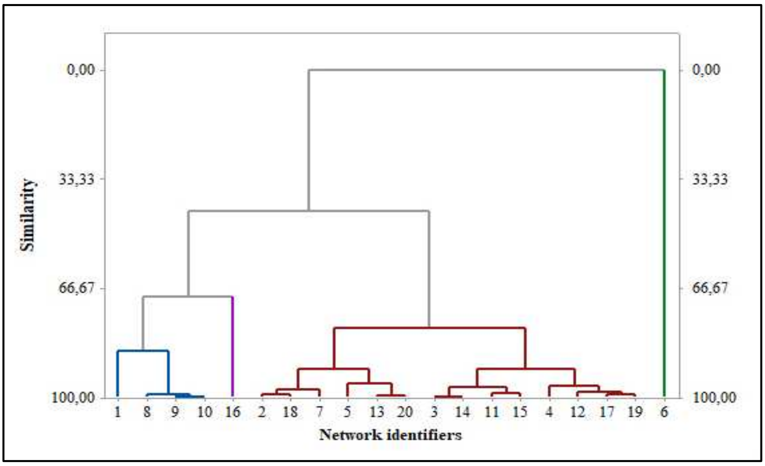


Figure 3
Dendrogram for the clustered feeders

For the graphical representation of the eigenvalues of principal components for the clustered feeders see Figure 4.

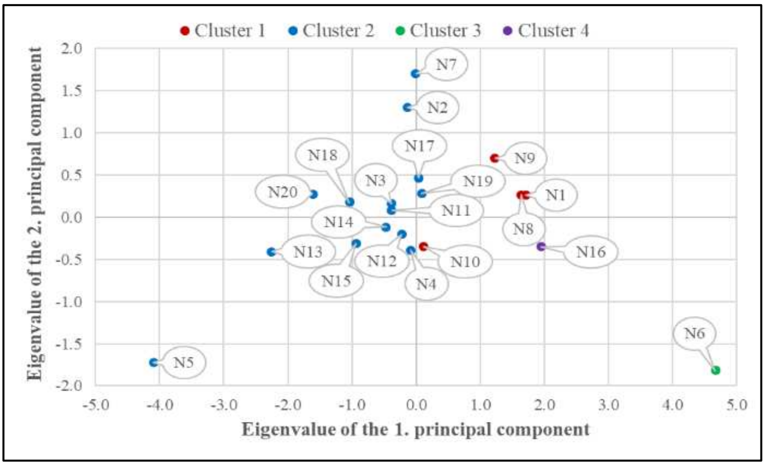


Figure 4
The score plot for the eigenvalues of principal components for the clustered feeders

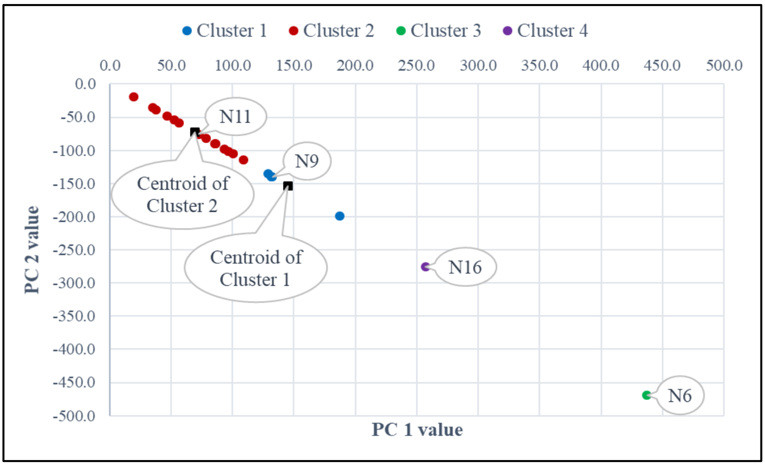


Figure 5

The score plot for the values of principal components for the clustered feeders

In Figure 5, the centroids of the non-single element clusters are marked with black square markers, and the markers of real networks closest to the centroids are labeled too. The values of the centroids (see Table 6), were determined as the average of the data points in a cluster with them. The centroids of the one element clusters are the feeders, included in each cluster.

Table 6

Calculated centroids of clusters

Variable	Cluster1	Cluster2	Cluster3	Cluster4
PCA 1st component	145.229	69.346	437.294	257.125
PCA 2nd component	-153.753	-72.359	-468.968	-275.406

The final partition of clustering and the various distance metrics of the clusters can be seen in Table 7.

Table 7

Final partition of clustering

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster 1	4	5134.500	30.969	61.933
Cluster 2	14	21507.800	34.968	72.514
Cluster 3	1	0.000	0.000	0.000
Cluster 4	1	0.000	0.000	0.000

For the average numeric values of the variables, see Table 8.

Table 8

The average value of parameters in the four clusters

	Node number	Average node degree	CC	CPL
Cluster 1	266.250	2.008	0.729	25.741
Cluster 2	122.214	1.982	0.737	16.309
Cluster 3	835.000	2.022	0.723	48.448
Cluster 4	491.000	2.000	0.724	27.919

In order to illustrate the characteristics of the typical network topology of clusters, the representation of sample networks, which are the closest to the previously defined centroids, are presented. These networks with their identifier are shown in Figure 5, and their topology can be seen in Figures 6-9.

In Cluster 1, there are 4 weakly looped networks. While the element number of the cluster is not too large (4) and the feeders are fairly similar, the sum of squares of distances within the cluster is approximately the quarter of the same value in Cluster 2. The distances from the centroids are in the same range for Clusters 1 and 2, so it can be said that these clusters are compact. The graphical representation of feeder N9 is shown in Figure 6.

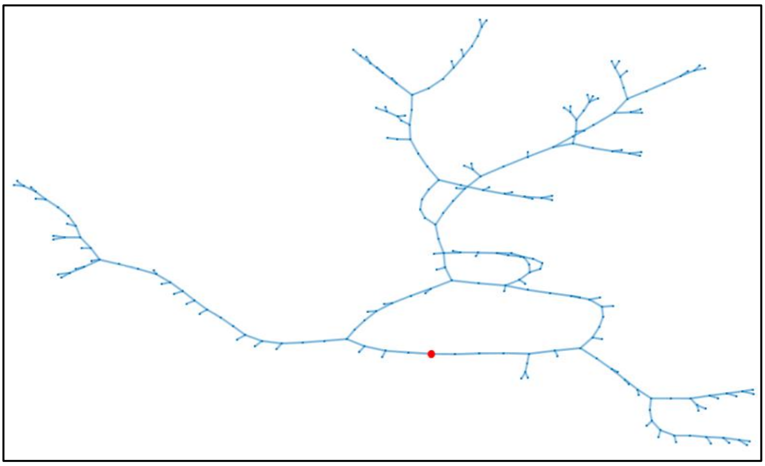


Figure 6

The topological representation of feeder N9 in Cluster 1

Topology N9 is a medium-sized, weakly looped medium voltage (22 kV) overhead network, located in a suburban area. In Figure 6, the HV/MV (132 kV/22 kV) substation is marked with red.

