

Analytical Model for the Material Flow during Cold Rolling

János György Bátorfi^{1,2*}, Mátyás Andó¹, Jurij J. Sidor¹

¹Savaria Institute of Technology, Faculty of Informatics, Eötvös Loránd University, Károlyi Gáspár tér 4, H-9700 Szombathely, Hungary, bj@inf.elte.hu, am@inf.elte.hu, js@inf.elte.hu

²Doctoral School of Physics, Faculty of Natural Sciences, Eötvös Loránd University, Pázmány Péter sétány 1/A, H-1117 Budapest, Hungary, bj@inf.elte.hu

*Corresponding author, e-mail: bj@inf.elte.hu

Abstract: An analytical description of conventional cold rolling, is developed with a novel mathematical algorithm, based on finite element (FEM) and flow-line (FLM) method calculations. A new function, enabling the formulation of deformation, accurately describes the possible reverse displacement near the surface of a rolled sheet. The corresponding value of deformation can be determined for various friction coefficients and roll gap geometries by employing the expressions developed. Knowing the material flow, the amount of deformation and stress distribution along the thickness of the rolled sheet can be calculated. The results obtained were compared to the counterparts computed by FEM and FLM. It was shown that the extended model ensures accurate description of the material flow for small thickness reductions and low friction coefficients, where the phenomena of reverse displacement are observed, and many numerical approaches fail to capture this type of deformation pattern. The model was tested on both experimentally measured results and data obtained from various literature sources.

Keywords: Cold rolling; Symmetric rolling; Shear deformation; Reverse displacement; FEM; FLM

1 Introduction

Properties of the conventionally produced flat-rolled products are strongly affected by the following technological parameters: angular velocity and diameter of rolls, friction coefficient, yield stress of materials, reduction and deformation temperature. In view of the complexity of rolling process, the material flow is generally examined by experimental measurements [1] [2], finite element modeling (FEM) [3], flow-line models (FLM) [3-8] or other analytical methods [9] [10].

In most general case, the simplest approximation called plane strain compression (PSC) is employed which considers only the normal component of deformation, while the contribution of friction with corresponding shear components is neglected. Here, the amount of strain component in the thickness direction is approximated by the following expression, this component can be called the reduction too:

$$r = \frac{h_0 - h}{h_0} \quad (1)$$

where h_0 and h are initial and final thicknesses of a sheet subjected to rolling.

This simple approach does not account for accurate estimation of equivalent strain, rolling force and torque, whereas the reasonable estimate of the technological parameters can be done by FEM, where the properties of a material, friction conditions and parameters of a roll gap should be set. The well-established FEM models are based on the theory of plasticity with specified material parameters and can be used for simulating numerous mechanical problems. In the simulation, the material subjected to deformation is subdivided to numerous volume elements by a network of points whereas the stiffness equation system is imposed to specified boundary conditions, which define the displacement of a given group of points. Knowing the deformation of each element, both strains and stresses can be calculated, which makes possible the evaluation of these quantities in diverse planes and directions. Application of this numerical approach requires significant computational capacity. The calculation time depends on the material model used, boundary conditions imposed, type and size of mesh generated, and a number of steps defined. For instance, simulation of rolling process by employing a very fine mesh and nonlinear material model can take approximately 2 weeks or even longer, depending on the capacity of the personal computer used.

In contrast to FEM, flow line models (FLM) [3-8] offer fast and relatively accurate analytical solutions for specific deformation processes such as sheet rolling, bending or extrusion. Employing FLM approximations requires the definition of model parameters, which might be related to the boundary conditions of deformation process, however, in many instances, the fitting parameters do not reveal physical meaning, which complicates the implementation of this numerical approach in industrial practice. In case of rolling, FLMs engage streamlines which enforces the material to flow along these predefined directions with specific velocity, whereas the heterogeneity of strain/stress distribution is predefined by model parameters. For instance, in the model of Decroos *et al.* [7], the model parameters α and n ensures diverse deformation velocities along various streamlines and as it turned out both values are functions of friction coefficient and roll gap geometry and can be defined as it is described in Ref. [6]. In many instances, the FLM [7] ensures results comparable to FEM [5-8]. In addition to FLMs, alternative approximations [9, 10] provide accurate solutions exclusively for small values of friction coefficient.

It should be noted that the friction coefficient of Coulomb model (μ) is not constant in cold rolling, and the temperature of material changes either but the temperature variation has a negligible effect on material flow. In cold rolling of Al alloys, an increase of temperature in one deformation pass is far below one needed for recovery or recrystallization, however, the smallest changes in friction tend to induce strong strain/stress heterogeneities across the thickness of rolled sheet, which in turn has a strong influence on recrystallization phenomena during the subsequent annealing process. In this view, the determination of μ is of crucial importance. It is generally known that rolling is possible if μ exceeds the minimum value μ_{\min} necessary for the process to be completed [11]:

$$\mu_{\min} = \frac{1}{2} \sqrt{\frac{h}{R}} \frac{\ln\left(\frac{h_0}{h}\right) + \frac{1}{4} \sqrt{\frac{h_0 - h}{R}}}{\tan^{-1} \sqrt{\frac{h_0}{h} - 1}} \quad (2)$$

where R is a roll radius.

There are other literatures, such as the one described in [14], which also takes horizontal forces into account.

To evaluate the strain in rolling, Inoue [15] employs three basic assumptions: 1) the value of shear does not grow linearly; 2) the shear strain can decrease after reaching neutral point; 3) there is a plane strain state in the sheet and the normal strain is uniform through the thickness. Apart from mentioned approximation, there are many literature sources dealing with the strain heterogeneities which evolve in rolled materials across the thickness [5] [12] [13], however, there are still many issues which are not entirely understood or cannot be captured by numerical methods.

There are many other modelling methods, like the “Genetic Algorithm” based on “Artificial Neural Network” [14].

In this contribution, we present a new mathematical formulation allowing fast and accurate estimation of strain/stress heterogeneities.

2 Computational Procedure

The evolution of strain in rolling was investigated by FEM simulations. Since cold rolling does not account for widening, the calculations were performed with Deform 2D© software. Cold rolling is a symmetric process and therefore, the boundary conditions are defined to the rolls and symmetry line. All simulations were carried out with the constant friction coefficient for the Coulomb model, the

value of which exceeds μ_{\min} . This minimal value is changing between 0.015 and 0.05. The behavior of a material is described by the plastic-multilinear model with a constant Young's modulus of 68.9 GPa and Poisson's ratio of 0.33. The plastic properties are the same as applied in [5] and as it is described later in Eq. (15). The mesh of a rolled sheet is divided into 50x11 elements, while half of the thickness is separated into 10 layers. In the calculation procedure, 59 different parameter set were used, and 11 points were taken along the half-thickness of the sheet, so the mathematical function introduced below was fitted to the results of 649 points. The amount of reduction on the material flow was analyzed by changing the degree of deformation and friction conditions. The applied geometrical values are the following: 250 mm is the radius of the roll, the velocity is 2 m/min, the initial half-thickness is 2 mm, the final thickness is changing between 0.6 and 0.9 mm, the friction coefficient is changing between 0.025 and 0.25.

3 New Mathematical Formulation of Rolling

As Fig. (1) shows, the material flow in rolling, revealing diverse patterns, is strongly correlated to the roll gap geometry. It is obvious that the distorted patterns of Fig. (1) can be successfully described by analytical approximations where the displacement is approximated by a quadratic function in the horizontal direction, Eq. (3). This function is capable of describing the experimental patterns as well [1]. Analyzing the experimentally observed [1] displacement of initially vertical line, it becomes clear that the stress/strain state during rolling can be considered as a plane strain one.

$$x = Az^2 \quad (3)$$

where x and z are parallel to rolling (x) and normal (z) directions, respectively, and A is a constant.

The precise description of deformation is important since it allows understanding the evolution of crystallographic texture in rolled materials [3-7]. Knowing the variations of texture evolved enables the evaluation of mechanical properties and their anisotropy.

In most general case, the distortion of the initially rectangular grid can be described by a simple analytical expression, Eq. (4). This type of displacement function is predicted by the flow line models, FEM and likewise observed experimentally [6] [7] [15]. It turns out that the model parameters α and n are functions of roll gap geometry and friction coefficient [6] [7].

$$x = \alpha z^n \quad (4)$$

It should be noted that mathematical expressions of Eqs. (3) and (4) describes the displacement fields with relatively good accuracy, however, a new exponential function Eq. (5) was introduced for more complex shaped curves, since the function presented by Eq. (4), is not capable of characterizing the phenomenon of reverse displacement, described in detail below. In some instances, [6] [16] [17], the maximal shear strain is localized in the subsurface layers and this phenomenon cannot be captured by a simple polynomial expression. Fig. (2) shows several examples of deformation patterns which can be reproduced by the extended function.

$$x = B_1(e^{-B_2z^2} - 1) + B_3z^2 \quad (5)$$

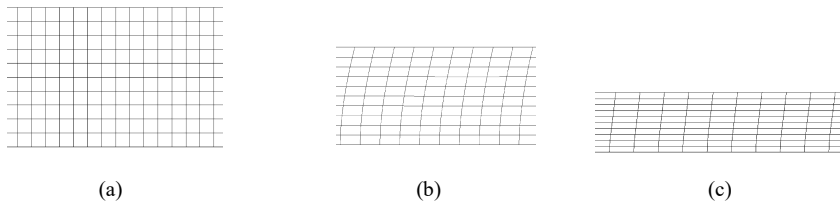


Figure 1

Distortion of mesh used in FEM simulations: a) initial mesh; b) mesh after 30% reduction with friction coefficient of 0.08; c) mesh after 50% reduction with friction coefficient of 0.08 (the initial thickness of Al sheet is 2 mm while the roll diameter is 250 mm; only the half-thickness is revealed due to symmetry imposed by rolling)

This complex equation is particularly advantageous in the cases when the reverse displacement occurs near the surface of the sheet leading to a strong deviation from the parabolic curve.

The shape of curve expressed by Eq. (5) can be controlled by varying the model parameters B_1 , B_2 and B_3 . As Fig. (2) shows, in some instances the maximum displacement is observed not on the surface of a rolled sheet but in the subsurface region. These types of curves are typically observed in FEM simulations, whereas the model parameters (B_1 , B_2 and B_3) can easily be determined for each particular case by fitting procedure. Comparing Eqs. (3) and (5), it becomes obvious that the simplified deformation model (SDM) can reproduce the quasi parabolic function of Eq. (3) if $B_1=0$ and $A=B_3$. Since the material flow is controlled by the roll gap geometry and friction coefficient, it is reasonable to suggest the model parameters will also be. The SDM model use the diameter of the roll, the initial thickness, the final thickness and the friction coefficient as inputs for calculating different parameters of the model.

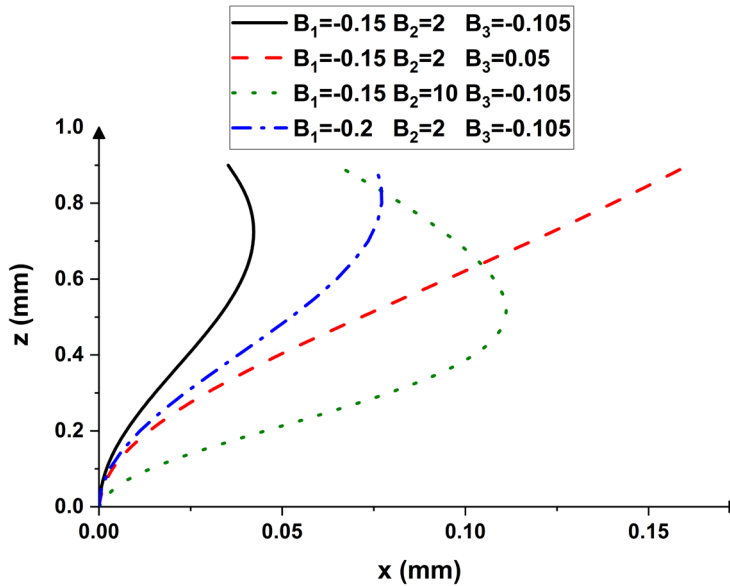
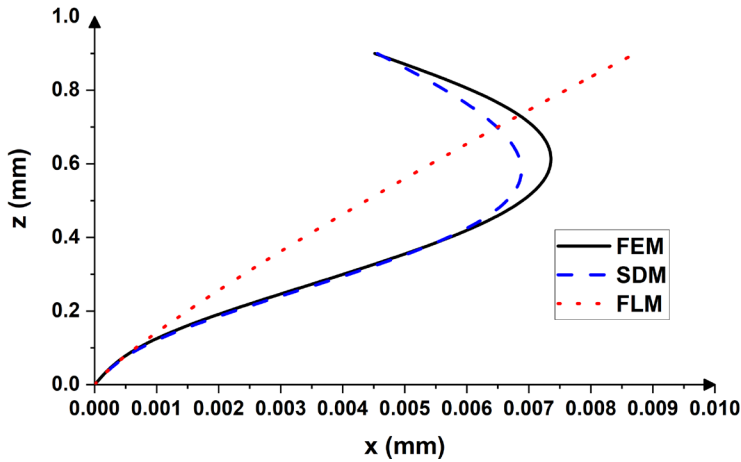


Figure 2

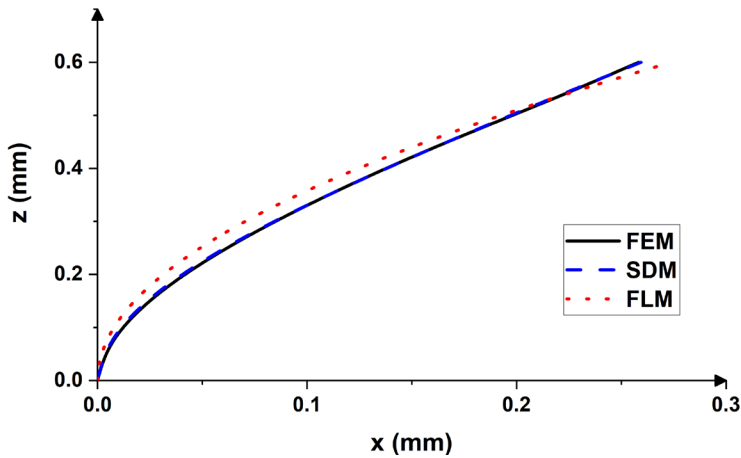
Distortion of initially vertical line as predicted by the extended function with specified model parameters

4 Results

Fig. (3) reveals the two ends of the deformation spectrum: A) small reduction with a small value of friction coefficient ($R=250$ mm, $h_0=1$ mm, $h=0.9$ mm, $r=0.1$, $\mu=0.08$, $v=2$ m/min) and B) relatively large straining with the larger value of friction coefficient ($R=250$ mm, $h_0=1$ mm, $h=0.6$ mm, $r=0.4$, $\mu=0.25$, $v=2$ m/min). In both cases, the linear velocity of rolls was identical $v=2$ m/min, this means, that this rolling process is a symmetric rolling. As it is shown in Fig (3a), the application of a small reduction degree with low values of μ accounts for reverse displacement and this phenomenon can be captured by the extended function (SDM, Eq. (5)), while the flow line model [7] fails to reproduce this deformation pattern. In the case of relatively large straining with larger μ , both the FLM and SDM can successfully reproduce the displacement curve calculated by FEM. It can be concluded here that the reverse displacement tends to vanish by increasing the degree of deformation.



(a)



(b)

Figure 3

Displacement patterns calculated by FEM, FLM [7] and SDM for various roll gap geometries: a) $R=250$ mm, $h_0=1$ mm, $h=0.9$ mm, $r=0.1$, $\mu=0.08$, $v=2$ m/min; b) $R=250$ mm, $h_0=1$ mm, $h=0.6$ mm, $r=0.4$, $\mu=0.25$, $v=2$ m/min

4.1 Determining the Model Parameters

In order to make the model practically attainable, the model parameters (B_1 , B_2 and B_3) of Eq. (5) should be determined. It is suggested here by Eqs. (6-8) that parameters B_1 - B_3 are functions of friction coefficient μ while the polynomial coefficients p_{ij} depend on the reduction degree r . In Eq. (9), the s_{ijk} coefficients are fitting parameters.

$$B_i = p_{i3} \cdot \mu^3 + p_{i2} \cdot \mu^2 + p_{i1} \cdot \mu + p_{i0} \quad (6)$$

$$p_{ij} = s_{ij3} \cdot r^3 + s_{ij2} \cdot r^2 + s_{ij1} \cdot r + s_{ij0} \quad (7)$$

To determine the s_{ijk} , the square sum of the difference between the SDM patterns and ones calculated by FEM need to be minimalized according to Eq. (8):

$$f(\mu, r, z, \bar{s}) = \sum_{\mu} \sum_{\bar{s}} [X_{\text{appr}}(\mu, r, z) - X_{\text{FEM}}]^2 \quad (8)$$

The minimum can be determined by the multivariable extreme value defined in Eq. (9):

$$\frac{\partial f(\mu, r, z, \bar{s})}{\partial \bar{s}} = \bar{0} \quad (9)$$

Due to a large number of variables, a system of 12 equations with 12 variables was created. The solution for this equation-system was found by the Nonlinear programming method. The s_{ijk} parameters, presented in Table 1 were defined from the results presented in [6] [8] [9]. It should be underlined that the parameters of Table 1 are applicable for aluminum alloys, while the same procedure can be repeated for other metals as well.

Table 1
Parameters fitted for Eq. (5)

	S _{xx3}	S _{xx2}	S _{xx1}	S _{xx0}
S _{13x}	-4.78	315.92	-85.11	0.58
S _{12x}	-8	-381.3	154.26	-12.97
S _{11x}	-23.94	99.61	-38.84	3.35
S _{10x}	-2.58	-2.53	1.4	-0.15
S _{23x}	-0.01	237042	-131188.6	15426.6
S _{22x}	-0.21	-133149.5	71118.7	-7884.67
S _{21x}	-0.91	22114.4	-11439.6	1186.25
S _{20x}	3.94	-1009.74	514.72	-44.56
S _{33x}	6.27	5.04	0.35	-15.24
S _{32x}	19.32	-391.02	172.8	-11.45
S _{31x}	37.88	101.43	-44.28	3.67
S _{30x}	-6.36	-2.29	1.6	-0.17

Knowing the s_{ijk} values enables the calculation of deformation distribution across the thickness of rolled sheets under various conditions. Fig. (4) shows the displacement patterns of initially vertical lines calculated by FEM and approximated by the simplified mathematical model of Eq. (5) with s_{ijk} coefficients of Table 1 for various thickness reductions and diverse friction

conditions. It is evident that the mathematical model employed provides very satisfactory results since the deviations between the FEM and SDM curves are negligibly small. It is obvious that the phenomenon of reverse displacement can be neglect for rolling under relatively dry condition (higher μ).

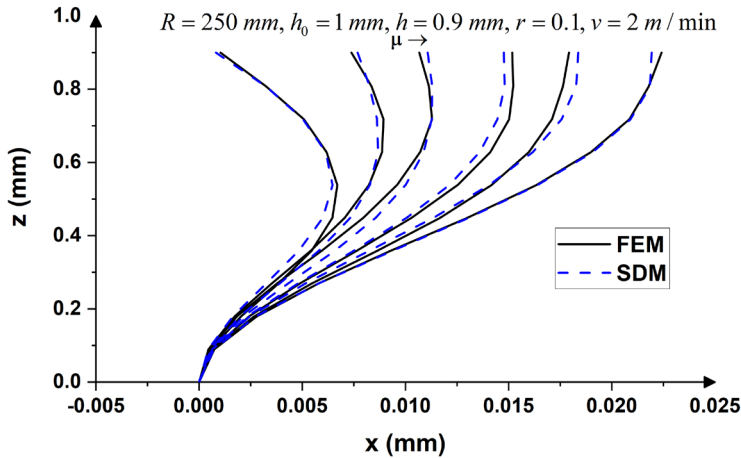


Figure 4

Displacement curves of initially vertical lines calculated by FEM (continuous line) and simplified mathematical model (Eq. (5), dashed lines) for thickness reduction $R=250$ mm, $h_0=1$ mm, $h=0.9$ mm, $r=0.1$, $\mu=(0.075, 0.1, 0.15, 0.175, 0.2, 0.25)$, $v=2$ m/min

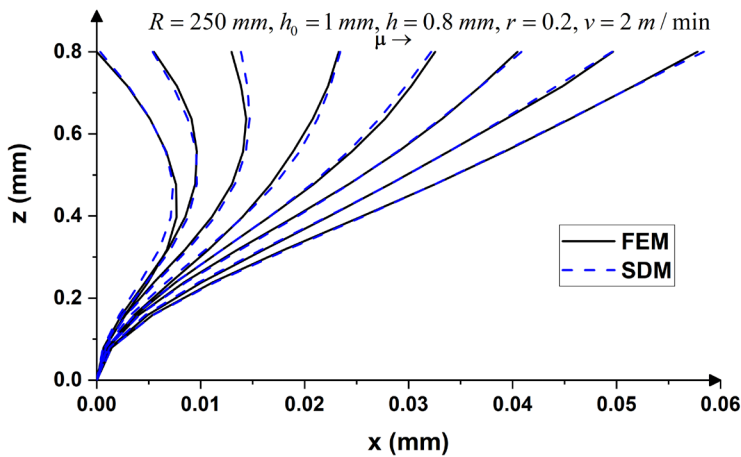


Figure 5

Displacement curves of initially vertical lines calculated by FEM (continuous line) and simplified mathematical model (Eq. (5), dashed lines) for thickness reduction $R=250$ mm, $h_0=1$ mm, $h=0.8$ mm, $r=0.2$, $\mu=(0.025, 0.06, 0.1375, 0.15, 0.175, 0.2, 0.225, 0.25)$, $v=2$ m/min

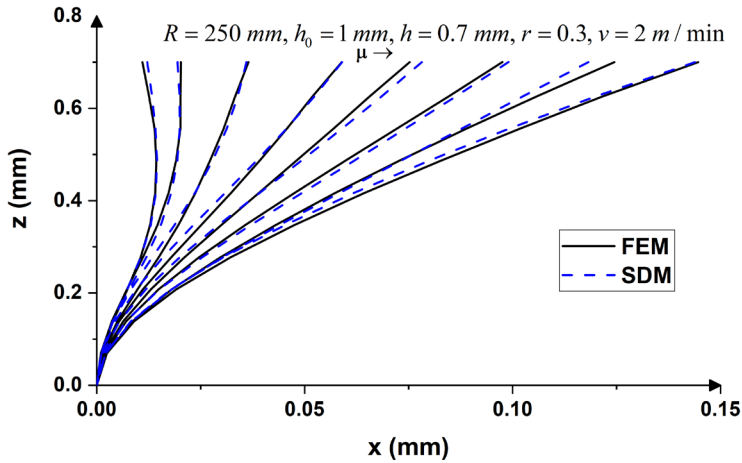


Figure 6

Displacement curves of initially vertical lines calculated by FEM (continuous line) and simplified mathematical model (Eq. (5), dashed lines) for thickness reduction $R=250$ mm, $h_0=1$ mm, $h=0.7$ mm, $r=0.3$, $\mu=(0.08, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25)$, $v=2$ m/min

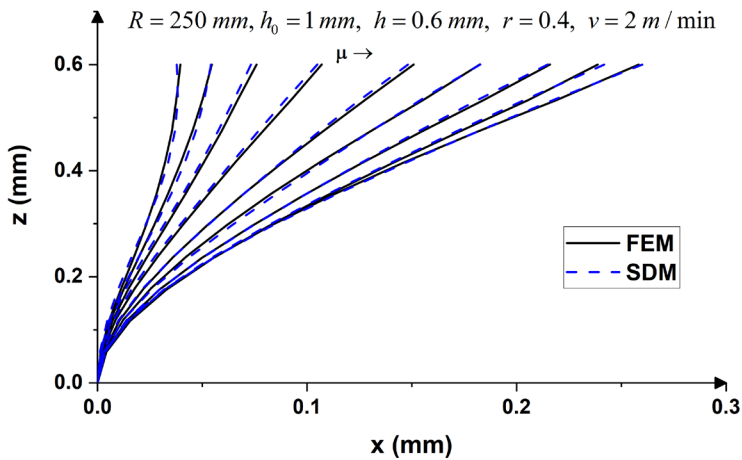


Figure 7

Displacement curves of initially vertical lines calculated by FEM (continuous line) and simplified mathematical model (Eq. (5), dashed lines) for thickness reduction $R=250$ mm, $h_0=1$ mm, $h=0.6$ mm, $r=0.4$, $\mu=(0.075, 0.9, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25)$, $v=2$ m/min

To calculate the equivalent (von Mises) strain, it is necessary to estimate the deformation in both horizontal and vertical directions, which are typically considered uniform along the cross-section of a rolled sheet. The normal and shear components of deformation as well as the equivalent strain [13] can be computed according to Eqs. (10-13):

$$\varepsilon_{zz}(z) = -\varepsilon_{xx}(z) = -\ln\left(\frac{h_0}{h}\right) \quad (10)$$

$$\gamma_{zx}(z) = \gamma_{xz}(z) = \gamma(z) = \frac{dx(z)}{dz} = -2B_1 \cdot B_2 \cdot ze^{-B_2 z^2} + 2B_3 \cdot z \quad (11)$$

$$\varepsilon_{vM}(z) = \frac{1}{\sqrt{3}} \sqrt{4 \cdot \varepsilon_{xx}^2(z) + \gamma^2(z)} \quad (12)$$

The shear distribution across the thickness is not homogeneous (Fig. (8)) implying that the equivalent strain will also reveal heterogeneous character along the normal direction (Fig. (9)). As can be seen in Fig. (8), the shear strain is not linear in investigated cases A and B. In case of condition A, the maximum is observed in the subsurface region due to phenomenon of reverse shear, while in case of B the amount of γ first linearly increases from the mid-thickness to the sub-surface and afterward tends to saturate within the surface layers. The analytical model developed in this work is capable of reproducing the evolutionary patterns of strain evolution presented in Figs. (8) and (9). The deviations observed between the SDM and FEM are attributed to the simplifications made in the proposed model.

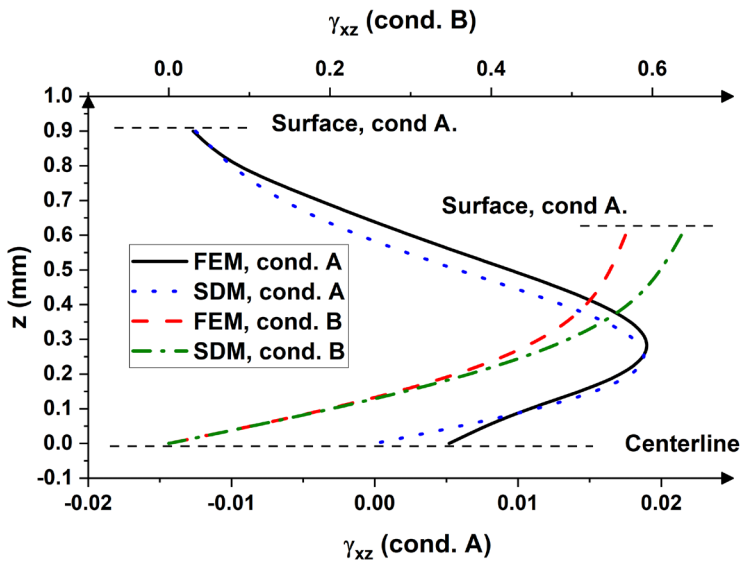


Figure 8

Distribution of shear strain for different rolling conditions as predicted by the new analytical model:

a) condition A ($R=250$ mm, $h_0=1$ mm, $h=0.9$ mm, $r=0.1$, $\mu=0.08$, $v=2$ m/min); b) condition B

($R=250$ mm, $h_0=1$ mm, $h=0.6$ mm, $r=0.4$, $\mu=0.25$, $v=2$ m/min)

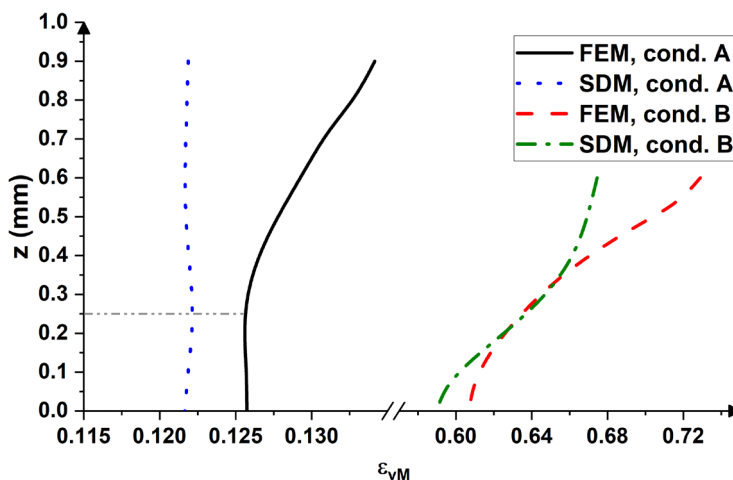


Figure 9

Distribution of von Mises strain for different rolling conditions as predicted by the new analytical model: a) condition A ($R=250$ mm, $h_0=1$ mm, $h=0.9$ mm, $r=0.1$, $\mu=0.08$, $v=2$ m/min); b) condition B ($R=250$ mm, $h_0=1$ mm, $h=0.6$ mm, $r=0.4$, $\mu=0.25$, $v=2$ m/min)

Alternatively, to Eqs. (10-12), the assessment of both shear and equivalent strains can be calculated by employing Eqs. (13) and (14). The accuracy of this method was tested on the ultrafine-grain structured metals [16]. Since the shear strain is mainly localized within the surface layer for a higher thickness reduction as a result of the reverse flow phenomenon, it was suggested to calculate the surface strain ε_s , while the ε_{vM} is computed by substituting Eqs. (13) and (14):

$$\varepsilon_s = \frac{2(1-r)^2}{r(2-r)} \gamma \ln\left(\frac{1}{1-r}\right) \quad (13)$$

$$\varepsilon_{vM} = \sqrt{\frac{4}{3} \left(\ln\left(\frac{1}{1-r}\right) \right)^2 + \frac{\varepsilon_s^2}{3}} \quad (14)$$

Both Eq. (14) and (12) require the amount of γ , imposed by rolling, which can be computed using Eq. (11). Fig. (10) suggests that these two methods, Eqs. (12) and (14), provide similar results.

Table 2
Shear and equivalent strain values according to different methods and calculations

Condition	Equation	FEM		FLM		SDM	
		ε_s	ε_{eq}	ε_s	ε_{eq}	ε_s	ε_{eq}
A	Eq. (12)	-0.01669	0.1159	0.01167	0.1157	-0.1149	0.1157
A	Eq. (14)	-0.1499	0.1219	0.01048	0.1218	-0.01032	0.1218
B	Eq. (12)	0.5852	0.5723	0.8675	0.6812	0.6279	0.5872
B	Eq. (14)	0.3363	0.6210	0.4985	0.6263	0.3609	0.6256

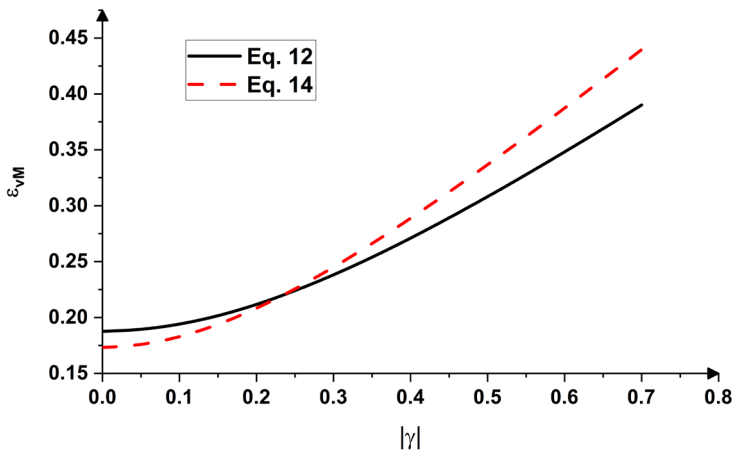
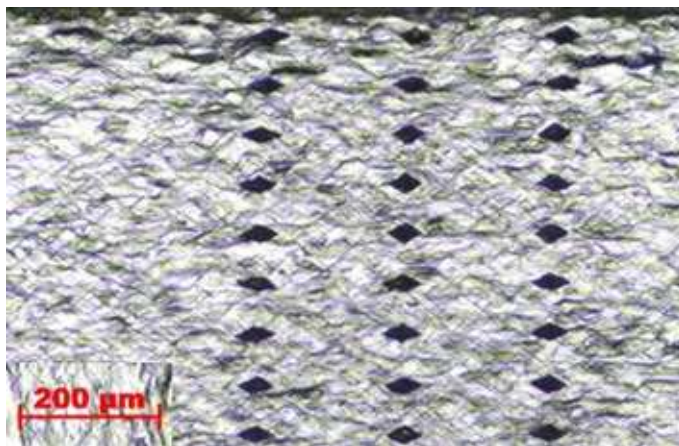
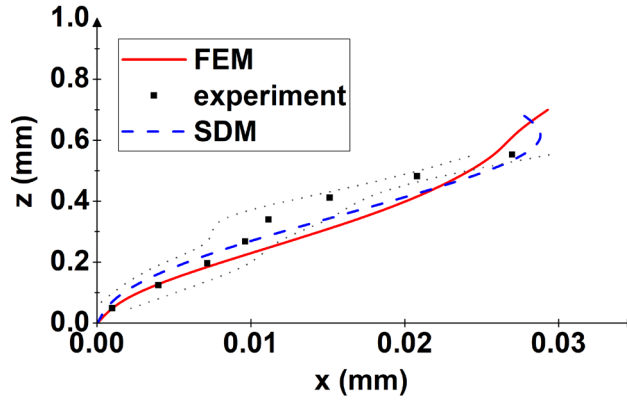


Figure 10

Measured (a) and calculated (b) displacement patterns for 30% thickness reduction ($h=0.7$ mm and $\mu=0.075$). The dotted lines on graph (b) represent the upper and lower bounds of experimental variations



(a)



(b)

Figure 11

Displacement for $R=250$ mm, $h_0=1$ mm, $h=0.7$ mm, $r=0.3$, $\mu=0.075$, $v=2$ m/min, (a) measured values, (b) data

Table 2 summarizes both shear and equivalent strain values calculated for various roll gap geometries by various equations. As it follows from Table 2, the FEM and SDM methods ensure similar values for ε_s and ε_{eq} , while the FLM is less accurate in capturing the shear strain.

Fig. (11a) presents the distortion patterns of initially vertical lines, made by microindentation, after $r=30\%$ thickness reduction in 1050 Al alloy. As it turns out (Fig. (11b)), both FEM and SDM outputs are in good agreement with the experimentally measured counterparts.

4.2 Calculation of Stress Values

Knowing the equivalent strains and material's hardening law [18], the von Mises stress distribution can be evaluated by Eq. (15). The calculated stress distribution for cases A and B reveals that stress variations across the thickness are approximately 5%. The calculated stresses for different rolling conditions as predicted by both FEM and the new analytical model for conditions A ($R=250$ mm, $\mu=0.08$, $h_0=1$ mm, $h=0.9$ mm, $v=2$ m/min) and B ($R=250$ mm, $\mu=0.25$, $h_0=1$ mm, $h=0.6$ mm, $v=2$ m/min) are as following: $\sigma_{vM,FEM}$ (A-surface)= 99.03 MPa, $\sigma_{vM,SDM}$ (A-surface)= 97.15 MPa, $\sigma_{vM,FEM}$ (A-mid-thickness)= 97.79 MPa, $\sigma_{vM,SDM}$ (A-mid-thickness)= 97.12 MPa, $\sigma_{vM,FEM}$ (B-surface)= 138.93 MPa, $\sigma_{vM,SDM}$ (B-surface)= 136.79 MPa, $\sigma_{vM,FEM}$ (B-mid-thickness)= 133.93 MPa, $\sigma_{vM,SDM}$ (B-mid-thickness)= 133.17 MPa. This change seems to be negligibly small, but it should be noted that even small stress diversities are capable of triggering the evolution of microstructural heterogeneities in variously oriented grains.

$$\sigma_{vM} = 148 \cdot \varepsilon_{vM}^{0.2} \text{ [MPa]} \quad (15)$$

Apart from the above formulation, there are other material models, such as the cubic polynomial strain-stress function described in Ref. [14] or the Ramberg-Osgood model [19].