Cluster 2 is the highest element number cluster with 14 feeders. The networks in Cluster 2 are small and medium size and have throughout radial topology, located in a rural area. For the graphical representation of feeder N11, see Figure 7.

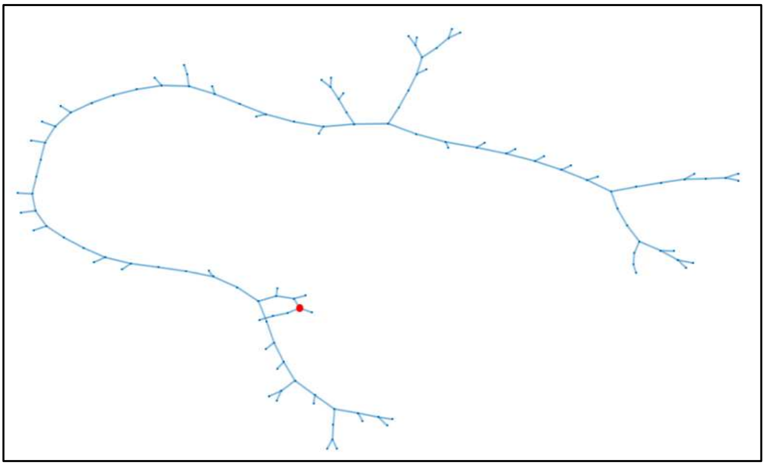


Figure 7
The topological representation of feeder N11 in Cluster 2

Topology N11 is a radial network, placed in the rural area. From the four reference networks, this is the smallest one (smallest node number and CPL value). In Figure 7, the HV/MV (132kV/22kV) substation is marked with cyan.

Clusters 3 and 4 are single clusters. The networks in these clusters are fairly large, and the CCs are the biggest ones too. These networks are located in a suburban area (Cluster 3). These clusters cannot be as relevant as Clusters 1 and 2, because the element number is only 1. The graphical representation of feeder N6 can be seen in Figure 8.

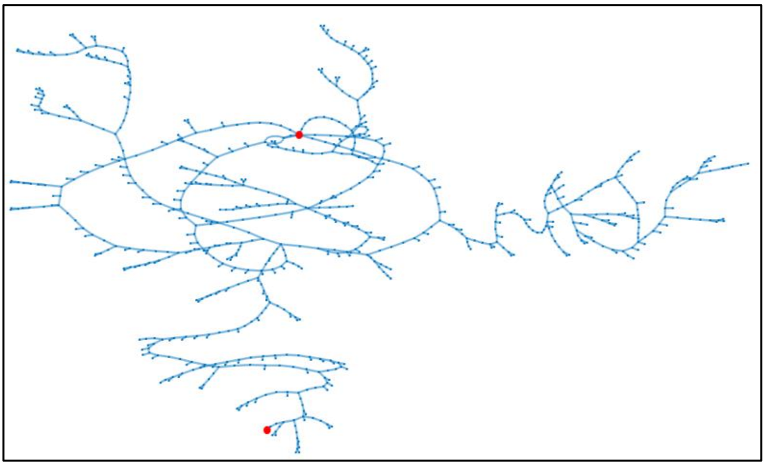


Figure 8
The topological representation of feeder N6 in Cluster 3

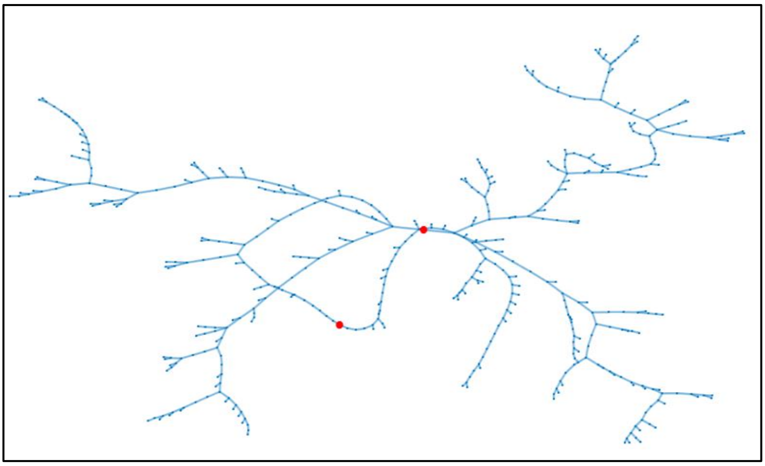


Figure 9
The topological representation of feeder N16 in Cluster 4

Topology N6 is a large, and heavily looped medium voltage network, located near the suburban area of the capital city of the county. Out of the four reference networks, this is the biggest one (highest node number and CPL value). In Fig. 8, the HV/MV (132 kV/22 kV) substations are marked with red (the two substations are different).

Topology N16 is similar to N9, but the node number is much higher. This topology is a large (not as large as N6), weakly looped medium voltage network, in a suburban area of a town. As in the case of N6 (Cluster 3), this topology also contains loops, but less. In Figure 9, the HV/MV (132 kV/22 kV) substations are marked with red (the two substations are different).

Conclusions

In this study, a network clustering method on Hungarian medium voltage distribution feeders has been displayed, which is suitable for the efficient processing of smaller or larger amounts of a data array. Found on the international studies published in the literature, at the formulation of network groups the agglomerative hierarchical clustering and PCA were used. With PCA, the number of the original variables space was reduced from four to two, and this two-dimensional component space was clustered. At the first step, as an input parameter of the clustering algorithm, the optimal number of the clusters was described using the Davies-Bouldin and Silhouette criterions. Both methods led to the same result, the optimal cluster number is 4. The results of the clustering were presented in Section 4. As the topologies of feeders are fairly varied, distinct clusters have been formulated. Two of the clusters (Clusters 3 and 4) are single clusters because in each of these clusters the is only one feeder.

The data processing and clustering method presented in this paper can be well used for clustering networks that cover a physically large area (eg. a country), formulating network topologies specific to the examined area.

Herein, a method has been created to generate Hungarian medium voltage distribution network models, which can be used to simulate the effects of the growing photovoltaic penetration within the Hungarian distribution network. In addition, these results can also help in modeling the voltage and power changing effects on these networks. On the reference networks the effects of growing electrical car numbers, the energy storage penetration and the opportunities for smart grid development can also be simulated.

Acknowledgment

The VEKOP-2.3.2-16-2016-00011 grant is supported by the European Structural and Investment Funds, financed jointly by the European Commission and the Hungarian Government.