The presented analytical model can be employed for the estimation of the strain/stress distribution as well as the amount of shear strain/stress in cold rolled aluminum sheets or other metals. There are many pros and cons regarding the implementation of SDM. Results of experimental observations and FEM simulations clearly demonstrate that the major advantage of the analytical model developed is that it can accurately capture the evolution of both strain and stress heterogeneity across the thickness of a rolled sheet. Compared to time costly FEM simulations, this approach is capable of providing a solution within a fraction of a second. It should be noted that this model cannot predict the strain/stress evolution over time and neglects the effect of temperature and rolling velocity, however, the effect of these technological parameters on properties of cold rolled materials is negligibly small. The model can be extended by taking into account the deflection of the rolls as it is described elsewhere [20]. The disadvantage of the presented approach is that the model parameters should be defined for each roll diameter individually, though, the fitting parameters can be defined for various rolling stands according to the procedure, presented in this contribution, and this will ensure a very fast and accurate simulation of cold rolling process.

Conclusions

The normal and shear components of material strain, introduced by rolling, can be accurately modelled, by the presented analytical description. The model parameters were defined from previously published experimental data, found in various literature sources, and results of finite element calculations.

The new model reveals very reasonable agreement with the data calculated by FEM and ones measured experimentally. The approach can be used for the description of evolution of strain components in cold rolling. The presented analytical solution was successfully tested for thickness reductions ranging from 10% to 40% while the friction coefficient varied between 0.025 and 0.25. The discrepancies observed between the FEM calculations and ones produced by the analytical model developed are attributed to the simplifications made in the SDM.

Compared to finite element or flow line models, the developed approach neglects the kinetics of material flow, nonetheless, it can guarantee a high accuracy of strain evolution in cold rolling process. The extended model can be used for rapid analysis of symmetric rolling and is capable of capturing the phenomenon of reverse displacement in Al alloys. By determining corresponding model parameters, this approach can be employed for strain/stress evolution in various metals. Results of model calculations suggest that the reverse displacement tends to disappear, by increasing both the degree of deformation and friction coefficient.

Acknowledgement

Project no. TKP2021-NVA-29 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

References

- [1] Boldetti, C., Pinna, C., Howard, I. C., Gutierrez, G., “Measurement of deformation gradients in hot rolling of AA3004”, *Experimental Mechanics*, Vol. 45, pp. 517-525, 2005, DOI: 10.1007/BF02427905
- [2] Roumina, R., Sinclair, C., “Deformation geometry and through-thickness strain gradients in asymmetric rolling”, *Metallurgical and Materials Transactions A – Physical Metallurgy and Materials Science*, Vol. 39, pp. 2495-2503, 2008, DOI: 10.1007/s11661-008-9582-6
- [3] Sidor, J. J., Petrov, R., Kestens, L., “Texture control in aluminium sheets by conventional and asymmetric rolling”, In: *Comprehensive Materials Processing*, Vol. 3: Advanced Forming Technologies, Elsevier, pp. 447-498, 2014, DOI: 10.1016/B978-0-08-096532-1.00324-1
- [4] Beausir B., Tóth L., “A New Flow Function to Model Texture Evolution in Symmetric and Asymmetric Rolling”, In: Haldar A., Suwas S., Bhattacharjee D. (eds), *Microstructure and Texture in Steels*, Springer, London, 2009, pp. 415-420, DOI: 10.1007/978-1-84882-454-6_25
- [5] Sidor, J. J., “Deformation texture in Al alloys: Continuum mechanics and crystal plasticity aspects”, *Modelling and Simulation in Materials Science and Engineering*, Vol. 26(8), 085011, 2018, DOI: 10.1088/1361-651X/aae886
- [6] Sidor, J. J., “Assessment of flow-line model in rolling texture simulations”, *Metals* Vol. 9(10), 1098, 2019, DOI: 10.3390/met9101098
- [7] Decroos, K., Sidor, J. J., Seefeldt, M., “A new analytical approach for velocity field in rolling processes and its application in through-thickness texture prediction”, *Metallurgical and Materials Transactions A*, Vol. 45, pp. 948-961, 2014, DOI: 10.1007/s11661-013-2021-3
- [8] Sidor, J. J., Xie, Q., “Deformation texture modelling by mean-field and full-field approaches”, *Advanced Materials Letters*, Vol. 10(9), pp. 643-650, 2019, DOI: 10.5185/amlett.2019.0030
- [9] Cawthorn, C. J., Loukaides, E., Allwood, J., “Comparison of analytical models for sheet rolling”, *Procedia Engineering*, Vol. 81, pp. 2451-2456, 2014, DOI: 10.1016/j.proeng.2014.10.349
- [10] Minton, J., Cawthorn, C. J., Brambley, E., “Asymptotic analysis of asymmetric thin sheet rolling”, *International Journal of Mechanical Sciences*, Vol. 113, pp. 36-48, 2016, DOI: 10.1016/j.ijmecsci.2016.03.024

- [11] Avitzur, B., “Friction-aided strip rolling with unlimited reduction”, *International Journal of Machine Tool Design and Research*, Vol. 20(3-4), pp. 197-210, 1980, DOI: 10.1016/0020-7357(80)90004-9
- [12] Szűcs, M., “Többszintű modellezés alkalmazása a szimmetrikus és az aszimmetrikus hengerlési folyamatok vizsgálatára” (The use of multi-scale modelling to study symmetric and asymmetric rolling processes), PHD Thesis, University of Miskolc, Hungary, 2017 (in Hungarian) [online] Available at: http://193.6.1.94:9080/JaDoX_Portlets/documents/document_25524_section_20959.pdf [Accessed: 25 09 2021]
- [13] Pesin, A., Pustovoytov, D., “Influence of process parameters on distribution of shear strain through sheet thickness in asymmetric rolling”, *Key Engineering Materials*, Vol. 622-623, pp. 925-935, 2014, DOI: 10.4028/www.scientific.net/KEM.622-623.929
- [14] Ďurovský, F., Zboray, L., Ferková, Ž., “Computation of rolling stand parameters by genetic algorithm”, *Acta Polytechnica Hungarica*, Vol. 5, No. 2, pp. 59-70, 2008
- [15] Inoue, T., “Strain variations on rolling condition in accumulative roll-bonding by finite element analysis”, In: David Moratal (ed.), *Finite Element Analysis*, InTech, London pp. 598-610, 2010, DOI: 10.5772/10233
- [16] Ma, C., Hou, L., Zhang, J., Zhuang, L., “Experimental and numerical investigation of plastic deformation during multi-pass asymmetric and symmetric rolling of high-strength aluminium alloys”, *Material Science Forum*, Vol. 794-796, pp. 1157-1162, 2014, DOI: 10.4028/www.scientific.net/MSF.794-796.1157
- [17] Inoue, T., Qiu, H., Ueki, R., “Through-thickness microstructure and strain distribution in steel sheets rolled in a large-diameter rolling process”, *Metals*, Vol. 10(1) 91, 2020, DOI: 10.3390/met10010091
- [18] Van Haafden, V. M., Magnin, B., Kool, W. H., Katgerman, L., “Constitutive behavior of as-cast AA1050, AA3104, and AA5182”, *Metallurgical and Materials Transactions A*, Vol. 33A, pp. 1971-1980, 2002, DOI: 10.1007/s11661-002-0030-8
- [19] Zsiha, K., “Stress-strain interaction model of plasticity”, *Acta Polytechnica Hungarica*, Vol. 12, No. 1, pp. 41-54, 2015, DOI: 10.12700/APH.12.1.2015.1.3
- [20] Kucséra, P., Béres, Zs., “Hot rolling mill hydraulic gap control (HGC) thickness control improvement”, *Acta Polytechnica Hungarica*, Vol. 12, No. 6, pp. 93-106, 2015, DOI: 10.12700/aph.12.6.2015.6.6

Sample-in-the-Loop Laser Speckle Contrast Imaging Based on Optimization

Máté Siket^{1,2}, Imre Jánoki¹, Ádám Nagy¹, and Péter Földesy¹

¹Institute for Computer Science and Control, Kende utca 13-17, H-1111, Budapest, Hungary; siket.mate@sztaki.hu; janoki.imre.gergely@sztaki.hu; nagy.adam@sztaki.hu; foldesy.peter@sztaki.mta.hu

²Physiological Controls Research Center, Óbuda University, Bécsi út 96/b, Budapest, H-1034

Abstract: Laser Speckle Contrast Imaging (LSCI) is an optical method mainly used for creating blood flow maps. Despite its beneficial properties, the technique is yet to find a place in clinical practice. In this work, we propose a setup for LSCI to overcome some of the disadvantages associated with the method. We call the setup the sample-in-the-loop LSCI as it is based on the feedback of the captured image, which is determined by the properties of the sample and the experimental setup. We investigate and demonstrate the method in three exemplary scenarios: optimization to specific contrast setpoint, sensitivity maximization and dynamic range maximization. These goals are achieved by using optimization on the laser light pulse sequence and on the exposure time of the digital camera.

Keywords: laser speckle contrast imaging; sample-in-the-loop; time varied illumination; dynamic range maximization

1 Introduction

Despite recent years show a decrease or slowed growth in the utilization rates of non-invasive diagnostic imaging modalities, they are still cornerstones in terms of clinical applications and are the main focus of scientific research [1]. The medical imaging practice has its traditional, proven, widely applied techniques such as X-ray, computed tomography, magnetic resonance imaging, or ultrasound. However, there are modalities yet to become a part of clinical practice, one of which is Laser Speckle Contrast Imaging (LSCI) [2, 3]. A possible drawback of LSCI is the sensitivity to calibration, dynamic range, and exposure [4, 5]. In our current work, we propose a way to alleviate these by partly automating processes based on feedback. LSCI utilizes the interference pattern caused by the reflection of coherent light from a medium with static and dynamic scatterers [6]. In the specific case when the reflective medium is completely static, the imaging device observes a frozen pattern of the so-called speckles. However, if motion occurs (e.g. blood flows) the pattern changes and

decorrelates in time. The level of decorrelation can be quantified by calculating the contrast in a region of interest (ROI). Carrying out the calculation using a sliding window results in a contrast map, which tells us about the relative flow speeds in vessels with various sizes, the progression of diabetes-induced vascular complications [7] or regeneration of a burned tissue [8, 9]. Among others, one of the disadvantages of the technique is the relatively low dynamic range. Previously multi-exposure techniques have been investigated to increase the dynamic range and to suppress the detrimental effect of static scatterers [10–12]. Furthermore, under- or overexposure can also greatly affect the observed contrast values. To this end, [13] introduced a correction term to counteract the effect of underexposed images. [14] recommends adjusting the average intensity in the image to be around or slightly below the middle of the dynamic range of the sensor.

A recent study [15] showed that the dynamic range of LSCI can be greatly improved by utilizing a novel, time-varied illumination during camera exposure. Discrete pulsetrains were used for the realization of the time-varied illumination. Different shapings of the discrete pulsetrains have been investigated, however, they share the common aspect of predefinition or model-based optimization of the sequence, and both are done offline. The offline approach is disadvantageous because (i) it requires an accurate model, (ii) two measurements are needed (before and after identification of the model), and (iii) the sample is subject to changes between measurements.

The experienced drawbacks motivated the creation of an alternative method. Here, we propose to optimize the discrete pulse sequence in an online, model-free manner, calling it the sample-in-the-loop approach. Although the concept originates from the time-varied illumination approach we do not limit ourselves to that specific case. In contrast to the varied illumination, hereinafter, we refer to the constant illumination as the continuous wave operation of the laser light. Compared to multiexposure methods [11, 12, 16], the online, optimization-based approach can take into account the average intensity; instead of making post-measurement corrections [13, 17] can keep the intensity in a favorable range. Also, the method does not need an extensive calibration process, as it was proposed in [18].

We demonstrate the concept in a custom-developed channel slide through three experimental scenarios: optimization to specific contrast setpoint, sensitivity maximization, and dynamic range maximization. The three distinct scenarios touch on different aspects; continuous wave versus time-varied operation dictates the usage of different optimization algorithms and scenarios can differ in the number of required ROIs.

The paper is structured as follows. First, in Section 2 we detail the sample in the loop setup with the designed channel slide. Next, the approach to the laser pulse optimization is given, defining the optimization goals and methods. In the Results section, the three experimental scenarios are investigated separately, and finally, in Section 4.3 conclusions are drawn.

2 Sample in the loop setup

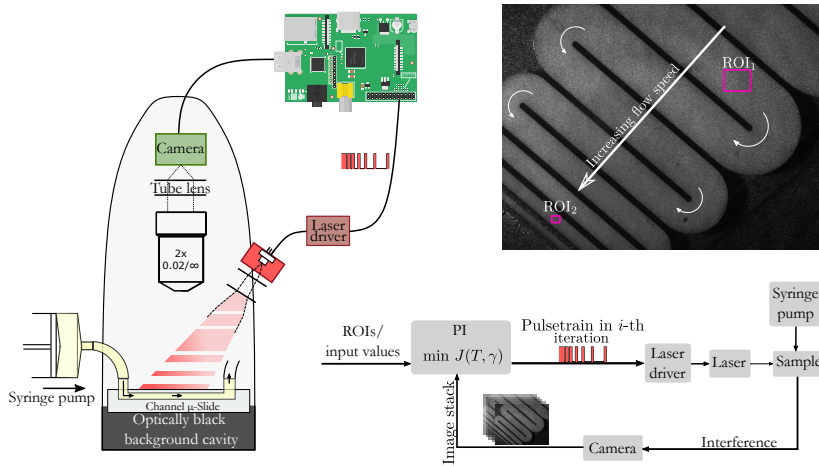


Figure 1

The experimental setup can be seen on the left-hand side. A medical syringe pump provides a constant flow speed, while the laser and the digital camera are controlled by the Raspberry Pi. A raw speckle image with a typical ROI selection is illustrated in the upper right corner. The schematic diagram in the lower right corner summarizes the sample-in-the-loop LSCI method.

A Raspberry Pi 4 Model B carries out the digital pulse wave generation, the camera control, and the image processing tasks through a custom-developed software. Fig. 1 depicts the sample-in-the-loop experimental setup. The protocol of an experiment can be summarized as follows: First, the user selects the optimization scenario and the related thresholds and reference values. Second, an initial pulse wave is generated to capture a single image for ROI selection purposes. Afterward, the optimization algorithm updates the pulse sequences based on 10 averaged contrast images (with a resolution of 2000x1500 pixels). It continues to do so until a maximum number of iterations or termination limit is reached. The pulsetrains (leaving the GPIO pins of the Pi) control the 660 nm 50 mW laser diode through a laser driver (LDP-VRM 01-12 CA, PicoLAS, Germany) and at the same time trigger the monochrome digital camera (Basler acA2040-55um). Based on preliminary investigations we limited the minimum pulse width to 10 μs . With shorter pulses the proportion of laser transients becomes dominant, which introduces two unwanted effects: average intensity significantly lowers, and contrast lowers because of the larger proportion of incoherent light [19]. Furthermore, the laser is mounted in a thermally stabilized mount (LDM21, Thorlabs, Germany) driven by a Thorlabs TED200C temperature controller. The constant temperature improves contrast by reducing temperature-induced laser mode hopping.

2.1 Designed channel slide

For demonstration purposes, we designed a channel slide 2, where the cross sections range linearly from $100\ \mu\text{m}$ to $1000\ \mu\text{m}$ in $100\ \mu\text{m}$ increments. We created a simulation of the flow in the channel design using COMSOL Multiphysics [20], which showed that the velocities at the center of the straight sections are roughly proportional to the diameter. The microfluidic channel slide was created using PDMS bonded to glass with a depth of $100\ \mu\text{m}$.

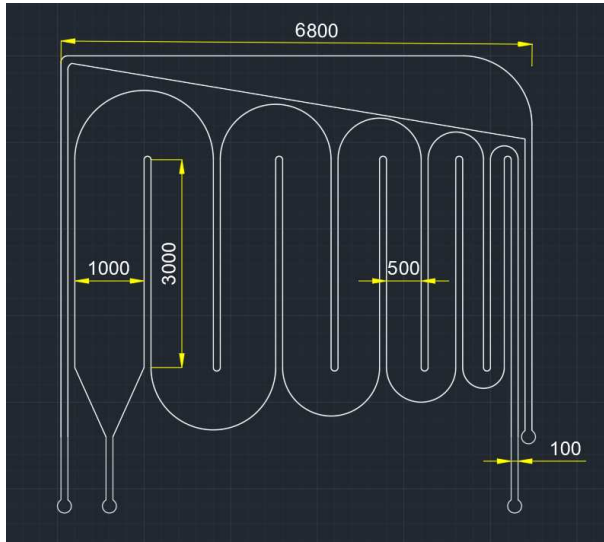


Figure 2

The channel slide design we used for evaluation. Based on the simulated model, the flow speeds at the center of each straight section is proportional to their diameter. The annotated sizes on the image are in μm .

3 Laser pulse optimization

The validity of pulsed speckle contrast imaging has already been proven in [21]. In this method, uniform intensity and uniform length discrete laser pulses with uniform pulse rates are used instead of continuous wave illumination. Instead of the well-known multi-exposure methods [11, 12] that were developed to extend the dynamic range, we use time-varied illumination laser speckle contrast imaging [15] which is based on pulsed speckle contrast imaging. It uses laser pulsetrains of uniform intensity and uniform length discrete pulses, however, it changes the density of the pulses that ultimately simulates a varying intensity during a single exposure. Using a pulse sequence made of multiple density pulsetrains (e.g. 3 pulsetrains of equal length with 75%, 50% and 25% duty cycles concatenated) enables high dynamic range flow rate imaging during this

single exposure.

We made this discrete pulse sequence to be a function of a single variable. This function limits the number of feasible sequences but helps with the optimization problem by introducing only one more variable besides the exposure length. The additional parameter affects the spacing – in a form of a power function – between two consecutive high states. The sequence starts with an initial high state. The next high state is defined by the number of low states (n_{LOW}) following the initial high state as:

$$n_{LOW}(i) = \text{round}(i^\gamma) - 1, \quad (1)$$

where the discrete sequence is loaded with binary values from $i = 0$ to $i = \frac{T}{10 \mu s}$, T is the exposure length, $\gamma \geq 0$ is the spacing factor. If $n_{LOW} \leq 0$ the next state will remain a high state. The continuous wave laser setup is a limiting case when $\gamma = 0$, typically we evaluated $0 \leq \gamma \leq 1.4$. Fig. 3 illustrates how the spacing factor affects the average ROI intensity for different exposure lengths.

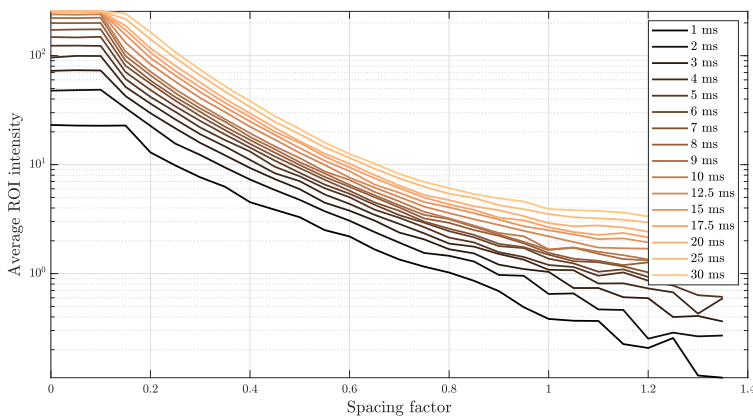


Figure 3

Measured average intensities in an ROI are visualized as a function of the spacing factor. It can be observed that after an initial interval between 0 and 0.1, a spacing factor of 1 reduces the intensity by more than an order of magnitude. Small values of the spacing factor would have only a visible effect on exposures above 30 ms.

The typical way of quantifying flow speed in LSCI is by means of contrast calculation. The speckle contrast is calculated on a single digital image or sequence of consecutive images (to decrease noise level), in both cases using a sliding window. The contrast is calculated by taking the quotient of the standard deviation and the mean of the intensities in the current window. The window size is usually $n \times n$, where n is in the range of 5 to 15 [22]. The spatial speckle

contrast is defined as:

$$\kappa = \frac{\sigma(I_s)}{\langle I_s \rangle}, \quad (2)$$

where $\sigma(I_s)$ and $\langle I_s \rangle$ are the sample standard deviation and the sample mean in the current window. The choice of window size affects the noisiness and the resolution of the resulting contrast map. The contrast value is averaged on 10 consecutive images in order to reduce noise and improve the consistency of the optimization.

We defined three major optimization problems: contrast setpoint, sensitivity maximization, and dynamic range maximization. Previously it was shown in [13] that the intensity has a significant effect on the contrast values. Thus, besides the major objectives, all three share an optional penalization for underexposed images; a linear term penalizes the given parameter/parameter set when the average intensity falls below a predefined threshold in the ROIs. In our experimental setup ROI sizes ranged between 10x10 to 50x50 depending on the size of the evaluated cross-section in the channel slide.

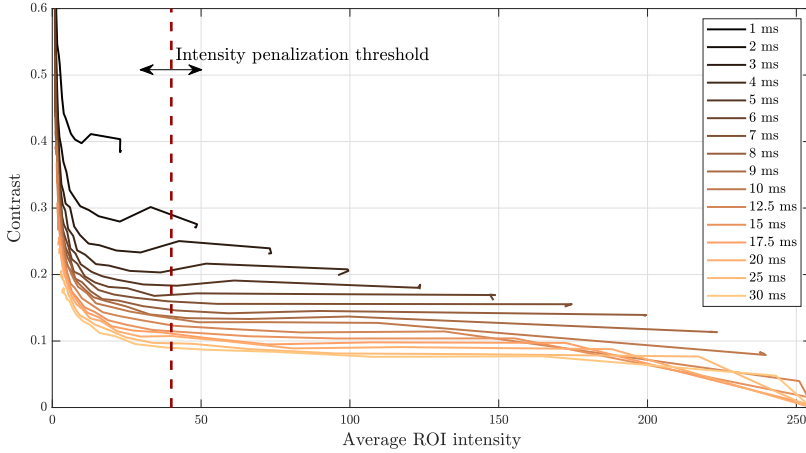


Figure 4

Observed speckle contrasts plotted with respect to the average ROI intensity. Low intensities can cause falsely observed large contrast values, while high intensities can cause falsely observed low contrast values. In order to avoid the potentially detrimental effects, a method of intensity penalization is implemented during the optimization. The figure indicates that a threshold around an intensity level of 40-50 guarantees a linear, flat response.

3.1 Contrast setpoint

Optimization based methods are widely used in control problems [23–26], we defined the first scenario as an optimization to setpoint contrast value. The user

can select an arbitrary ROI and define a desired average contrast value in that particular ROI. Then the gradient descent algorithm iteratively modifies the continuous wavelength, which also determines the exposure length of the camera. This first scenario aims to approach the sample-in-the-loop concept in the simplest manner. For this reason, we decided to optimize a single variable with the gradient descent method as follows:

$$T_{i+1} = T_i - \alpha \frac{\Delta J}{\Delta T}, \quad (3)$$

where α is the step size, Δ denotes the differences in cost J and exposure length T between the i -th and $i - 1$ -th iteration.

$$\begin{aligned} \min_T \quad & J(T) = |\kappa(T) - \kappa_{ref}| + J_{exp.}, \\ \text{s.t.} \quad & T \in [1, 30], \\ & J_{exp.} = \begin{cases} 0, & \text{if } \mu \geq \mu_{ref}, \\ \frac{\mu - \mu_{ref}}{\lambda}, & \text{else.} \end{cases} \end{aligned} \quad (4)$$

where κ is the average calculated contrast in the ROI, κ_{ref} is the user-defined contrast setpoint, and μ_{ref} is the average intensity threshold. The offset from the reference intensity μ_{ref} is normalized by λ in order to scale the additional penalization in the range of the unpenalized cost, in our experimental setup, a reasonable value for the λ was around 40. The parameters of the control: α , the number of iterations, and the λ were determined based on multiple experiments with different channel slides and flow speeds.

3.2 Sensitivity maximization

The cost is defined as the inverse of the difference:

$$\begin{aligned} \min_T \quad & J(T) = \frac{1}{|\kappa_1(T) - \kappa_2(T)| + \varepsilon} + J_{exp.}, \\ \text{s.t.} \quad & T \in [1, 30], \\ & J_{exp.} = \begin{cases} 0, & \text{if } \mu \geq \mu_{ref}, \\ \frac{\mu - \mu_{ref}}{\lambda}, & \text{else.} \end{cases} \end{aligned} \quad (5)$$

where κ_1 and κ_2 are the average contrast values in the respective ROI and ε avoids division by zero.

3.3 Dynamic range maximization

Study [15] showed that by utilizing time variant illumination during the camera exposure the dynamic range of the laser speckle contrast imaging can be greatly improved.

$$\begin{aligned}
& \min_{T, \gamma} J(T, \gamma) = |\kappa_1(T, \gamma) - \kappa_{ref}| + |\kappa_2(T, \gamma) - \kappa_{ref}| + J_{exp.}, \\
& \text{s.t. } T \in [1, 30], \quad \gamma \in [0, 1.4], \\
& J_{exp.} = \begin{cases} 0, & \text{if } \mu \geq \mu_{ref}, \\ \frac{\mu - \mu_{ref}}{\lambda}, & \text{else.} \end{cases}
\end{aligned} \tag{6}$$

where κ_{ref} is a user-defined contrast value, around which the sensitivity ought to be the least.

4 Results

4.1 Optimization to contrast setpoint

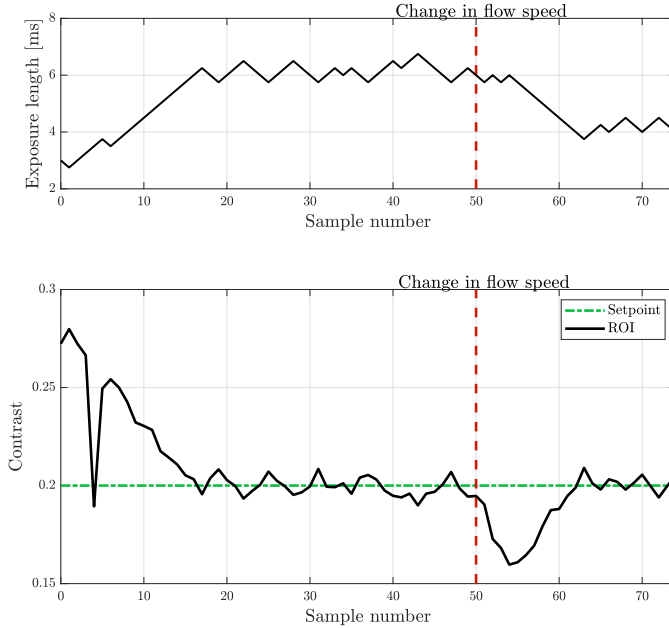


Figure 5

The set exposure length and measured contrast with respect to the current sample number are illustrated by solid black lines. The flow speed of the syringe pump has been modified during the experiment (denoted by the dashed red line), creating an external disturbance on the system. The contrast setpoint is depicted by the dashed green line.

Fig. 5 demonstrates the result of the first scenario, namely the optimization for a specific contrast value. A contrast value of 0.27 is observed with an initial exposure length of 3 ms. Then the gradient descent algorithm iteratively increases the exposure length to match the reference 0.2 contrast value. Besides the fluctuation caused by the measurement noises, the algorithm settles for an exposure length of 6 ms. When the disturbance occurs (in a form of flow speed change), the observed contrast drops by 0.05, but the algorithm compensates for it after a couple of iterations and settles for a new 4 ms exposure length. A change in flow speed induces a change in the measured contrast values, which are based on the captured image. The larger deviance in the contrast value ultimately results in a larger cost. A new optimum can be found by either increasing or reducing the exposure time so that deviance can be counteracted. The compensation can be done only in a feasible range of the parameters which determine the operation of the laser and the camera.

4.2 Sensitivity maximization

Despite that the optimization of the exposure length would be sufficient for sensitivity maximization, we carried out an experiment using both of the variables for demonstration purposes: the spacing factor and the exposure length.

Fig. 6 demonstrates well that the optimizer achieved the lowest costs when the spacing factor was close to zero. The small spacing factor in practical perspective translates to a continuous wave-like operation, which is otherwise expected to yield the best result in terms of sensitivity.

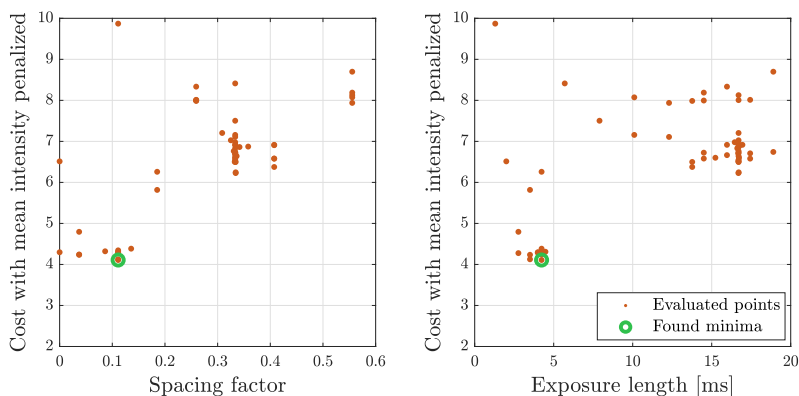


Figure 6

Evaluated spacing factor-exposure length combinations are illustrated during the sensitivity maximization scenario. It can be seen that the lowest costs are achieved if the optimizer converges the pulse sequence to the continuous wave operation.

4.3 Dynamic range maximization

In Fig. 7 we showcase an exemplary optimization using the grid search algorithm [27]. The algorithm aims to minimize the cost function given in (6). It can be seen that it converges to a spacing factor of around 0.3 and to an exposure length of 8 ms, where a minimum is found. In this multivariate scenario, the grid search algorithm was preferred in order to avoid the possibility of quick convergence to a local minimum.

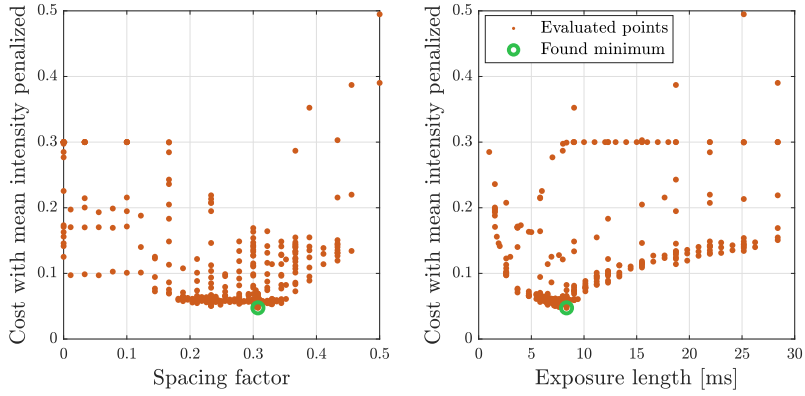


Figure 7

Evaluated spacing factor-exposure length combinations are illustrated during the dynamic range maximization scenario. It can be seen that the cost can be lowered by applying time-varying illumination. An optimal solution is found around a spacing factor of 0.3 and an exposure length of 8 ms.

We explored the effect of the spacing factor and exposure length in order to ascertain the cost surface and that the algorithm indeed found a minimum. The exposure length is varied in a range from 1 ms to 30 ms and the spacing factor is from 0 to 1.4. Fig. 8 and 9 depict the optimization surfaces given two different reference contrast values. The reference value of (6) is set to 0.1 in Fig. 8 and 0.3 in Fig. 9.

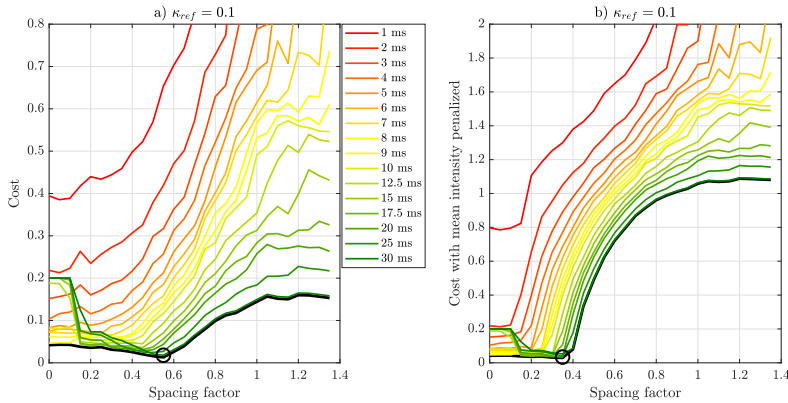


Figure 8

The measured optimization surface when the reference value is set to 0.1 is approximated with a family of curves. Subplot a) represents the "raw" cost, while subplot b) the cost when the average intensity is penalized.

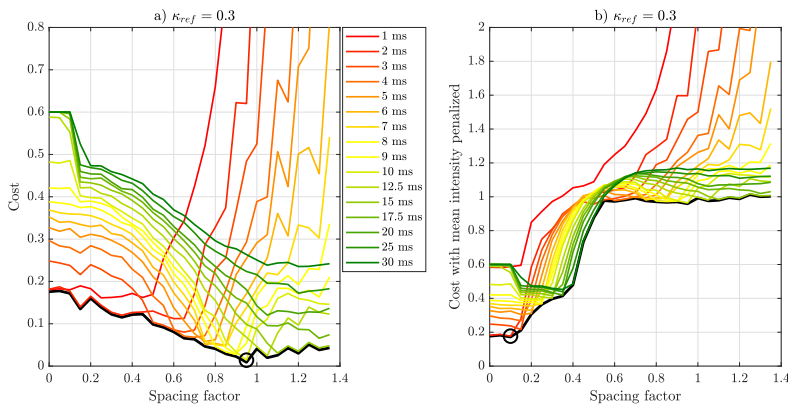


Figure 9

The measured optimization surface when the reference value is set to 0.3 is approximated with a family of curves. Subplot a) represents the "raw" cost, while subplot b) the cost when the average intensity is penalized.

The behavior of the cost function towards the minimum (0) and maximum (1.4) of the spacing factor is dominated by the phenomenon related to the average intensities. Firstly, for small values of the spacing factor, the laser switches to continuous wave-like operation, which means increased exposure. The increased exposure (mostly in the case of longer exposure times) leads to saturation, which in return lowers the observed contrast (as indicated in Fig. 4). Zero contrast values consequently saturate the cost function, as it can be seen in Fig. 8 and 9 subplots a) for exposure times above 20 ms. Secondly, the

countereffect of high spacing factors introduces numerical instability into the calculation of the speckle contrast. Falsely observed high contrast values increase the cost; numerical instability appears at smaller spacing factors for shorter exposure times. Subplots b) showcase the intensity penalized cost surface. The significant difference between the raw and the penalized versions is that the cost towards the larger spacing factors is affected by an additional term, which penalizes the underexposed images in a linear way. The start of the penalization – when the exposure falls below the predefined threshold – is dependent on the actual exposure length as expected.

The theoretical minimum moves toward the continuous wave operation mode for larger values, as it is indicated by the black circle on the envelope. Such behavior is expected since larger spacing factors have a great effect on the dynamic range, making the response so wide that the contrast values fall below the reference values, hence increasing the cost. Fig. 10 depicts the Pareto front around the minimum; it can be seen that similar costs are achieved with different combinations of exposure length and spacing factor.