References

[1] A. Whiteman, J. Esparrago, S. Rueda, S. Elsayed, I. Arkhipova, Renewable capacity statistics, International Renewable Energy Agency (IRENA) 2019

[2] Data of license exemptioned small power plants and household size small power plants between 2008 and 2017, Hungarian Energy and Public Utility Regulatory Authority, 2018

[3] A. S. Kazsoki, B. Hartmann, Data Analysis and Data Generation Techniques for Comparative Examination of Distribution Network Topologies, in: International Review of Electrical Engineering (IREE) Vol. 14, 2019: pp. 32-42

[4] A. S. Kazsoki, B. Hartmann, Typologization of medium voltage distribution networks using data mining techniques: A case study, 2019 7th International Youth Conference on Energy (IYCE), pp:1-8, 2019

[5] G. A. Pagani, From the Grid to the Smart Grid, Topologically, University of Groningen, 2014

[6] G. A. Pagani, M. Aiello, Towards Decentralization: A Topological Investigation of the Medium and Low Voltage Grids, 2011

[7] V. Lenz, Generation of Realistic Distribution GridTopologies Based on Spatial Load Maps, Swiss Federal Institute of Technology (ETH) Zurich, 2015

[8] P. Hines, S. Blumsack, E. C. Sanchez, C. Barrows, The Topological and Electrical Structure of Power Grids, in: 2010 43rd Hawaii Int. Conf. Syst. Sci., 2010: pp. 1-10

[9] Y. Wang, J. Zhao, F. Zhang, B. Lei, Study on structural vulnerabilities of power grids based on the electrical distance, in: IEEE PES Innov. Smart Grid Technol., 2012: pp. 1-5

[10] K. P. Schneider, YousuChen, D. P. Chassin, R. G. Pratt, D. W. Engel, S. E. Thompson, Modern Grid InitiativeDistribution Taxonomy Final Report, Pacific Northwest National Laboratory, 2008

[11] K. Vill, A. Rosin, Identification of Estonian weak low voltage grid topologies, in: 2017 IEEE Int. Conf. Environ. Electr. Eng. 2017 IEEE Ind. Commer. Power Syst. Eur. (EEEIC / I&CPS Eur., 2017: pp. 1-5

[12] F. Dehghani, H. Nezami, M. Dehghani, M. Saremi, Distribution feeder classification based on self-organized maps (case study: Lorestan province, Iran), in: 2015 20th Conf. Electr. Power Distrib. Networks Conf., 2015: pp. 27-31

[13] Y. Li, P. Wolfs, Preliminary statistical study of low voltage distribution feeders under a representative HV network in Western Australia, in: AUPEC 2011, 2011: pp. 1-6

[14] A. Méffe, C. Oliveira, Classification techniques applied to electrical energy distribution systems, in: CIRED 2005 - 18th Int. Conf. Exhib. Electr. Distrib., 2005: pp. 1-5, Z. Ilonczai, Klaszter-analízis és alkalmazásai, Eötvös Loránd University, Budapest, 2014

[15] Z. Ilonczai, Cluster analysis and their applications (M.Sc. thesis), Eötvös Loránd University, Budapest, 2014

[16] Y. Li, P. Wolfs, Statistical identification of prototypical low voltage distribution feeders in Western Australia, in: 2012 IEEE Power Energy Soc. Gen. Meet., 2012: pp. 1-8

[17] R. J. Broderick, J. R. Williams, Clustering methodology for classifying distribution feeders, in: 2013 IEEE 39th Photovolt. Spec. Conf., 2013: pp. 1706-1710

[18] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining: Pearson New International Edition, Pearson Education Limited, 2013

[19] R. J. Broderick, K. Munoz-Ramos, M. J. Reno, Accuracy of clustering as a method to group distribution feeders by PV hosting capacity, in: 2016 IEEE/PES Transm. Distrib. Conf. Expo., 2016: pp. 1-5

[20] J. Watson, N. Watson, D. Santos-Martin, S. Lemon, A. Wood, A. Miller, Low Voltage Network Modelling, EEA Conf. Exhib. (2014) 15

[21] C. Gonzalez, J. Geuns, S. Weckx, T. Wijnhoven, P. Vingerhoets, T. De Rybel, J. Driesen, LV distribution network feeders in Belgium and power quality issues due to increasing PV penetration levels, in: 2012 3rd IEEE PES Innov. Smart Grid Technol. Eur. (ISGT Eur., 2012: pp.

[22] J. Dickert, M. Domagk, P. Schegner, Benchmark Low Voltage Distribution Networks Based on Cluster Analysis of Actual Grid Properties, 2013

[23] J. Cale, B. Palmintier, D. Narang, K. Carroll, Clustering distribution feeders in the Arizona Public Service territory, in: 2014 IEEE 40th Photovolt. Spec. Conf., 2014: pp. 2076-2081

[24] I. Borlea, R. Precup, F. Dragan, and A. Borlea, Centroid Update Approach to K-Means Clustering, Adv. Electr. Comput. Eng., Vol. 17, No. 4, pp. 3-10, 2017

[25] S. Chakraborty and S. Das, k − Means clustering with a new divergence-based distance metric: Convergence and performance analysis, Pattern Recognit. Lett., Vol. 100, pp. 67-73, 2017

[26] S. Zahra, M. A. Ghazanfar, A. Khalid, M. A. Azam, and U. Naeem, Novel Centroid Selection Approaches for KMeans-Clustering Based Recommender Systems, 2015

[27] R. Zall, M. R. Kangavari, On the Construction of Multi-Relational Classifier Based on Canonical Correlation Analysis, International Journal of Artificial Intelligence, Vol. 17, No. 2, pp. 23-43, 2019

[28] Q. Zhao, Cluster Validity in Clustering Methods, University of Eastern Finland, 2012

[29] I. Bonet, A. Escobar, A. Mesa-múnera, and F. Alzate, Clustering of Metagenomic Data by Combining Different Distance Functions, Acta Polytech. Hungarica, Vol. 14, No. 3, pp. 223-236, 2017

[30] A. Hamouda, Improvement of the Power Transmission of Distribution Feeders by Fixed Capacitor Banks, Acta Polytech. Hungarica, Vol. 4, No. 2, pp. 47-62, 2007

[31] MATLAB R2018b, Davies-Bouldin evaluation, (n.d.). https://www.mathworks.com/help/stats/clustering.evaluation.daviesbouldin evaluation-class.htm

[32] Davies, D. L., and D. W. Bouldin. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. PAMI-1, No. 2, 1979, pp. 224-227

[33] MATLAB R2018b, Silhouette evaluation, (n.d.) https://www.mathworks. com/help/stats/clustering.evaluation.silhouetteevaluation-class.html

[34] Kaufman L. and P. J. Rouseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: John Wiley & Sons, Inc., 1990

[35] Rouseeuw, P. J., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics. Vol. 20, No. 1, 1987, pp. 53-65

[36] MINITAB 18.0, Principal Component Analyzis, (n.d.). https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/methods-and-formulas/methods-and-formulas/

[37] MINITAB 18.0, Linkage clustering methods, (n.d.). https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-observations/methods-and-formulas/linkage-methods/