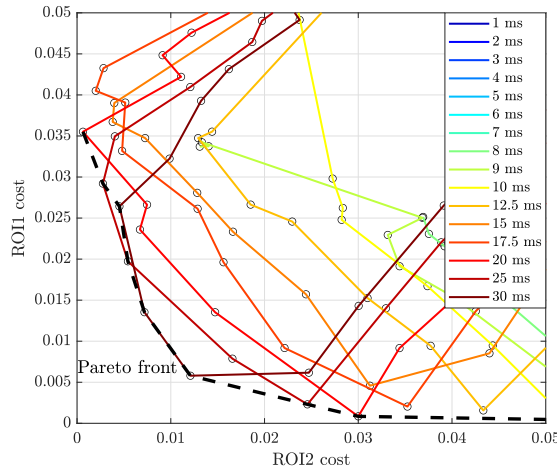


Figure 10

As the dynamic range maximization in (6) is formulated as a multi-objective optimization, a Pareto front can be found, where ROI1 and ROI2 costs represent the distances from the reference contrast value.

Conclusions

In this study, we proposed the sample-in-the-loop approach for LSCI applications. We defined three use cases in order to demonstrate the applicability of the method. We did not aim to prove or validate the utilization of the three use cases but to demonstrate the concept through different scenarios.

In the first scenario, the algorithm realized a setpoint optimization for the

reference contrast value and was able to compensate for an external disturbance coming from flow speed change. The solution can be a help running automated measurements, however, it is important to note that the method can provide information only about the relative flow speeds in a given time instant or if the external factors are constant. This is the case since the algorithm will compensate (in a feasible range) for changes in external factors such as flow speed.

Previous studies indicated and our findings confirmed that the sensitivity maximization is best achieved as a continuous wave operation, where we optimized the exposure length using the gradient descent algorithm. Regarding the dynamic range maximization, we demonstrated that the proposed algorithm can find a minimum, and this minimum is dependent on the actual reference contrast value. Furthermore, with the current laser setup, the exposure penalization leaves small room for dynamic range improvements. This highlights the importance of sufficiently high laser power as large spacing factors reduced the total exposure by more than an order of magnitude. The explored cost surface indicates that it is sufficient to use gradient descent in the multivariate case as well since the cost surface is a convex function with respect to the free variables. The application of the gradient descent can then accelerate the convergence to a minimum and the fulfillment of the measurement protocol.

Acknowledgement

The work was supported in part by the Eötvös Loránd Research Network Secretariat (Development of cyber-medical systems based on AI and hybrid cloud methods) under Agreement ELKH KÖ-37/2021.

References

- [1] R. Smith-Bindman, M. L. Kwan, E. C. Marlow, M. K. Theis, W. Bolch, S. Y. Cheng, E. J. A. Bowles, J. R. Duncan, R. T. Greenlee, L. H. Kushi, J. D. Pole, A. K. Rahm, N. K. Stout, S. Weinmann, and D. L. Miglioretti. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *JAMA*, 322(9):843–856, 09 2019.
- [2] W. Heeman, W. Steenbergen, G. M. van Dam, and E. C. Boerma. Clinical applications of laser speckle contrast imaging: a review. *Journal of Biomedical Optics*, 24(8):1 – 11, 2019.
- [3] D. Briers, D. Duncan, E. Hirst, S. Kirkpatrick, M. Larsson, W. Steenbergen, T. Stromberg, and O. Thompson. Laser speckle contrast imaging: Theoretical and practical limitations. *Journal of biomedical optics*, 18:66018, 06 2013.
- [4] C. Wang, Z. Cao, X. Jin, W. Lin, Y. Zheng, B. Zeng, and M. Xu. Robust quantitative single-exposure laser speckle imaging with true flow speckle contrast in the temporal and spatial domains. *Biomed. Opt. Express*, 10(8):4097–4114, Aug 2019.
- [5] P. Földesy, M. Siket, I. Jánoki, K. Demeter, and Ádám Nagy. Ensemble averaging laser speckle contrast imaging: statistical model of improvement as function of static scatterers. *Opt. Express*, 29(18):29366–29377, Aug


- 2021.
- [6] A. Fercher and J. Briers. Flow visualization by means of single-exposure speckle photography. *Optics Communications*, 37(5):326–330, 1981.
 - [7] O. A. Mennes, J. J. van Netten, J. G. van Baal, and W. Steenbergen. Assessment of microcirculation in the diabetic foot with laser speckle contrast imaging. 40(6):065002, jul 2019.
 - [8] B. S. Lertsakdadet, G. T. Kennedy, R. Stone, C. Kowalczewski, A. C. Kowalczewski, S. Natesan, R. J. Christy, A. J. Durkin, and B. Choi. Assessing multimodal optical imaging of perfusion in burn wounds. *Burns*, 2021.
 - [9] K. Zheng, E. Middelkoop, M. Stoop, P. van Zuijlen, and A. Pijpe. Validity of laser speckle contrast imaging for the prediction of burn wound healing potential. *Burns*, 48(2):319–327, 2022.
 - [10] G. J. Richards and J. D. Briers. Capillary-blood-flow monitoring using laser speckle contrast analysis (LASCA): improving the dynamic range. In V. V. Tuchin, H. P. M.D., and B. Ovrin, editors, *Coherence Domain Optical Methods in Biomedical Science and Clinical Applications*, volume 2981, pages 160 – 171. International Society for Optics and Photonics, SPIE, 1997.
 - [11] A. B. Parthasarathy, W. J. Tom, A. Gopal, X. Zhang, and A. K. Dunn. Robust flow measurement with multi-exposure speckle imaging. *Optics Express*, 16(3):1975–1989, 2008.
 - [12] T. Dragojević, D. Bronzi, H. M. Varma, C. P. Valdes, C. Castellvi, F. Villa, A. Tosi, C. Justicia, F. Zappa, and T. Durduran. High-speed multi-exposure laser speckle contrast imaging with a single-photon counting camera. *Biomedical Optics Express*, 6(8):2865–2876, 2015.
 - [13] L. Song and D. S. Elson. Effect of signal intensity and camera quantization on laser speckle contrast analysis. *Biomed. Opt. Express*, 4(1):89–104, Jan 2013.
 - [14] S. Sunil, S. Zilpelwar, D. A. Boas, and D. D. Postnov. Guidelines for obtaining an absolute blood flow index with laser speckle contrast imaging. *bioRxiv*, 2021.
 - [15] M. Siket, I. Jánoki, K. Demeter, M. Szabó, and P. Földesy. Time varied illumination laser speckle contrast imaging. *Opt. Lett.*, 46(4):713–716, Feb 2021.
 - [16] M. Hultman, M. Larsson, T. Strömberg, and I. Fredriksson. Real-time video-rate perfusion imaging using multi-exposure laser speckle contrast imaging and machine learning. *Journal of Biomedical Optics*, 25(11):116007, 2020.
 - [17] P. Földesy, M. Siket, Ádám Nagy, and I. Jánoki. Correction of overexposure in laser speckle contrast imaging. *Opt. Express*, 30(12):21523–21534, Jun 2022.
 - [18] C. Wang, Z. Cao, X. Jin, W. Lin, Y. Zheng, B. Zeng, and M. Xu. Robust quantitative single-exposure laser speckle imaging with true flow speckle contrast in the temporal and spatial domains. *Biomed. Opt. Express*, 10(8):4097–4114, Aug 2019.

- [19] M. Ikeda. Switching characteristics of laser diode switch. *IEEE Journal of Quantum Electronics*, 19(2):157–164, 1983.
- [20] C. Inc. Comsol, 2020. <http://www.comsol.com/products/multiphysics/>.
- [21] Y. Zhao, K. Wang, W. Li, H. Zhang, Z. Qian, and Y. Liu. Laser speckle contrast imaging system using nanosecond pulse laser source. *Journal of Biomedical Optics*, 25(05):1–10, 2020.
- [22] O. Thompson, M. Andrews, and E. Hirst. Correction for spatial averaging in laser speckle contrast analysis. *Biomed. Opt. Express*, 2(4):1021–1029, Apr 2011.
- [23] H. Khan and J. K. Tar. On the Implementation of Fixed Point Iteration-based Adaptive Receding Horizon Control for Multiple Degree of Freedom, Higher Order Dynamical Systems. *Acta Polytechnica Hungarica*, 16(9):20, 2019.
- [24] A. J. Babqi and B. Alamri. A Comprehensive Comparison between Finite Control Set Model Predictive Control and Classical Proportional-Integral Control for Grid-tied Power Electronics Devices. *Acta Polytechnica Hungarica*, 18(7):67–87, 2021.
- [25] R.-C. Roman, R.-E. Precup, E.-L. Hedrea, S. Preitl, I. A. Zamfirache, C.-A. Bojan-Dragos, and E. M. Petriu. Iterative Feedback Tuning Algorithm for Tower Crane Systems. *Procedia Computer Science*, 199:157–165, 2022.
- [26] H. Redjimi and J. K. Tar. Multiple Components Fixed Point Iteration in the Adaptive Control of Single Variable 2nd Order Systems. *Acta Polytechnica Hungarica*, 18(9):69–86, 2021.
- [27] S. G. Johnson. The nlopt nonlinear-optimization package. <https://github.com/stevengj/nlopt>.

Integrated Force/Motion Trajectory Design of Parallel Robots for Singularity Robustness during Contact Tasks


Mustafa Özdemir^{1,*} and Sıtkı Kemal İder²

¹ Department of Mechanical Engineering, Faculty of Engineering, Marmara University, Recep Tayyip Erdoğan Campus, 34854 Maltepe, İstanbul, Türkiye
E-mail: mustafa.ozdemir@marmara.edu.tr

ORCID iD:  <https://orcid.org/0000-0002-4981-9573>

* Corresponding author

² Department of Mechanical Engineering, Faculty of Engineering, Çankaya University, 06790 Etimesgut, Ankara, Türkiye
E-mail: kider@cankaya.edu.tr

ORCID iD:  <https://orcid.org/0000-0001-5869-893X>

Abstract: Parallel robots have an increasing use in industrial and medical applications. Many of these applications require the execution of contact tasks. However, parallel robots possess drive singularities, which act as invisible barriers inside their workspace. In this paper, we develop an integrated force and motion trajectory planning method for removing drive singularities of parallel robots which perform contact tasks. The method is based on satisfaction of a consistency condition at the singularity, which is stated in terms of the generalized velocities, accelerations and contact forces, provided that the derivative of the associated determinant with respect to time does not simultaneously vanish. It is shown that, in the presence of singularity crossing, either the motion or the force trajectory can be arbitrarily chosen while the other is planned to satisfy the necessary conditions.

Keywords: parallel robot; contact task; motion trajectory; force trajectory; drive singularity; singularity removal

1 Introduction

In order to increase profitability and market share, firms seek ways to improve efficiency and competitiveness in their manufacturing processes. The use of robotics emerges as one of the most effective tools for achieving these goals [1, 2], especially at an accelerated pace in the era of Industry 4.0 [3-6]. Another area where the utilization of robots is of vital importance is medicine [7, 8]. Robotic-

assisted systems enable to perform high-precision surgical operations with minimal invasion [9, 10].

Conventional robotic arms have a serial kinematic architecture. However, parallel robots offer better accuracy and precision, increased rigidity, larger load capacity, lower inertias, and higher accelerations and speeds compared to the serial ones [11-14]. Owing to these advantages they have received attention as motion simulators [11, 15]. In addition, they have been widely applied to various industrial purposes, including, but not limited to, pick-and-place tasks [16], welding applications [17], spray-painting [18], and machining operations [12, 19]. They have been also increasingly used as medical and surgical robots [20].

One of the most serious handicaps of parallel robots is the limitation in the usability of their workspace due to the “type II singularity” loci within it [21]. The actuator forces tend to infinity in magnitude near singularities of this type. Due to this fact, they are alternatively called “drive singularities” [22, 23].

Since their avoidance during path planning would confine the robot to only a small portion of the workspace, there has been a growing interest in developing different methods for dealing with drive singularities. The approaches in this regard fall into two main categories. The first of these is to use actuation redundancy, which is well known in the literature to decrease or eliminate singularities [24]. The second one focuses on nonredundant parallel robots with an aim to obtain bounded inverse dynamics solutions near singularities.

This relatively new second approach enables parallel robots to pass in a controllable fashion through drive singular configurations and hence to use their entire workspace at no extra cost. The motion trajectory must be planned to sustain the consistency of the equations of motion at the singular configuration to be passed through [23, 25, 26]. The necessary “consistency conditions” were derived by Ider [23], and Briot and Arakelian [26] with different physical interpretations. As another recent effort in this regard, Ozgoren [27] obtained a similar condition by using the virtual work method.

However, as shown by Özdemir [28-30] there exist also “high-order” or “hyper-” drive singularities where boundedness of the inverse dynamics solution cannot be guaranteed only via the said consistency considerations. Özdemir [28] proved that time derivatives of the vanishing determinant should also be taken into account for removal of drive singularities. In accordance with this desingularization principle, Özdemir and İder [31] developed a motion trajectory planning method for flexible-joint parallel robots.

Although there are a number of unconstrained motion tasks such as pick-and-place and spray-painting operations [32-34], numerous advanced applications require the end-effector to move along a prescribed trajectory on a constraint surface while exerting a specified contact force profile onto that constraint surface. Some typical examples of these constrained motion tasks are machining processes (e.g.,

cutting, grinding, deburring), assembly operations and surgical procedures [32-37]. Indeed, it is essential to control the contact forces for executing such interaction tasks. Thus, in the last decade there has been a considerable research focus on force/position control of parallel robots [38-43].

However, in the previous studies on parallel robots performing contact tasks, singularity crossing problem is not considered. In order to fill this gap in the literature, the aim of the present paper is to develop an integrated force and motion trajectory planning method for enabling parallel robots to perform contact tasks in the presence of drive singularities. A condition is formulated for ensuring the consistency of the motion and force trajectories at the singularity. In accordance with the literature [28-30], the occurrence of high-order singularities is also prevented. Hence, boundedness of the actuator torques and forces near the singularity time of the contact task is guaranteed. We believe that the current study will facilitate the prevalence of parallel robots in industrial and medical applications.

2 Mathematical Modelling of a Parallel Robot Performing a Contact Task

Consider an n degree-of-freedom rigid-link rigid-joint parallel robot with n actuators. For modelling purposes, let this robot be transformed into an $m > n$ degree-of-freedom tree-like open system by cutting some passive joints. We denote the vector of the tree-like system's joint variables by $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \dots \ \theta_m]^T$. By reconnecting the cut joints, the loop-closure equations can be written as

$$f_i(\boldsymbol{\theta}) = 0, \quad i = 1, 2, \dots, m - n \quad (1)$$

Differentiating equations (1) with respect to time t and rearranging into matrix form, we get

$$\mathbf{A}\dot{\boldsymbol{\theta}} = \mathbf{0} \quad (2)$$

where the elements of the $(m - n) \times m$ matrix $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta})$ are given by

$$A_{ij} = \frac{\partial f_i}{\partial \theta_j}, \quad i = 1, 2, \dots, m - n, \quad j = 1, 2, \dots, m \quad (3)$$

Assuming that the environment is stationary and rigid, the constraints due to the contact of the end-effector with the environment can be expressed as

$$g_u(\boldsymbol{\theta}) = 0, \quad u = 1, 2, \dots, k \quad (4)$$

where k is the number of contact constraints such that $k < n$. These contact constraints can be written at velocity level as

$$\mathbf{B}\dot{\boldsymbol{\theta}} = \mathbf{0} \quad (5)$$

where the elements of the $k \times m$ matrix $\mathbf{B} = \mathbf{B}(\boldsymbol{\theta})$ are

$$B_{uj} = \frac{\partial g_u}{\partial \theta_j}, \quad u = 1, 2, \dots, k, \quad j = 1, 2, \dots, m \quad (6)$$

It is worth mentioning that the contact constraints considered here are equality constraints. This is because the robot is assumed not to lose its contact with the constraint surface during the whole task [44].

Selecting $\boldsymbol{\theta}$ as the vector of generalized coordinates, neglecting the impact and friction effects, and using the Lagrangian method, the equations of constrained motion of the parallel robot can be obtained in the following form:

$$\mathbf{M}\ddot{\boldsymbol{\theta}} + \mathbf{N} = \mathbf{T} + \mathbf{A}^T \boldsymbol{\lambda} + \mathbf{B}^T \boldsymbol{\mu} \quad (7)$$

where $\mathbf{M} = \mathbf{M}(\boldsymbol{\theta})$ is the $m \times m$ generalized mass matrix, $\mathbf{N} = \mathbf{N}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})$ the m -dimensional vector of generalized Coriolis, centrifugal and gravity forces, and \mathbf{T} the m -dimensional vector of nonconservative generalized forces applied by the actuators. In these equations, $\boldsymbol{\lambda} = [\lambda_1 \ \lambda_2 \ \dots \ \lambda_{m-n}]^T$ is the vector of the Lagrange multipliers associated with the loop-closure constraints, whereas $\boldsymbol{\mu} = [\mu_1 \ \mu_2 \ \dots \ \mu_k]^T$ is the vector of the Lagrange multipliers due to the contact constraints.

Let $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T$ be the vector of independent motion variables of the end-effector when it is in free motion. However, only $n-k$ of them can be arbitrarily prescribed along the contact surface. In other words, the number of motion degrees of freedom reduces to $n-k$. Denote by $\mathbf{y} = [y_1 \ y_2 \ \dots \ y_{n-k}]^T$ the vector of independent variables of contact motion, which are related to the joint variables by

$$y_v = h_v(\boldsymbol{\theta}), \quad v = 1, 2, \dots, n-k \quad (8)$$

By taking the time derivative of equations (8) and putting into matrix form, we obtain

$$\dot{\mathbf{y}} = \mathbf{C}\dot{\boldsymbol{\theta}} \quad (9)$$

where the elements of the $(n-k) \times m$ matrix $\mathbf{C} = \mathbf{C}(\boldsymbol{\theta})$ are given by

$$C_{vj} = \frac{\partial h_v}{\partial \theta_j}, \quad v = 1, 2, \dots, n-k, \quad j = 1, 2, \dots, m \quad (10)$$

3 Conditions for Singularity Robust Driving During a Contact Task

The vector $\boldsymbol{\theta}$ can be assumed to be constructed such that its first n elements are the actuated joint variables. Notice that this assumption yields no loss of generality since it deals only with ordering of the vector elements. Under this assumption, the vector \mathbf{T} is of the form

$$\mathbf{T} = \begin{bmatrix} \boldsymbol{\tau} \\ \mathbf{0} \end{bmatrix} \quad (11)$$

where τ_w denotes the generalized actuator force that is associated with the generalized coordinate θ_w ($w = 1, 2, \dots, n$) and $\boldsymbol{\tau} = [\tau_1 \quad \tau_2 \quad \dots \quad \tau_n]^T$. The above form of \mathbf{T} suggests the following partitioning of the matrices \mathbf{M} , \mathbf{A} , \mathbf{B} and the vector \mathbf{N} :

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{n \times m}^a \\ \mathbf{M}_{(m-n) \times m}^u \end{bmatrix} \quad (12)$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{(m-n) \times n}^a & \mathbf{A}_{(m-n) \times (m-n)}^u \end{bmatrix} \quad (13)$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{k \times n}^a & \mathbf{B}_{k \times (m-n)}^u \end{bmatrix} \quad (14)$$

$$\mathbf{N} = \begin{bmatrix} \mathbf{N}_{n \times 1}^a \\ \mathbf{N}_{(m-n) \times 1}^u \end{bmatrix} \quad (15)$$

where the sizes of the submatrices and subvectors are shown as subscripts. Then, using equations (11)-(15), one can rewrite equation (7) in the following form:

$$\boldsymbol{\tau} = \mathbf{M}^a \ddot{\boldsymbol{\theta}} + \mathbf{N}^a - (\mathbf{A}^a)^T \boldsymbol{\lambda} - (\mathbf{B}^a)^T \boldsymbol{\mu} \quad (16)$$

$$(\mathbf{A}^u)^T \boldsymbol{\lambda} = \mathbf{M}^u \ddot{\boldsymbol{\theta}} + \mathbf{N}^u - (\mathbf{B}^u)^T \boldsymbol{\mu} \quad (17)$$

In order to determine the joint motions required for a given motion trajectory $\mathbf{y}(t)$ along the constraint surface, equations (2), (5) and (9) can be merged into the following equation:

$$\mathbf{D}\dot{\boldsymbol{\theta}} = \mathbf{z} \quad (18)$$

where $\mathbf{D} = \mathbf{D}(\boldsymbol{\theta})$ is an $m \times m$ matrix constructed as

$$\mathbf{D} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \end{bmatrix} \quad (19)$$

and $\mathbf{z} = \mathbf{z}(t)$ is an m -dimensional vector defined as

$$\mathbf{z} = \begin{bmatrix} \mathbf{0} \\ \dot{\mathbf{y}} \end{bmatrix} \quad (20)$$

As long as \mathbf{D} is nonsingular, equation (18) constitutes a system of m first-order differential equations that can be rewritten as

$$\dot{\boldsymbol{\theta}} = \mathbf{D}^{-1}\mathbf{z} \quad (21)$$

By time-differentiating equation (18) and rearranging, one can also get

$$\ddot{\boldsymbol{\theta}} = \mathbf{D}^{-1}(-\dot{\mathbf{D}}\dot{\boldsymbol{\theta}} + \dot{\mathbf{z}}) \quad (22)$$

Thus, once the system of equation (21) is solved for the generalized coordinates $\boldsymbol{\theta}(t)$ and the generalized velocities $\dot{\boldsymbol{\theta}}(t)$ by a suitable numerical integration method, the generalized accelerations $\ddot{\boldsymbol{\theta}}(t)$ can be computed from equation (22).

Substituting the calculated $\boldsymbol{\theta}$, $\dot{\boldsymbol{\theta}}$ and $\ddot{\boldsymbol{\theta}}$ together with the given force trajectory $\boldsymbol{\mu}(t)$ into equation (17) and solving for $\boldsymbol{\lambda}(t)$ gives the following equation, provided that \mathbf{A}^u is nonsingular:

$$\boldsymbol{\lambda} = (\mathbf{A}^u)^{-T} \left[\mathbf{M}^u \ddot{\boldsymbol{\theta}} + \mathbf{N}^u - (\mathbf{B}^u)^T \boldsymbol{\mu} \right] \quad (23)$$

Finally, the required actuator forces can be computed from equation (16).

During the implementation of the above procedure, inverse kinematic singularities occur when the determinant of the \mathbf{D} matrix becomes zero. However, such singularities are in general on the boundaries of the workspace [21]. Therefore, they are not a major concern and are left out of the scope of this study.

Drive singularities arise when the determinant of the \mathbf{A}^u matrix vanishes. As it is apparent from equation (23), the Lagrange multipliers associated with the loop-closure constraints grow without bounds in the neighborhood of such singularities. Let us assume that \mathbf{A}^u is rank deficient by one at the drive singular configuration to be passed through. This assumption is quite realistic since higher rank deficiencies would be rather rare in practice [23, 25]. By writing $(\mathbf{A}^u)^{-T}$ in terms of the adjoint matrix and determinant of $(\mathbf{A}^u)^T$, equation (23) can be reexpressed as

$$\lambda = \frac{1}{\det\left(\left(\mathbf{A}^u\right)^T\right)} \text{adj}\left(\left(\mathbf{A}^u\right)^T\right) \left[\mathbf{M}^u \ddot{\boldsymbol{\theta}} + \mathbf{N}^u - \left(\mathbf{B}^u\right)^T \boldsymbol{\mu} \right] \quad (24)$$

or, recalling that $\det\left(\left(\mathbf{A}^u\right)^T\right) = \det\left(\mathbf{A}^u\right)$ and $\text{adj}\left(\left(\mathbf{A}^u\right)^T\right) = \left(\text{adj}\left(\mathbf{A}^u\right)\right)^T$,

$$\lambda = \frac{1}{\det\left(\mathbf{A}^u\right)} \left(\text{adj}\left(\mathbf{A}^u\right)\right)^T \left[\mathbf{M}^u \ddot{\boldsymbol{\theta}} + \mathbf{N}^u - \left(\mathbf{B}^u\right)^T \boldsymbol{\mu} \right] \quad (25)$$

By inspecting equation (25), the condition that should be satisfied for the dynamic equations of the robot to be consistent at a drive singularity can be stated as

$$\left(\text{adj}\left(\mathbf{A}^u\right)\right)^T \left[\mathbf{M}^u \ddot{\boldsymbol{\theta}} + \mathbf{N}^u - \left(\mathbf{B}^u\right)^T \boldsymbol{\mu} \right] = \mathbf{0} \quad (26)$$

If this consistency condition holds and the first-order time derivative of the determinant $\det\left(\mathbf{A}^u\right)$ does not vanish at the singularity time t_s (i.e., the singularity is not of high order), then it follows from l'Hôpital's Rule that $\lim_{t \rightarrow t_s} \lambda_i(t)$ is finite for all $i = 1, 2, \dots, m - n$, which further implies that the required actuator forces will remain bounded as the singularity is approached. However, the inverse dynamics solution is still indeterminate at time t_s since satisfaction of the condition given by equation (26) for maintaining the consistency of the robot's dynamic equations yields $0/0$ in equation (25) for all $\lambda_i(t_s)$. This indeterminacy can be removed by setting

$$\lambda_i(t_s) = \lim_{t \rightarrow t_s} \lambda_i(t), \quad i = 1, 2, \dots, m - n \quad (27)$$

where the limits $\lim_{t \rightarrow t_s} \lambda_i(t)$ are evaluated via l'Hôpital's Rule.

If the consistency condition given by equation (26) is not satisfied, then at least one of the limits $\lim_{t \rightarrow t_s} \lambda_i(t)$ is not finite, which yields an unbounded growth of the inverse dynamics solution in the vicinity of the singularity. Besides, it is useful to note that among the $m - n$ individual equations of the matrix equation (26), only one is linearly independent. This is due to the fact [45] that the adjoint matrix of a matrix that is rank deficient by one has rank one.

4 Case Study

In this section, the application of the developed method is exemplified by considering the planar 5R parallel robot, which is shown in Figure 1. Link 0 is the fixed link. Each moving link p ($p = 1, 2, 3, 4$) has mass m_p , mass center at G_p , and centroidal moment of inertia I_{G_p} . The link lengths are denoted by $L_0 = |R_1R_2|$, $L_1 = |R_1R_3|$, $L_2 = |R_2R_4|$, $L_3 = |R_3R_5|$ and $L_4 = |R_4R_5|$. The locations of the mass centers are given by $r_1 = |R_1G_1|$, $\alpha_1 = \angle R_3R_1G_1$, $r_2 = |R_2G_2|$, $\alpha_2 = \angle R_4R_2G_2$, $r_3 = |R_3G_3|$, $\alpha_3 = \angle R_5R_3G_3$, $r_4 = |R_4G_4|$ and $\alpha_4 = \angle R_5R_4G_4$. The origin of the fixed rectangular coordinate system xy is at joint R_1 . The gravitational acceleration g acts in the negative y -direction. The robot has two degrees of freedom and is actuated by two motors that are located at joints R_1 and R_2 . The endpoint P is given by $b = |R_3P|$ and $\beta = \angle R_5R_3P$.

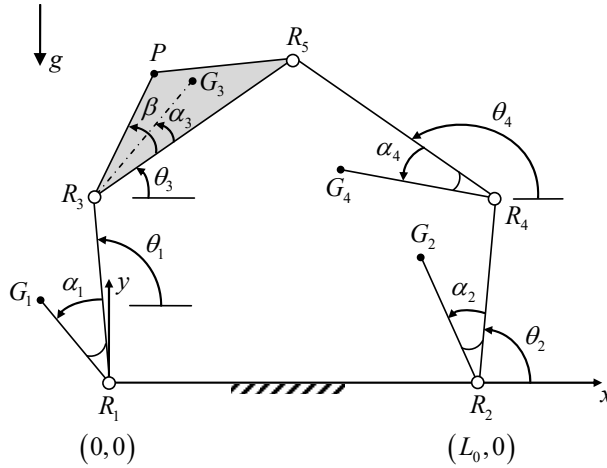


Figure 1
The considered robot

By virtually cutting the closed-loop at joint R_5 and choosing the generalized coordinates vector as $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \theta_3 \ \theta_4]^T$, the generalized mass matrix and the vector of generalized nonlinear inertial and gravity forces of the resulting open-loop system can be obtained as

$$\mathbf{M} = \begin{bmatrix} M_{11} & 0 & M_{13} & 0 \\ 0 & M_{22} & 0 & M_{24} \\ M_{31} & 0 & M_{33} & 0 \\ 0 & M_{42} & 0 & M_{44} \end{bmatrix} \quad (28)$$

$$\mathbf{N} = \begin{bmatrix} N_1 \\ N_2 \\ N_3 \\ N_4 \end{bmatrix} \quad (29)$$

where

$$M_{11} = m_1 r_1^2 + I_{G_1} + m_3 L_1^2 \quad (30)$$

$$M_{13} = M_{31} = m_3 r_3 L_1 \cos(\theta_1 - \theta_3 - \alpha_3) \quad (31)$$

$$M_{22} = m_2 r_2^2 + I_{G_2} + m_4 L_2^2 \quad (32)$$

$$M_{24} = M_{42} = m_4 r_4 L_2 \cos(\theta_2 - \theta_4 - \alpha_4) \quad (33)$$

$$M_{33} = m_3 r_3^2 + I_{G_3} \quad (34)$$

$$M_{44} = m_4 r_4^2 + I_{G_4} \quad (35)$$

$$N_1 = m_3 r_3 L_1 \dot{\theta}_3^2 \sin(\theta_1 - \theta_3 - \alpha_3) + m_1 g r_1 \cos(\theta_1 + \alpha_1) + m_3 g L_1 \cos \theta_1 \quad (36)$$

$$N_2 = m_4 r_4 L_2 \dot{\theta}_4^2 \sin(\theta_2 - \theta_4 - \alpha_4) + m_2 g r_2 \cos(\theta_2 + \alpha_2) + m_4 g L_2 \cos \theta_2 \quad (37)$$

$$N_3 = -m_3 r_3 L_1 \dot{\theta}_1^2 \sin(\theta_1 - \theta_3 - \alpha_3) + m_3 g r_3 \cos(\theta_3 + \alpha_3) \quad (38)$$

$$N_4 = -m_4 r_4 L_2 \dot{\theta}_2^2 \sin(\theta_2 - \theta_4 - \alpha_4) + m_4 g r_4 \cos(\theta_4 + \alpha_4) \quad (39)$$

Then, by reconnecting joint R_5 , the loop-closure equations can be written as

$$f_1(\boldsymbol{\theta}) = L_1 \cos \theta_1 + L_3 \cos \theta_3 - L_0 - L_2 \cos \theta_2 - L_4 \cos \theta_4 = 0 \quad (40)$$

$$f_2(\boldsymbol{\theta}) = L_1 \sin \theta_1 + L_3 \sin \theta_3 - L_2 \sin \theta_2 - L_4 \sin \theta_4 = 0 \quad (41)$$

Thus, the Jacobian matrix of the loop-closure constraint equations is

$$\mathbf{A} = \begin{bmatrix} -L_1 \sin \theta_1 & L_2 \sin \theta_2 & -L_3 \sin \theta_3 & L_4 \sin \theta_4 \\ L_1 \cos \theta_1 & -L_2 \cos \theta_2 & L_3 \cos \theta_3 & -L_4 \cos \theta_4 \end{bmatrix} \quad (42)$$

As illustrated in Figure 2, let the constrained-motion task of the robot be moving the endpoint P according to a prescribed trajectory $x_p(t)$ on the frictionless plane surface given by $y = y^*$ while simultaneously applying a specified normal contact force $\mu(t)$ onto it. The surface is rigid and fixed in space. The robot will be in contact with the surface only at point P throughout the entire duration, t_f , of the task.

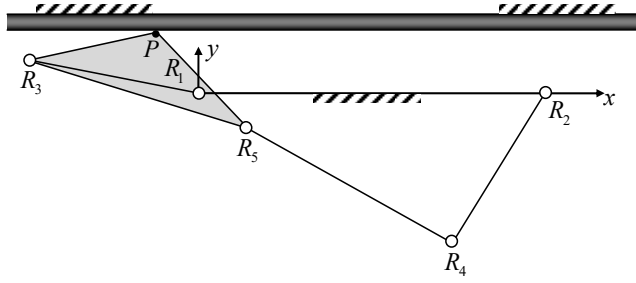


Figure 2
The given contact task

The motion trajectory can be expressed as

$$x_p(t) = x_0 + (x_f - x_0)s(t) \quad (43)$$

where x_0 and x_f are the initial and final values of the x -coordinate of point P on the constraint surface, and $s(t)$ is a timing function, which is chosen to be the following quintic polynomial in order to have zero initial and final velocities and accelerations:

$$s(t) = 6\left(\frac{t}{t_f}\right)^5 - 15\left(\frac{t}{t_f}\right)^4 + 10\left(\frac{t}{t_f}\right)^3 \quad (44)$$

Let the force trajectory $\mu(t)$ be trapezoidal. The desired constant value of the contact force in the plateau phase is μ^* . The force will be linearly increased from zero to this value in the first Δt time units of motion and will be decreased linearly back to zero in the last Δt time units. Thus,

$$\mu(t) = \begin{cases} \frac{t}{\Delta t} \mu^*, & 0 \leq t < \Delta t \\ \mu^*, & \Delta t \leq t < t_f - \Delta t \\ \frac{t_f - t}{\Delta t} \mu^*, & t_f - \Delta t \leq t \leq t_f \end{cases} \quad (45)$$

The surface contact constraint can be expressed in the task space as

$$g_1(x_p, y_p) = y^* - y_p = 0 \quad (46)$$

or in the joint space as

$$g_1(\boldsymbol{\theta}) = y^* - L_1 \sin \theta_1 - b \sin(\theta_3 + \beta) = 0 \quad (47)$$

Thus, the vector of generalized constraint forces acting on the robot due to its contact with the surface at point P is

$$\mathbf{F}_c = \mathbf{B}^T \mu \quad (48)$$

where

$$\mathbf{B} = [-L_1 \cos \theta_1 \quad 0 \quad -b \cos(\theta_3 + \beta) \quad 0] \quad (49)$$

The motion variable x_p can be related to the joint variables as

$$x_p = h_1(\boldsymbol{\theta}) = L_1 \cos \theta_1 + b \cos(\theta_3 + \beta) \quad (50)$$

Then, it follows from equations (10) that

$$\mathbf{C} = [-L_1 \sin \theta_1 \quad 0 \quad -b \sin(\theta_3 + \beta) \quad 0] \quad (51)$$

After constructing the matrix \mathbf{D} given by equation (19), the condition for the occurrence of an inverse kinematic singularity can be expressed as follows:

$$\det(\mathbf{D}) = L_1 L_2 L_4 b \sin(\theta_1 - \theta_3 - \beta) \sin(\theta_4 - \theta_2) = 0 \quad (52)$$

With $n = 2$, $m = 4$ and $k = 1$, the \mathbf{M} , \mathbf{A} and \mathbf{B} matrices and the \mathbf{N} vector are partitioned according to equations (12)-(15) as given below:

$$\mathbf{M}^a = \begin{bmatrix} M_{11} & 0 & M_{13} & 0 \\ 0 & M_{22} & 0 & M_{24} \end{bmatrix} \quad (53)$$

$$\mathbf{M}^u = \begin{bmatrix} M_{31} & 0 & M_{33} & 0 \\ 0 & M_{42} & 0 & M_{44} \end{bmatrix} \quad (54)$$

$$\mathbf{A}^a = \begin{bmatrix} -L_1 \sin \theta_1 & L_2 \sin \theta_2 \\ L_1 \cos \theta_1 & -L_2 \cos \theta_2 \end{bmatrix} \quad (55)$$

$$\mathbf{A}^u = \begin{bmatrix} -L_3 \sin \theta_3 & L_4 \sin \theta_4 \\ L_3 \cos \theta_3 & -L_4 \cos \theta_4 \end{bmatrix} \quad (56)$$

$$\mathbf{B}^a = [-L_1 \cos \theta_1 \quad 0] \quad (57)$$

$$\mathbf{B}^u = [-b \cos(\theta_3 + \beta) \quad 0] \quad (58)$$

$$\mathbf{N}^a = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} \quad (59)$$

$$\mathbf{N}^u = \begin{bmatrix} N_3 \\ N_4 \end{bmatrix} \quad (60)$$

Then the equation describing the drive singularity locus in the joint space can be obtained as

$$\det(\mathbf{A}^u) = L_3 L_4 \sin(\theta_3 - \theta_4) = 0 \quad (61)$$

Readers can be referred to numerous studies [29, 46-48] for more details on the singularities and workspace of the planar 5R mechanism.