Tree Growth Simulation based on Ray-Traced Lights Modelling

Bence Tamás Tóth

John von Neumann Faculty of Informatics
Óbuda University
Bécsi street 96/B, Budapest, H-1034, Hungary,
toth.bence@nik.uni-obuda.hu

Sándor Szénási

Faculty of Economics
J. Selye University
Bratislavská cesta 3322, 945 01, Komárno, Slovakia,
szenasis@ujs.sk

Abstract: In the fields of forestry and horticulture, it is necessary to have forecasts about the growth of trees. This process is affected by a lot of external factors like weather, light conditions, other objects, etc. There are several already existing methods for this purpose, but these can give only rough estimations. This paper presents a novel solution, based on the simulation of the growth of the tree. During this process, the application takes into account the environment of the tree and the properties of the species, which parameters are all easily configurable. The presented application can simulate not just one, but a group of trees parallel, estimating their effects to each other (shadows, etc.). The result of the simulation is a three dimensional model of the tree(s) at any time of the growth process. The distortion effects of these external factors are well visible on this model, giving a realistic estimation about the integration of the tree(s) into the given environment. The simulation has high computational demand; therefore, the most computationally intensive steps of the simulation are implemented on graphics accelerators using the CUDA framework.

Keywords: Computer Simulation; Computer Graphics; Ray-tracing; GPU programming

1 Introduction

In practice, experts of the fields of forestry and horticulture would like to see the state of a given tree in the next decades. This should be essential for gardeners to design the estimated landscape of an area and for foresters to maximise the profit from the trees. The simulation of tree growth based on the environmental effects can help giving answers for further scientific questions related to climate change, air pollution [1], education [2], and other fields [3].

But the growth of trees is a very complex process because it is affected by several external factors (weather, light conditions, other objects, etc.). Although, every tree can be distinguished by its type traits, but in reality, these can be heavily altered by external forces. After a given tree is rooted in a given place, it has to evolve to withstand a wide range of environmental effects. The main guiding principle of the authors is that to simulate a natural phenomena, the best way is to follow the steps of nature. Nowadays, there are some popular algorithms following this approach, and the method presented in this paper is also based on this principle. It simulates the growth of a given tree from the first day. During this, it takes into account the properties of the given species and the surrounding environment [4].

In some cases, it is not enough to reflect on the static surrounding factors, because it should be also important to take into account the neighbouring trees. In the case of planting multiple trees at once, taking into consideration their effects on each other are also necessary. The presented algorithm is also suitable for this kind of simulations, the final result shows a good estimation of the future shape of all trees.

The main mechanism of the presented model consists of two separate steps:

- The determination of the shape of the tree without any environmental effect.
- Deformation of this shape based on the environmental factors.

It is necessary to repeat this process iteratively every year and recursively for every branch of the tree. After the basic growth step (given by the species properties), the deforming modifiers are applied one by one from a list which can be easily updated or expanded.

The presented model contains the following potential factors to configure the tree growing:

- Type traits
- Heliotropism
- Collisions
- (Self) shades

The main goal of this research is the development of an easily configurable model which creates visualisation of trees in a given environment. That can help professionals and hobbyists to plan their garden or even to choose the correct tree species for populating a new forest.

The rest of this paper is structured as follows: Section 2 contains a state-of-the-art overview of the already existing theoretical methods and implemented systems. The next section shows the methodology of how to simulate the growth of a given tree. Section 4 presents the experimental validity and runtime test results; and finally, the last section contains the conclusions, limitations and further plans.

2 Related Work

The simulated growth of trees is a main task in the field of systems biology and mathematical biology to reproduce plant morphology with computer simulations. There are several related attempts to create realistic tree models [5, 6, 7]. In the past, these were only ray-traced renders; but later, simulations also appear for this purpose. A significant difference should be made between the main approaches of the latter alternatives [8]:

- Developmental models start with a single root element and produce further elements by adding and dividing already existing items of the model. These are very accurate and well usable models with high computational demand. The real-time usage of these is not a real alternative.

- Non-developmental methods are based on prefabricated models. After the insertion of these, the model undergoes a series of necessary modifications and adaptations. This approach has the advantage of very fast response time, but the result is usually less realistic. Nowadays, it is common to use non-developmental models to simulate plants in animations and computer games.

The main objective of this paper is the generate as realistic models as possible (the runtime is less important); therefore, the non-development approach is not acceptable. The growth of a plant is a complex, continuous process altered by several external factors. There are already existing computer simulation based methods to replicate this behaviour with various models. The most commonly used one is the Lindenmayer System model (L-Systems) [9].

The earlier variants of this model handled only the branching structure of the trees in a non-developmental way [10, 11]. The newer variants use several environmental factors to modify this process and create highly detailed developmental models [12] [13].

Based on its recursive approach, it is well applicable for simulation of self-organising objects. The process is based on a simple starting state and a rule set like (1).

$$\begin{aligned} &S\{A,B\} \\ &\alpha\{A\} \rightarrow \{B,B\} \\ &\beta\{B\} \rightarrow \{A\} \end{aligned} \tag{1}$$

After that, the algorithm executes recursive steps applying the given rules in all iterations. Based on this model, it is possible to generate realistic plants.

The original L-Systems implementation contains the following modification factors:

- Effect of gravitation

- Heliotropism

- Geotropism

- Longitudinal and transverse growth
- Collisions
- Rotation of branches
- Angle of branches

Our presented method uses some of the recursive properties of the L-Systems, but makes the model applicable for simulating higher level plants.

The "Structural simulation of tree growth and response" research project [14] had very similar objectives. It presented a mathematical model taking into account the energy consumption of the growth, the weight of branches, and the energy input given by photosynthesis. The developed system gave a good overall estimation of the shape of the tree, but it was not able to handle environmental effects.

It is also worth mentioning the work of Jason Weber and Joseph Penn [5]. Their approach was not simulation based. The shape of the generated tree was described by a simple rule set (shape of the tree, number of branches, subdivision of branches, angle of branches, etc.). According to these rules, it was able to build a tree body made of pyramids and having leaves as surface items. This method mainly focused on visualisation; therefore, it did not take into consideration any environmental effects.

From the visualisation point of view, it is also worth mentioning the already existing advanced modelling and rendering methods, like SpeedTree [15]. These are not simulations, but modelling tools. The end user has to set all necessary properties, and the application can generate a model according to these. Trees generated by this approach are the key components of 3D animations, computer games, and augmented reality applications [16].

Applications from the field of forestry have a very different approach than above. These are detailed simulations supporting the estimation of cost/benefit of tree production. There are very accurate and take into consideration all available environmental factors, but does not have any graphical output.

3 Methodology

3.1 Type Rules

To simulate the most defining traits of a tree, information should be collected from the field of Dendrology. The problem is that this field can help identify trees but not to create them. Therefore, as a preliminary step, we had to find the traits defining the look of a tree. The following properties have been identified:

- Longitudinal growth
- Transverse growth
- Branching angle
- Branching rotation

- Branching type which can be “whorled”, “opposite”, “alternate”, and “spiral”

These properties are not constant in the whole lifetime of a tree. It is necessary to differentiate several life stages as follows:

- 0-1 years: seedling
- 1-5 years: sapling
- 5-15 years: young tree
- 15-70 years: mature tree
- 70-150 years: seed-growing tree

The set of used rules reflects on this by having different parameter values for every life stage. The life stages may vary with different tree types.

As an additional option, different branch levels can also have different parameter values. In this context, branch level refers to the distance from the root. The algorithm is able to handle an arbitrary number of levels, but in practice, it is enough to use three of them:

- 1st level: the trunk of the tree
- 2nd level: branches grown from directly the trunk
- 3rd level: further branches

It is possible to set different parameter values for each branch level.

3.2 Light Detection

Heliotropism is the most decisive modifier because light is the most important resource for every species. Trees can observe light in multiple ways. The simulation uses virtual sensors located in the branch tip buds detecting the direction and amount of light. This information significantly influences the grow direction of this branch. Sensors use Monte Carlo ray-tracing to gain the requested information.

Ray-tracing algorithms are heavily used in three-dimensional graphics. These are used for rendering realistic images by tracing the paths of light as pixels in an image plane. These are capable of producing realistic results, but at the cost of very great computational. Basically, ray-tracing starts beams from one of the pixels of the image (camera) and tries to follow its path. It is able to simulate several optical effects, such as reflection, refraction, scattering, and dispersion. Based on this path, it is possible to determine the colour and intensity of the given pixel. The presented implementation is different from this basic process as it doesn’t produce a flat image, but explores the directions where light beams arrive at the virtual light sensors.

The special Monte Carlo ray-tracing [17] algorithm is used to simulate this behaviour of trees. In general, Monte Carlo algorithms are randomised procedures to estimate the value of very time-consuming computations. These algorithms are based on repeated randomised sampling, and they produce an output which might be incorrect with a certain probability. This probability can be reduced using a higher number of random samples.

As an optimisation step, the presented light detection algorithm first looks for a direct line to any light source. If any of these exists, the dominant direction of light is from this point. If there are no such direct lines, then it starts the path tracing. During this, a beam of ray is started into a random direction. If it hits a triangle (all objects in the model space are described by the triangles forming their surface) it continues to a direction according to the rules of reflection, and repeats this process for a given number of times or until it reaches a light source. After that, it is able to start further beams, and at the end of this process, it results in an array of vectors representing the directions of potential light sources.

Based on this array, the light source with maximum energy is selected (the dominant light source direction). This will modify the direction of the branch growth leading to that light source.

3.3 Collision Detection

The growth of a tree is significantly influenced by the objects in its environment. It is obvious that a branch cannot move across or into any solid obstacles. The simulation is executed in a three-dimensional model space; therefore, it is possible to create and place any additional objects into this. There are no limits to the shape and size of these objects. Furthermore, the neighbouring trees can also be considered as obstacles too.

The collision detection is another part of the iterative growing process. It is done for every branch after the altered growing direction and length is determined. The potential new branch interval is checked against every triangle in its neighbourhood. If there are no collisions, the potential branch becomes real. If it collides with any of the triangles, it is necessary to alter the direction of growth.

This new direction is based on the projection of the growth vector to the plane of the triangle. This projection gives a new growth direction if it does not cause further collisions. This method has some limitations, for example, if the growth vector is perpendicular to the surface of the triangle then the projection is a point which stops the growing process of the given branch.

3.4 Pruning

It is possible to dynamically modify the structure of the tree. Selected branches and their sub-branches can be removed from the model. This lets more light for other branches, modifying their behaviour. Using this technique, it is possible to give more realistic results.

There are also some automatic pruning mechanisms. In the case of some species, branches without enough light become inactive and withered. These are usually the lower/inner branches of a tree, especially in a multi-tree environment. It is possible to automatically remove these parts.

3.5 Simulating Multiple Trees

It is possible to simulate not just one but multiple trees parallel. These are simulated year by year, sequentially one after the another. This may cause some problems with the trees effecting to each other in an unnatural way. For example, trees at the beginning of the iteration can grow over the others blocking more light from them. The prevention of this phenomena is that shades are calculated and updated at the start of every iteration.

The partially separated simulations of trees have several benefits. There are no limits to the number, type and age of the trees in the same model space. It is possible to plant different species at different times and run the simulation. The result will show a good estimation of the whole group.

3.6 Acceleration with GPUs

3.6.1 GPU Acceleration of Ray-Tracing in General

The well-known disadvantage of ray-tracing methods is the very high computational demand compared to other three-dimensional rendering techniques [18]. Tracking the light beams from their source to the final destination needs several mathematical calculations like reflection angles, collision projections, etc. This is the reason, why the accepted view is that this method is not applicable for real-time purposes.

Using the Monte Carlo method makes it possible to significantly decrease the number of these tracings, but it is also worth considering that more rays usually gives more accurate results. This means thousands of rays from one sensor. In the first years, this is not an issue, but as the tree becomes older, the number of branches increases exponentially. On an average CPU, it becomes days to simulate the changes for further one-year iterations, which is unacceptable.

Fortunately, ray-tracing is a typical embarrassingly parallel algorithm. Thanks to this, there are several parallel implementations and it is obvious that it is suitable for data-parallel implementations. Because all pixels of the target image are independent of each other, it is possible to fully parallelise the process. GPU implementations are usually based on the idea that it is possible to assign each pixel (camera ray) to one thread of the GPU. This means thousands or millions of threads, but this is what the GPU for [19, 20], the GPU based implementations can achieve significant speedup over the serialised version.

3.6.2 GPU Acceleration of Tree Growth Simulation

The novel GPU accelerated tracing method works mostly the same as the presented regular implementation. The major difference is in the pluralisation phase. Ray-tracing algorithms are easy to parallelise because these are usually completely data parallel. This means that each photon paths are independently tracked from each other and the branch traces are also separate from each other. This helps the implementation, because it is possible to handle the branches in CUDA blocks where the paths are handled by independent threads within those blocks.

The other difference is that the CPU version calculates every branch sequentially, but the GPU version calculates these at once at the beginning of the iteration. This needs the following consecutive steps:

1. Before the calculation, all data required for the ray-tracing process is copied to the device from the host.
2. GPU kernels run and calculate the results.
3. The results should be copied back to the host from the GPU.

This data transfer is very time-consuming because it uses the standard PCI-e bus. It is a hardware limitation, therefore there is no way to significantly decrease the requested time.

This causes the behaviour presented in the evaluation part that in the first few years of the simulation, the GPU version is slower than the CPU because the amount of data copied is large compared to the number of calculations (where the GPU can offset this disadvantage).