The values of the model parameters are chosen as follows: $L_0 = 3$ m, $L_1 = L_2 = 1.5$ m, $L_3 = L_4 = 2$ m, $b = 1$ m, $\beta = (\pi/6)$ rad, $m_1 = m_2 = 0.4$ kg, $m_3 = m_4 = 0.6$ kg, $r_1 = r_2 = 0.75$ m, $r_3 = 1.5$ m, $r_4 = 1$ m, $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = \alpha_4 = (2\pi/3)$ rad, $I_{G_1} = I_{G_2} = 0.2$ kg·m², $I_{G_3} = I_{G_4} = 0.3$ kg·m².

The gravitational acceleration is assumed to be $g = 9.807$ m/s². The constraint surface is taken as $y^* = 0.5$ m. The x -coordinates of the starting and ending positions of the endpoint P along this surface are $x_0 = -0.5$ m and $x_f = -0.42$ m, respectively. The total duration of the task is selected to be $t_f = 2$ s. Thus, the timing function becomes

$$s(t) = 0.1875t^5 - 0.9375t^4 + 1.25t^3 \quad (62)$$

The motion trajectory obtained using equation (43) is shown in Figure 3. The values of the joint variables at the starting position of the endpoint are as follows: $\theta_1 = 169.4^\circ$, $\theta_2 = 237.5^\circ$, $\theta_3 = 343.0^\circ$, $\theta_4 = 151.5^\circ$. In the case studies presented below, the numerical integrations are based on the Dormand-Prince formula (see, e.g., [49] and the references cited therein) with a fixed step size of 0.002 s.

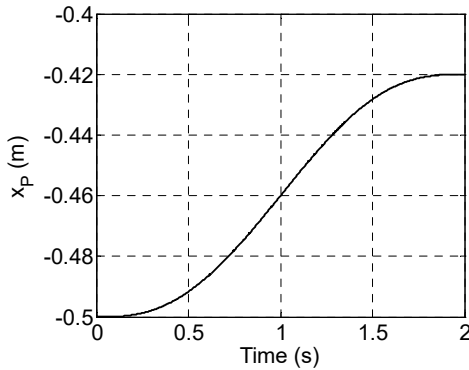


Figure 3

The desired motion trajectory

4.1 Case 1: A Case Where Motion and Force Trajectories Are Not Consistent with Each Other at the Singularity

As the first case, let the force trajectory be generated with $\mu^* = 1 \text{ N}$ and $\Delta t = 0.2 \text{ s}$, as shown in Figure 4.

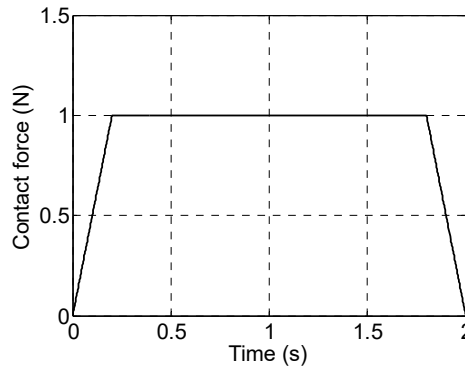


Figure 4

A force trajectory that is not consistent with the motion trajectory at the singularity

The inverse kinematic solution is singularity free. The time variations of the joint angular displacements that correspond to the prescribed motion trajectory of the endpoint on the constraint surface are shown in Figure 5.

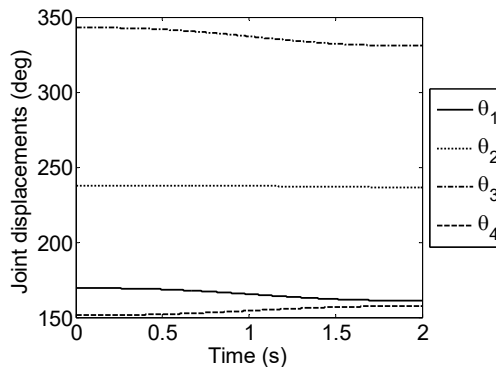


Figure 5

Time histories of the joint variables necessary for the desired constrained motion of the end-effector

However, although no inverse kinematic singularity is encountered, a drive singularity occurs when $x_p = -0.45 \text{ m}$ (i.e., $s = 0.65$). The values of the joint variables at this singular position are as follows: $\theta_1 = 164.2^\circ$, $\theta_2 = 237.4^\circ$, $\theta_3 = 335.3^\circ$, $\theta_4 = 155.3^\circ$. Both Lagrange multipliers become unbounded near this position. As can be read from Figure 6, the singularity time is $t_s = 1.164 \text{ s}$.

The limits of the required motor torques as t approaches this value are not finite, as can be seen in Figure 7.

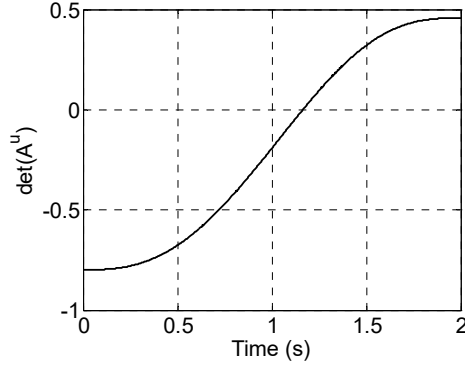


Figure 6

Time variation of the determinant whose vanishing implies the occurrence of a drive singularity

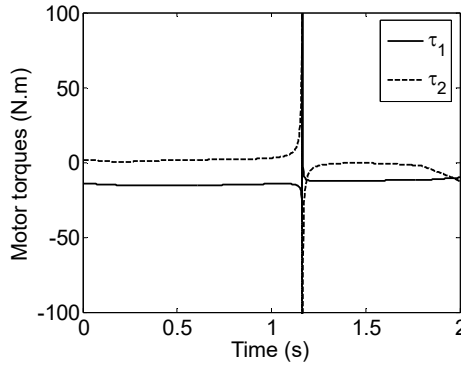


Figure 7

Motor torques required for the motion and force trajectories that are not consistent with each other at the singularity

4.2 Case 2: A Case Where Motion and Force Trajectories Are Consistent with Each Other at the Singularity

In order to overcome the unboundedness of the inverse dynamics solution, the motion and force trajectories should be such that the consistency of the dynamic equations is maintained at the singular configuration to be passed through. The consistency condition that should be satisfied at the singular position of interest can be derived as follows: The transpose of the adjoint matrix of \mathbf{A}^u is

$$\left(\text{adj}(\mathbf{A}^u)\right)^T = \begin{bmatrix} -L_4 \cos \theta_4 & -L_3 \cos \theta_3 \\ -L_4 \sin \theta_4 & -L_3 \sin \theta_3 \end{bmatrix} \quad (63)$$

Substituting equation (63) into equation (26) gives

$$-L_4 \cos \theta_4 \left[M_{31} \ddot{\theta}_1 + M_{33} \ddot{\theta}_3 + N_3 + \mu b \cos(\theta_3 + \beta) \right] - L_3 \cos \theta_3 \left(M_{42} \ddot{\theta}_2 + M_{44} \ddot{\theta}_4 + N_4 \right) = 0 \quad (64)$$

$$-L_4 \sin \theta_4 \left[M_{31} \dot{\theta}_1 + M_{33} \dot{\theta}_3 + N_3 + \mu b \cos(\theta_3 + \beta) \right] - L_3 \sin \theta_3 \left(M_{42} \dot{\theta}_2 + M_{44} \dot{\theta}_4 + N_4 \right) = 0 \quad (65)$$

The following relation in radians exists between θ_3 and θ_4 at the encountered singular configuration: $\theta_4 = \theta_3 - \pi$. Thus, at that singularity, $\cos \theta_4 = -\cos \theta_3$, $\sin \theta_4 = -\sin \theta_3$, the rank of $\left(\text{adj}(\mathbf{A}^u) \right)^T$ is one, and equations (64) and (65) are linearly dependent and can be reduced to

$$L_4 \left[M_{31} \ddot{\theta}_1 + M_{33} \ddot{\theta}_3 + N_3 + \mu b \cos(\theta_3 + \beta) \right] - L_3 \left(M_{42} \ddot{\theta}_2 + M_{44} \ddot{\theta}_4 + N_4 \right) = 0 \quad (66)$$

The above consistency condition can be satisfied at the singularity via a proper planning of either the motion or the force trajectory. The velocity- and acceleration-level inverse kinematic solutions are calculated at the encountered singularity as

$$\dot{\theta}_1(t_s) = -1.8461 \dot{x}_p(t_s) \quad (67)$$

$$\dot{\theta}_2(t_s) = -0.2892 \dot{x}_p(t_s) \quad (68)$$

$$\dot{\theta}_3(t_s) = -2.6766 \dot{x}_p(t_s) \quad (69)$$

$$\dot{\theta}_4(t_s) = 1.3388 \dot{x}_p(t_s) \quad (70)$$

$$\ddot{\theta}_1(t_s) = -4.4386 \left[\dot{x}_p(t_s) \right]^2 - 1.8461 \ddot{x}_p(t_s) \quad (71)$$

$$\ddot{\theta}_2(t_s) = -9.3659 \left[\dot{x}_p(t_s) \right]^2 - 0.2892 \ddot{x}_p(t_s) \quad (72)$$

$$\ddot{\theta}_3(t_s) = -4.3744 \left[\dot{x}_p(t_s) \right]^2 - 2.6766 \ddot{x}_p(t_s) \quad (73)$$

$$\ddot{\theta}_4(t_s) = 1.7241 \left[\dot{x}_p(t_s) \right]^2 + 1.3388 \ddot{x}_p(t_s) \quad (74)$$

In the above equations, the endpoint velocity is in m/s, endpoint acceleration is in m/s², angular joint velocities are in rad/s, and angular joint accelerations are in rad/s². Then, for the singular configuration of interest, we compute from equations (31) and (33)-(35) that $M_{31}(t_s) = 0.4854 \text{ kg} \cdot \text{m}^2$, $M_{33} = 1.65 \text{ kg} \cdot \text{m}^2$,

$M_{42}(t_s) = 0.7096 \text{ kg} \cdot \text{m}^2$, $M_{44} = 0.9 \text{ kg} \cdot \text{m}^2$, and from equations (38) and (39) that

$$N_3(t_s) = -4.2931[\dot{x}_p(t_s)]^2 - 0.8179 \quad (75)$$

$$N_4(t_s) = 0.0463[\dot{x}_p(t_s)]^2 + 0.5452 \quad (76)$$

where N_3 and N_4 are obtained in N·m for the endpoint velocity in m/s. By substituting these into equation (66), simplifying and multiplying both sides of the resulting equation by -1 , we get

$$12.6244\ddot{x}_p(t_s) + 17.2351[\dot{x}_p(t_s)]^2 - 1.9914\mu(t_s) + 2.7262 = 0 \quad (77)$$

Equation (77) shows that consistent values of the endpoint velocity, the endpoint acceleration and the contact force at the singularity time lie on a quadric surface as seen in Figure 8. In this figure, v_p and a_p represent the endpoint velocity and acceleration, respectively. It may be useful to note that $v_p = \dot{x}_p$ and $a_p = \ddot{x}_p$ since the endpoint moves along a constant y path.

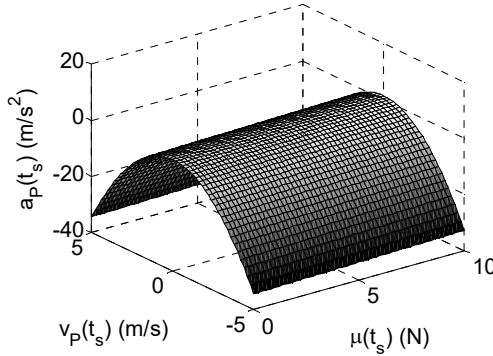


Figure 8

Consistent values of the endpoint velocity, the endpoint acceleration and the contact force at the singular point of interest

In this example, we prefer to plan the force trajectory for satisfying the consistency condition. Hence, the previously chosen desired motion trajectory becomes realizable, despite the presence of drive singularity. We first compute that $\dot{x}_p(t_s) = 0.0710 \text{ m/s}$ and $\ddot{x}_p(t_s) = -0.0478 \text{ m/s}^2$. Then, substituting these values into equation (77), we get $\mu(t_s) = 1.11 \text{ N}$. As our second case, we construct a new trapezoidal force trajectory with $\mu^* = 1.11 \text{ N}$ and $\Delta t = 0.2 \text{ s}$, as shown in Figure 9. With this new force trajectory, the dynamic equations are now

consistent at all times. Additionally, $\det(\mathbf{A}^u)$ has a nonzero first-order time derivative at the singularity time. Thus, the required motor torques remain bounded in the neighborhood of the singularity, as can be seen in Figure 10. It is useful to note that $\lim_{t \rightarrow t_s} \lambda_1(t) = 4.77 \text{ N}$ and $\lim_{t \rightarrow t_s} \lambda_2(t) = -1.93 \text{ N}$.

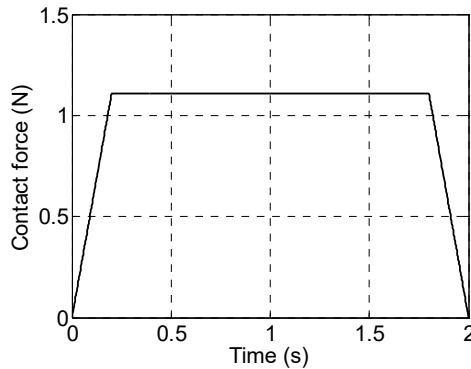


Figure 9

A force trajectory that is consistent with the motion trajectory at the singularity

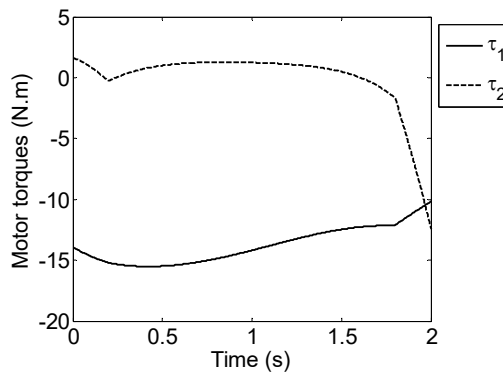


Figure 10

Motor torques required for the motion and force trajectories that are consistent with each other at the singularity

Conclusions

This paper presents an integrated force and motion trajectory planning approach for removing drive singularities of parallel robots performing contact tasks. Such tasks constitute the majority of the industrial and medical robotic applications. A consistency condition is derived in terms of the generalized velocities, accelerations and contact forces. This condition should be considered while planning the motion and force trajectories. Also, in accordance with the literature

[28-30], the singularity is prevented from being of high order. Thus, the boundedness of the inverse dynamics solution around the singularity is ensured.

The effectiveness of the proposed method is verified through a numerical case study where the planar 5R parallel robot is considered to perform a constrained motion task in the presence of drive singularities. It is shown that one of the motion and force trajectories can be arbitrarily chosen while the other is planned to satisfy the consistency condition at the singularity. The consistent values of the endpoint velocity, the endpoint acceleration and the contact force at the singularity are found to describe a quadric surface.