4 Evaluation

4.1 Validation

4.1.1 Validation of Technical Sub-steps

It is always hard to evaluate and validate the results of a nature-inspired simulation, because there are no gold standards to compare with. Because of the several randomised factors, it is also not possible to give an exact expected state from a given initial state (environmental and growth parameters).

The best thing to to do is checking the validity of the presented technical sub-steps and do visual examination on the simulation results compared to similar real-world examples.

Fig. 1 shows the result of the validation of the growth sub-step. All presented sub-steps are validated in a similar way (not detailed in this paper).

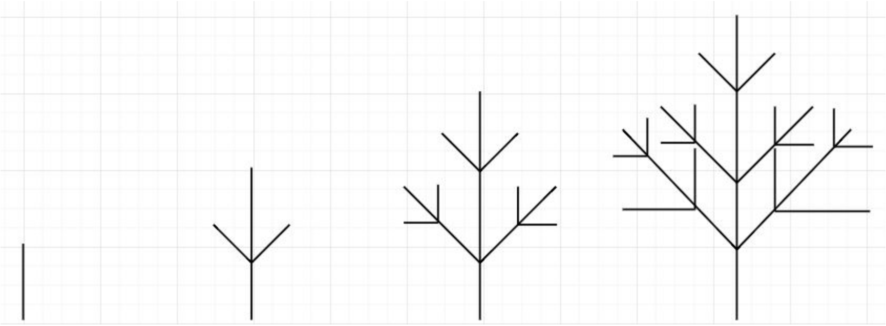


Figure 1
Validation of growing in the first 3 years.

4.1.2 Basic Tree without any Environmental Effects

Fig. 2 shows the result of a simulation of a basic tree without any environmental effects. The length and number of branches conform the given simulation parameters. As visible, it has the expected regular and symmetrical shape. It is also possible to display the state of the tree at the end of every simulation year (Fig. 3).

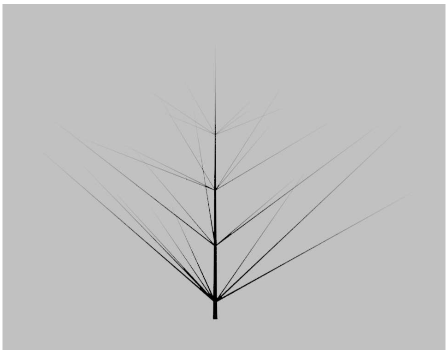


Figure 2
Tree without any environmental effects.

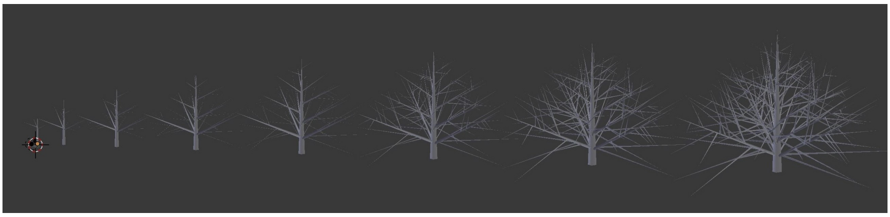


Figure 3
8 years long simulation without environmental effects.

4.1.3 Light Detection

Adding some light sources to the model makes it possible to validate the effect of light tracking. Fig. 4 shows a tree grown with an external light source (from the upper left corner of the image). As expected, the tree is blended toward the light source.

4.1.4 Collision Detection

To validate the collision detector, it is possible to place an additional obstacle object into the model space. Fig. 5 shows some examples for simulations of trees grown near one or more external object(s). As visible, all trees were growing straight upwards in the first years (there were no external light sources to modify this behaviour). After that, some of their branches could not grow to the desired direction. According to the expectations, they had to find another direction to grow after the

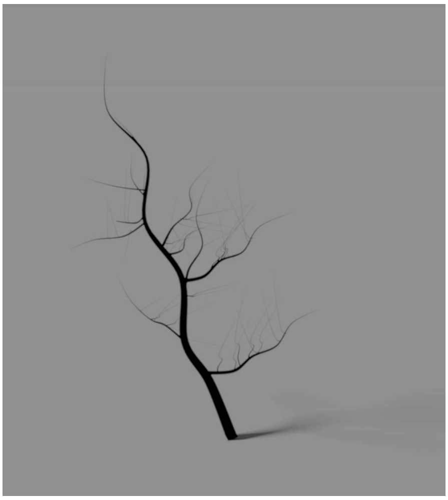


Figure 4
Tree with light effect.

collision. Overcoming the obstacles, they continued the growing process straight upwards.

4.1.5 Complex Examples

To validate the cooperation of modifiers, Fig. 6 shows a more complex example. In the first few years, the tree cannot see any direct light source, therefore it started growing towards the left inner side of the box, where some light reflection comes. But when is becomes taller, it turns towards the direction of the light source. As visible, it took care about the collisions and found the hole in the top of the box.

To summarise the effects of these modifiers, Fig. 7a shows the result of a 5 year long simulation. The parameters of teak trees were used. To help the visual validation, Fig. 7b presents a real-world 5 years old teak tree. As expected, the main visual attributes of these are very similar.

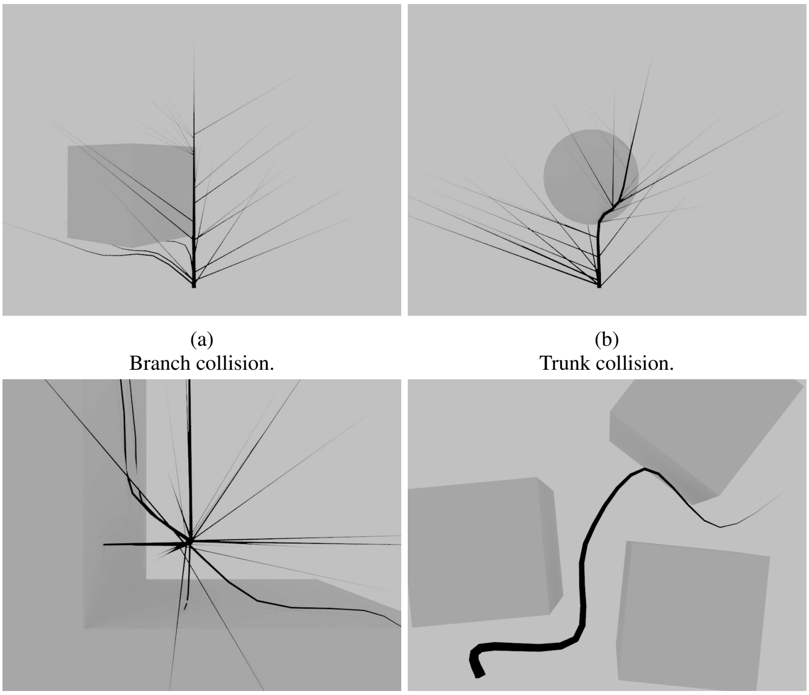
As a final test, Fig. 8 shows the result of a simulation on multiple trees. The rear one was deployed some years before the others, therefore it is larger.

4.2 Runtime

As mentioned, the tree growing simulation has very high computational demand. The most time-consuming part of the process is the ray-tracing step, which determines the direction and strength of dominant light in a given point. An efficient GPU based ray-tracing algorithm has been designed and implemented to speed-up this step.

The following hardware configurations were used for benchmarking:

- CPU configuration



(a)
Branch collision.

(b)
Trunk collision.

(c)
Upper view of collision.

(d)
Complex use case for multiple obstacles.

Figure 5
Validation of the collision detector.

  - CPU: Intel Core i5-2500K
  - Number of cores: 4
  - RAM: 8GB DDR3
  - TDP: 95W
- GPU configuration
  - GPU: NVIDIA GeForce GTX 1070
  - Number of CUDA cores: 1920
  - RAM: 8GB DDR5
  - MPC: 150W

Fig. 9 shows the runtime of the CPU and the GPU implementation of the ray-tracing algorithm for a one year simulation period. In the first few years, the tree is not enough complex to fully utilise all the available cores of the GPU. As a consequence, the execution time of the CPU implementation is smaller in the case of small (young) trees. But it is also visible, that the runtime increases faster in the

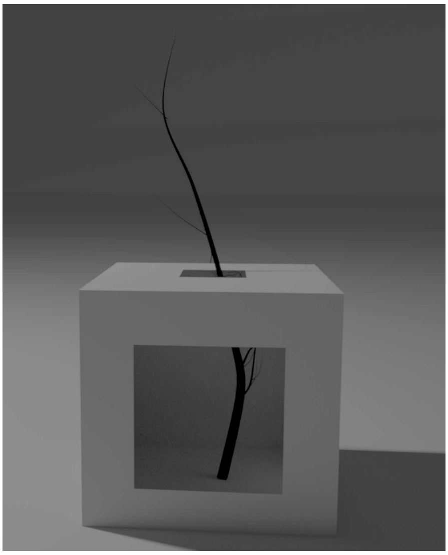


Figure 6
A complex example for light and collision detection.

case of the CPU. The measured runtimes are similar for 6 year old trees, and after that, the GPU clearly outperforms the CPU.

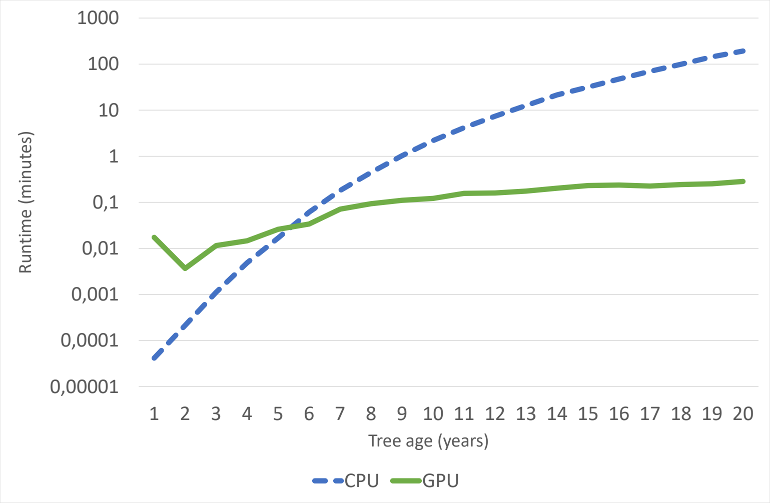
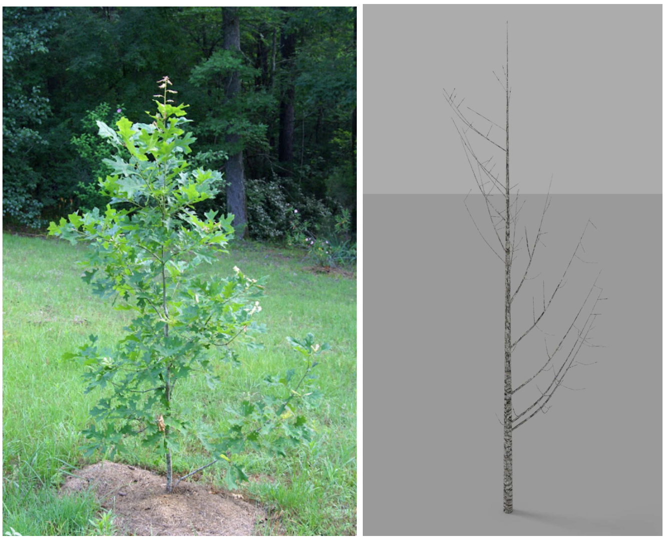


Figure 9
Year to year growth time



(a)
Real 5 year old teak tree.

(b)
Simulated 5 year old teak tree.

Figure 7
Comparison of simulated and real tree.

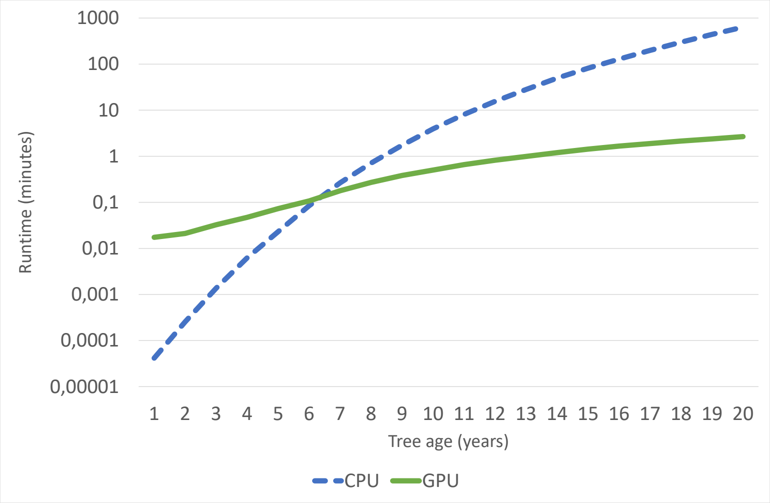


Figure 10
Total growth time

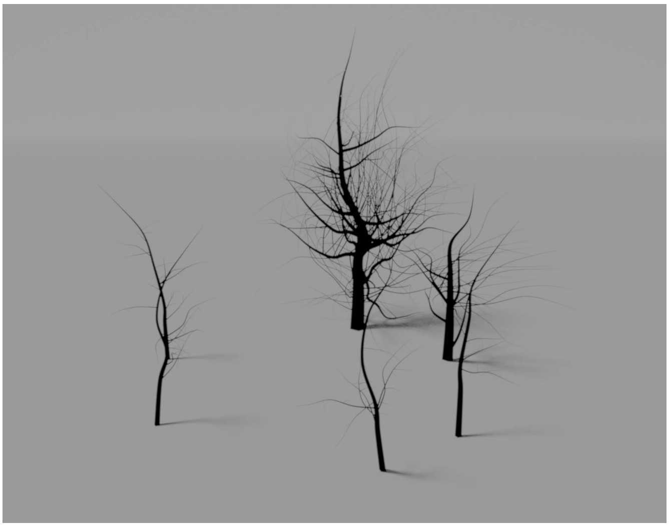


Figure 8
Simulating the growth of multiple trees.

The results of Fig. 10 are more interesting. These show the accumulated runtimes of the previous years, which are more important from the practical point of view. Obviously, the runtime of the CPU implementation is smaller for small trees. This changes after 7 years, when the overall runtime of the GPU becomes better. The runtime of both algorithms is exponential to the age of the tree (according to the exponential growth of branch count), but the runtime of the CPU implementation rises more steeply.

It was not necessary to compare the accuracy of the CPU and the GPU solution, because the underlying algorithms are the same. Therefore, the trees generated by the CPU and GPU are exactly the same. These share the same results regarding the validation steps.

In practice, only long-term simulations give valuable results. As a consequence, it is worth using the GPU implementation. It may worth considering to implement a hybrid approach which uses the CPU in the case of small trees and switch to the GPU for larger ones.

5 Conclusions

This paper presents a novel nature-inspired tree growth simulation algorithm to estimate the future states (shape, size, etc.) of a given tree according to the environmental parameters (light, obstacles, neighbouring trees, etc.), and its species characteristics (annual number of new branches; angle, length of branches, etc.).

It presents a novel method of simulating the growth of plants which is not entirely based on L-Systems. The main difference is that it takes a less direct approach as rule based generation. The environmental effects are not part of the growth rule system but a separate subsystem which modifies the result of the growth model. This leads to a dynamically and easily extendable model.

The validation section shows that this novel method can efficiently estimate the future characteristics of a given tree. It was able to give not just quantitative information about the tree, but also a complete three-dimensional model. It is possible to preview the states during the simulation at the end of each year.

Profiling showed that the most time-consuming part of the algorithm is the ray-tracing sub-process. To speed-up this part, a novel GPU based ray tracking algorithm was developed and implemented using the CUDA framework. Benchmarks show that this was slower than the CPU implementation in the case of small (young) trees, but it was significantly faster in the case of large (old) ones.

Unlike the already existing implementations, the presented system can simulate not just one but multiple trees at once (where each of these can be different species). Benchmarks show that the lack of hardware resources should be the main limit for this kind of simulations.

Thanks to the modular design of the application, the set of deforming factors of the simulation can be easily extended. As further plans, authors would like to implement the following modifiers:

- Regionalism by altitude
- Water consumption and need
- Soil quality
- Geotropism

Acknowledgements

The authors would like to say thanks for the help of the Hungarian National Talent Program (NTP-HHTDK-18-0031 and NTP-SZKOLL-18-0016). Sándor Szénási would like to thank EPAM Systems for their financial support. The authors would like to thank NVIDIA Corporation for providing graphics hardware through the CUDA Teaching Center program. The authors also wish to thank the members of the CUDA Teaching Center at Óbuda University for their constructive comments and suggestions. The content of the paper, however, does not necessarily express the views of these companies and persons, the authors take full responsibility for any errors or omissions.

References

[1] D. Stojcsics, Z. Domozi, and A. Molnar, “Air pollution localisation based on uav survey,” in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 2546–2551.

[2] K. Czakóová *et al.*, “Microworld environment of small language as, living laboratory” for developing educational games and applications,” in *Conference*

proceedings of eLearning and Software for Education (eLSE), vol. 1, no. 01, 2017, pp. 285–291.

[3] D. Kiss, "A model of heat exchange and accumulation in small-sized bioreactors during ethanol fermentation," in *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, 2017, pp. 000 313–000 316.

[4] D. Chamovitz, *What a plant knows: a field guide to the senses*. Scientific American/Farrar, Straus and Giroux, 2012.

[5] J. Weber and J. Penn, "Creation and rendering of realistic trees," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. ACM, 1995, pp. 119–128.

[6] P. Horácek, "Introduction to tree statics & static assessment," in *Tree statics and dynamics seminar, interpreting the significance of factors affecting tree structure & health, Westonbirt, UK*, 2003.

[7] K. Onishi, S. Hasuike, Y. Kitamura, and F. Kishino, "Interactive modeling of trees by using growth simulation," in *Proceedings of the ACM symposium on Virtual reality software and technology*. ACM, 2003, pp. 66–72.

[8] P. L. Jaworski, "Using simulations and artificial life algorithms to grow elements of construction," Ph.D. dissertation, UCL (University College London), 2006.

[9] A. Lindenmayer, "Mathematical models for cellular interactions in development i. filaments with one-sided inputs," *Journal of Theoretical Biology*, vol. 18, no. 3, pp. 280 – 299, 1968. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022519368900799

[10] C. Jirasek and P. Prusinkiewicz, "A biomechanical model of branch shape in plants," in *Proceedings of the western computer graphics symposium, Whistler, Canada*. Citeseer, 1998, pp. 23–26.

[11] C. A. Jirasek, *A biomechanical model of branch shape in plants expressed using L-systems*. Calgary, 2000.

[12] J.-F. Barczi, H. Rey, Y. Caraglio, P. de Reffye, D. Barthélémy, Q. X. Dong, and T. Fourcaud, "A structural whole-plant simulator based on botanical knowledge and designed to host external functional models," *Annals of Botany*, vol. 101, pp. 1125–1138, 2008.

[13] F. Tian-shuanga, I. Yi-binga, and S. Dong-xu, "Tree modeling and dynamics simulation," *Physics Procedia*, vol. 33, pp. 1710–1716, 2012.

[14] J. C. Hart, B. Baker, and J. Michaelraj, "Structural simulation of tree growth and response," *The Visual Computer*, vol. 19, no. 2, pp. 151–163, 2003.

[15] (2018) speedtree main page. [Online]. Available: https://store.speedtree.com/

[16] G. Molnár, Z. Szűts, and K. Biró, "Use of augmented reality in learning," *Acta Polytechnica Hungarica*, vol. 15, no. 5, 2018.

[17] H. W. Jensen, J. Arvo, P. Dutre, A. Keller, A. Owen, M. Pharr, and P. Shirley, "Monte carlo ray tracing," in *ACM SIGGRAPH*, 2003, pp. 27–31.

[18] G. Kertész and Z. Vámossy, "Current challenges in multi-view computer vision," in *2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics*. IEEE, 2015, pp. 237–241.

[19] P. Szántó and B. Fehér, "Hierarchical histogram-based median filter for gpus," *Acta Polytechnica Hungarica*, vol. 15, no. 2, 2018.

[20] V. Marković and Z. Konjović, "A contribution to software development quality management," *Acta Polytechnica Hungarica*, vol. 14, no. 8, 2017.