References

- [1] M. Smieszek, P. Dobrzanski and M. Dobrzanska. Comparison of the level of robotisation in Poland and selected countries, including social and economic factors. *Acta Polytechnica Hungarica*, 16(4):197-212, 2019
- [2] Z. Cséfalvay. Robotization in Central and Eastern Europe: catching up or dependence? *European Planning Studies*, 28(8):1534-1553, 2020
- [3] G. Haidegger and I. Paniti. Episodes of robotics and manufacturing automation achievements from the past decades and vision for the next decade. *Acta Polytechnica Hungarica*, 16(10):119-136, 2019
- [4] A. Okanović, B. Jokanović, V. Đaković, S. Vukadinović and J. Ješić. Innovating a model for measuring competitiveness in accordance with the challenges of Industry 4.0. *Acta Polytechnica Hungarica*, 17(7):67-88, 2020
- [5] W. S. Barbosa, M. M. Gioia, V. G. Natividade, R. F. F. Wanderley, M. R. Chaves, F. C. Gouvea and F. M. Gonçalves. Industry 4.0: examples of the use of the robotic arm for digital manufacturing processes. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 14(4):1569-1575, 2020
- [6] G. Fragapane, D. Ivanov, M. Peron, F. Sgarbossa and J. O. Strandhagen. Increasing flexibility and productivity in Industry 4.0 production networks with autonomous mobile robots and smart intralogistics. *Annals of Operations Research*, 308(1-2):125-143, 2022
- [7] Á. Takács, D. Á. Nagy, I. J. Rudas and T. Haidegger. Origins of surgical robotics: From space to the operating room. *Acta Polytechnica Hungarica*, 13(1):13-30, 2016
- [8] P. E. Dupont, B. J. Nelson, M. Goldfarb, B. Hannaford, A. Menciassi, M. K. O'Malley, N. Simaan, P. Valdastrì and G.-Z. Yang. A decade retrospective of medical robotics research from 2010 to 2020. *Science Robotics*, 6(60):eabi8017, 2021
- [9] M. J. Mack. Minimally invasive and robotic surgery. *JAMA-Journal of the American Medical Association*. 285(5):568-572, 2001

- [10] R. Nagyné Elek and T. Haidegger. Robot-assisted minimally invasive surgical skill assessment—manual and automated platforms. *Acta Polytechnica Hungarica*, 16(8):141-169, 2019
- [11] J.-P. Merlet. Parallel Robots, 2nd edition. In: Solid Mechanics and Its Applications, Volume 128, Series Editor: G. M. L. Gladwell, Springer, Dordrecht, 2006
- [12] Y. Jin, H. Chanal and F. Paccot. Parallel Robots. In: Handbook of Manufacturing Engineering and Technology, A. Y. C. Nee (Ed.), Springer, London, 2015, pp. 2091-2127
- [13] J. Somló, G. D. Varga, M. Zenkl and B. Mikó. The „Phantom” Delta robot A new device for parallel robot investigations. *Acta Polytechnica Hungarica*, 15(4):143-160, 2018
- [14] J. Somló. General triangle parallel robot (GTPR) Basic features of a new robot type - kinematics and related application issues. *Acta Polytechnica Hungarica*, 16(5):7-24, 2019
- [15] G. Liu, Z. Qu, J. Han and X. Liu. Systematic optimal design procedures for the Gough-Stewart platform used as motion simulators. *Industrial Robot*, 40(6):550-558, 2013
- [16] F. Bourbonnais, P. Bigras and I. A. Bonev. Minimum-time trajectory planning and control of a pick-and-place five-bar parallel robot. *IEEE/ASME Transactions on Mechatronics*, 20(2):740-749, 2015
- [17] J. Shi, Y. Wang, G. Zhang and H. Ding. Optimal design of 3-DOF PKM module for friction stir welding. *The International Journal of Advanced Manufacturing Technology*, 66(9-12):1879-1889, 2013
- [18] J. Wu, Y. Gao, B. Zhang and L. Wang. Workspace and dynamic performance evaluation of the parallel manipulators in a spray-painting equipment. *Robotics and Computer-Integrated Manufacturing*, 44:199-207, 2017
- [19] J. Jahanpour, M. Motallebi and M. Porghoveh. A novel trajectory planning scheme for parallel machining robots enhanced with NURBS curves. *Journal of Intelligent & Robotic Systems*, 82(2):257-275, 2016
- [20] H. Tian, C. Wang, X. Dang and L. Sun. A 6-DOF parallel bone-grinding robot for cervical disc replacement surgery. *Medical & Biological Engineering & Computing*, 55(12):2107-2121, 2017
- [21] C. Gosselin and J. Angeles. Singularity analysis of closed-loop kinematic chains. *IEEE Transactions on Robotics and Automation*, 6(3):281-290, 1990
- [22] S. K. Ider. Singularity robust inverse dynamics of planar 2-RPR parallel manipulators. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 218(7):721-730, 2004

-
- [23] S. K. Ider. Inverse dynamics of parallel manipulators in the presence of drive singularities. *Mechanism and Machine Theory*, 40(1):33-44, 2005
- [24] M. Luces, J. K. Mills and B. Benhabib. A review of redundant parallel kinematic mechanisms. *Journal of Intelligent & Robotic Systems*, 86(2):175-198, 2017
- [25] C. K. K. Jui and Q. Sun. Path tracking of parallel manipulators in the presence of force singularity. *ASME Journal of Dynamic Systems, Measurement, and Control*, 127(4):550-563, 2005
- [26] S. Briot and V. Arakelian. Optimal force generation in parallel manipulators for passing through the singular positions. *The International Journal of Robotics Research*, 27(8):967-983, 2008
- [27] M. K. Ozgoren. Kinematic and kinetostatic analysis of parallel manipulators with emphasis on position, motion, and actuation singularities. *Robotica*, 37(4):599-625, 2019
- [28] M. Özdemir. Removal of singularities in the inverse dynamics of parallel robots. *Mechanism and Machine Theory*, 107:71-86, 2017
- [29] M. Özdemir. High-order singularities of 5R planar parallel robots. *Robotica*, 37(2):233-245, 2019
- [30] M. Özdemir. Hypersingularities of 3-RRR planar parallel robots. *Proceedings of the Romanian Academy, Series A: Mathematics, Physics, Technical Sciences, Information Science*, 22(4):353-360, 2021
- [31] M. Özdemir and S. K. İder. Desingularization of flexible-joint parallel robots. *Acta Polytechnica Hungarica*, 18(6):85-106, 2021
- [32] G. de A. Barreto, A. F. R. Araújo and H. J. Ritter. Self-organizing feature maps for modeling and control of robotic manipulators. *Journal of Intelligent & Robotic Systems*, 36(4):407-450, 2003
- [33] M. Vukobratovic, V. Potkonjak and V. Matijevic. Dynamics of Robots with Contact Tasks. In: *International Series on Microprocessor-Based and Intelligent Systems Engineering*, Volume 26, Series Editor: S. G. Tzafestas, Springer, Dordrecht, 2003
- [34] M. Vukobratovic. Robot-environment dynamic interaction survey and future trends. *Journal of Computer and Systems Sciences International*, 49(2):329-342, 2010
- [35] E. Dombre, G. Duchemin, P. Poignet and F. Pierrot. Dermarob: A safe robot for reconstructive surgery. *IEEE Transactions on Robotics and Automation*, 19(5):876-884, 2003
- [36] N. Zemiti, G. Morel, T. Ortmaier and N. Bonnet. Mechatronic design of a new robot for force control in minimally invasive surgery. *IEEE/ASME Transactions on Mechatronics*, 12(2):143-153, 2007
-

- [37] A. Pappalardo, A. Albakri, C. Liu, L. Bascetta, E. De Momi and P. Poignet. Hunt–Crossley model based force control for minimally invasive robotic surgery. *Biomedical Signal Processing and Control*, 29:31-43, 2016
- [38] S. Bellakehal, N. Andreff, Y. Mezouar and M. Tadjine. Force/position control of parallel robots using exteroceptive pose measurements. *Meccanica*, 46(1):195-205, 2011
- [39] M. Madani and M. Moallem. Hybrid position/force control of a flexible parallel manipulator. *Journal of the Franklin Institute*, 348(6):999-1012, 2011
- [40] B. Achili, B. Daachi, Y. Amirat, A. Ali-Cherif and M. E. Daâchi. A stable adaptive force/position controller for a C5 parallel robot: a neural network approach. *Robotica*, 30(7):1177-1187, 2012
- [41] O. Korkmaz and S. K. Ider. Hybrid force and motion control of flexible joint parallel manipulators using inverse dynamics approach. *Advanced Robotics*, 28(18):1221-1230, 2014
- [42] J. Casalilla, M. Vallés, Á. Valera, V. Mata and M. Díaz-Rodríguez. Hybrid force/position control for a 3-DOF 1T2R parallel robot: Implementation, simulations and experiments. *Mechanics Based Design of Structures and Machines*, 44(1-2):16-31, 2016
- [43] C. Gao, D. Cong, X. Liu, Z. Yang and H. Tao. Hybrid position/force control of 6-dof hydraulic parallel manipulator using force and vision. *Industrial Robot*, 43(3):274-283, 2016
- [44] L. Villani and J. De Schutter. Force Control. In: Springer Handbook of Robotics, B. Siciliano and O. Khatib (Eds.), Springer, Berlin, Heidelberg, 2008, pp. 161-185
- [45] M. J. Tobias. Matrices in Engineering Problems. In: Synthesis Lectures on Mathematics and Statistics, Series Editor: S. G. Krantz, Morgan & Claypool Publishers, 2011
- [46] G. Alici. Determination of singularity contours for five-bar planar parallel manipulators. *Robotica*, 18(5):569-575, 2000
- [47] J. J. Cervantes-Sánchez, J. C. Hernández-Rodríguez and J. G. Rendón-Sánchez. On the workspace, assembly configurations and singularity curves of the RRRRR-type planar manipulator. *Mechanism and Machine Theory*, 35(8):1117-1139, 2000
- [48] E. Macho, O. Altuzarra, C. Pinto and A. Hernandez. Workspaces associated to assembly modes of the 5R planar parallel manipulator. *Robotica*, 26(3):395-403, 2008
- [49] A. D. Polyanin and V. F. Zaitsev. Handbook of Ordinary Differential Equations: Exact Solutions, Methods, and Problems. Chapman & Hall/CRC Press, Taylor & Francis Group, Boca Raton, Florida, USA, 2018

Overall Equipment Effectiveness (OEE) Complexity for Semi-Automatic Automotive Assembly Lines

Péter Dobra¹, János Jósmai²

¹Doctoral School of Multidisciplinary Engineering Sciences, Széchenyi István University, Egyetem tér 1, 9026 Győr, Hungary, e-mail: dobra.peter@sze.hu

²Department of Vehicle Manufacturing, Széchenyi István University, Egyetem tér 1, 9026 Győr, Hungary, e-mail: josvai@ga.sze.hu

Abstract: In industrial practice, measuring and monitoring production performance is an essential task. The production plan performance is monitored by middle and top management of companies daily, weekly and monthly and make short and long-term operational and strategic decisions when necessary. One of the most common ways of measuring the performance of production and, within this, of assembly lines, is to use the Overall Equipment Effectiveness (OEE) indicator. Although companies sometimes interpret and use this Key Performance Indicator (KPI) in their own way, it is the indicator that best reflects the development of the production efficiency for a given company. A high OEE percentage means high performance, which directly increases the company's profitability. This article explores the complexity of the OEE indicator, supported by the use of a cause and effect diagram. Firstly, a literature review demonstrates scientific relevance. Secondly, the factors affecting OEE are grouped and analyzed according to the following six aspects: man, environment, method, material, machine, and measurement. Each factor is further subdivided into five groups, and then these subgroups also cover five key factors of importance for the approachability of 100% OEE. The 150 aspects listed herein, provide a complete guideline for a semi-automatic assembly line, to consistently increase efficiency in industrial practice.

Keywords: KPI; OEE; assembly line; cause and effect diagram

1 Introduction

Today's automotive manufacturing environment is becoming increasingly complex thanks to Industry 4.0, Smart manufacturing, Big Data, Artificial Intelligence, Lean, IoT, among others. Production logistics systems are becoming increasingly complex in a turbulent industrial environment [1]. Efficiency and flexibility on the part of manufacturing companies are particularly important

especially due to periodic shortages of raw materials (e.g. semi-conductor, chip, metal, plastic) and other constraints (e.g. COVID situation).

The complex environment also adds complexity to performance indicators. The efficiency of production systems, including assembly lines, is increasingly effected by a number of components, both positively and negatively. Modularity, flexibility, digitalization, automation, autonomous processes, autonomous systems, autonomy of an equipment [2], cloud computing help to achieve higher efficiency and productivity, while higher product variety, growing product complexity, shortening product life cycle [3] and complex material flow hinder [4]. Increasingly, the question of efficiency arises: which scopes should be assembled in the final assembly and which ones in the pre-assembly line [5]?

In industrial practice, measuring and monitoring production performance is an essential task. The production plan performance is monitored by middle and top management of companies daily, weekly and monthly and make short and long-term operational and strategic decisions when necessary. One of the most common ways of measuring the performance of production and, within this, of assembly lines is to use the Overall Equipment Effectiveness (OEE) indicator. Although companies sometimes interpret and use this KPI in their own way, it is the indicator that best reflects the development of the production efficiency of a given company. Key Performance Indicators (KPIs) or also known as Key Success Indicators (KSIs) are quantitative measurement tools for the improvement of the machine or line performance [6]. A high OEE percentage means high performance, which directly increases the profitability of the company.

The aim of this paper is to reveal the complexity of OEE using cause and effect diagram. The paper is organized as follows. Section 2 focuses on the relevant scientific work regarding to OEE and cause and effect diagram. Following, Section 3 introduces and details the OEE complexity at a semi-automatic assembly line in automotive industry by fishbone diagram. Last section, Section 4, concludes the paper.

2 Literature Review

Higher expectations of the customers' and new industrial and IT developments have resulted is an increased complexity of Production System (PS) especially assembly systems [7] [8]. This also implies the complexity of the performance evaluation system. Okwir *et al.* define the following six forms of Performance Measurement Complexity (PMC): role, task, procedural, methodological, analytical and technical complexity [9].

Nowadays, the traditional Key Performance Indicator (KPI) system is still well managed due to the Manufacturing Execution System (MES) [10], but further

increasing the efficiency indicators such as Overall Equipment Effectiveness (OEE) is not a simple task in practice. It is becoming increasingly difficult to take real measures that will lead to significant improvements in the short term. Problems almost always have multiple root causes and this complexity is also increased especially at the hybrid assembly lines where automatic devices are combined with manual work in one system [11].

2.1 OEE at the Semi-Automatic Line

A plethora of publications shows the applicability of OEE in the domain of manufacturing [12], Corrales et al. collected almost 900 articles between 1996 and 2020 [13]. This standard indicator is widely used for internal efficiency at the semi-automatic assembly lines [14]. Within the concept of Total Productive Maintenance (TPM), OEE metric was introduced in 1988 by Nakajima [15]. The original formula for calculating OEE is written as:

$$OEE = A P Q \quad [\%] \quad (1)$$

Where:

A = Availability

P = Performance

Q = Quality

100% OEE means, that we exclusively produce high-quality products without stop at maximal capacity, although there are no machines with 100% reliability [16]. During the last few decades several performance indicators and techniques are developed from the basic OEE structure [17] among others: Overall Equipment Effectiveness of a Manufacturing Line (OEEML) [18] and Global Production Effectiveness (GPE) [19].

OEE can be characterized by the following items:

- Metric which shows the reliability of the production network [20]
- OEE is a mechanism to continuously monitor and improve the efficiency of a production processes, focus on zero loss, zero break downs, zero defects and zero accidents [21-23]
- Clearly shows current status of production [24] [25]
- Standard and best practice, can be used to compare with the other assembly line performance during the operation [26]
- Reduce or eliminate six major losses (equipment breakdown losses, setup and adjustment losses, minor stoppage losses, speed reduction losses, defective losses and startup losses) [15, 27] and increase efficiency in the production processes [28].

From other perspective, availability is influenced by the technical failures of workstations and changeover, performance is influenced by small stops and reduced speed, quality is influenced by scrap and rework [29]. Real example of OEE analysis using the waterfall chart at an assembly line shows Fig. 1. (Source: data collected by the authors on the semi-automatic assembly line of a Hungarian automotive supplier.)

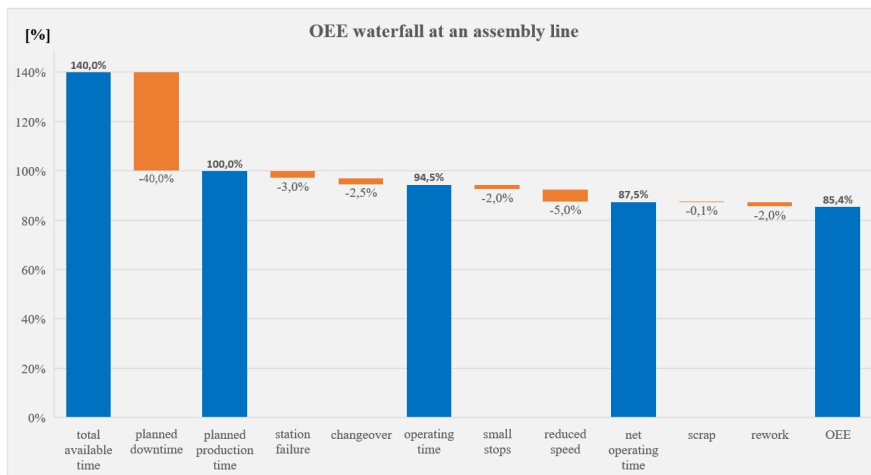


Figure 1
Real example of OEE waterfall chart at an assembly line

The main benefits of implementing and applying OEE are the reduced manufacturing cost, increased uptime, higher speed, minimized material waste, better asset utilization, lower overhead cost, additional sales capacity, reduced inventory and reliable assembly processes [30]. At an assembly line at least one of the workstations is the bottleneck. This article focus on this bottleneck station regarding OEE.

2.2 Cause and Effect Diagram

In manufacturing industry huge losses and/or waste occur in the production shop floor. These losses due to operators, maintenance personnel, process, tooling problems and lack of components in time, etc. [31]. In case of capacity problems, increasing overtime and shift numbers, purchasing new machines, equipment and tools can be a solution to fully meet customer demands, but a much better alternative is to make better use of existing resources, increase machine efficiency, keep bottlenecks under control, and reduce downtime and set-up times.

To decrease losses, several quality management concepts and tools such as Lean Manufacturing, Toyota Production System (TPS), Total Productive Maintenance

(TPM), and Failure Mode and Effect Analysis (FMEA) had been developed in order to achieve higher operational level. There are numerous quality improvement techniques available for improving equipment OEE among others as PDCA cycle, Failure Tree Analysis (FTA), why-why analysis, Value Stream Mapping (VSM), RADAR, DMAIC, EFQM, DFSS, Pareto chart and cause and effect diagram [32] [33].

Cause and effect diagram, Ishikawa or fishbone diagram is one of the seven tools in the quality control system. Firstly, it was presented as a casual diagram by Ishikawa in 1968 [34]. Fishbone diagrams have been constructed mostly based on the categories of man, machine, method, material measurement and environment. Ishikawa diagram is a useful tool to determine the possible causes for a problem, represents the relationship, but it directly does not identify the root causes of the problems [35] [36]. According to Czifra et al. Ishikawa diagram is the most used method on a regular basis in automotive industry in addition to FMEA, 8D, and 5 Why analysis [37].

In the manufacturing industry, several cause and effect research works were published related to OEE. Table 1 shows the articles over the last six years.

Table 1
List of used cause and effect diagrams for improving OEE

Author	Year	Ishikawa elements	Effect on OEE
[38]	2015	manpower, material, methods, milieu, machine	process deviation
[39]	2016	man, machine, material, method, environment	technical failure (part clamping)
[39]	2016	man, machine, material, method, environment	technical failure (hydraulic oil is mixed up with cutting oil)
[40]	2016	environment and social, lead time, machine, management, quality issues, man	poor OEE
[41]	2017	waiting, extra-processing, defects, workforce, environment	low performance
[35]	2017	man, machine, material, measure, management, environment	idling and minor stoppage losses
[35]	2017	man, machine, material, measure	breakdown losses
[36]	2017	man, machine, material, measure, management, environment	idling and minor stoppage losses
[36]	2017	man, machine, material, measure	breakdown losses
[42]	2017	people, work method, environment	technical failure (limit switches failure)

[43]	2018	man, environment, machine	technical failure (overheating of electric motors)
[44]	2018	equipment failure, reduced speed, defect and rework, setup and adjustment, idling and minor stoppage, startup issue	reduced OEE
[45]	2018	method, human, material, machine	low OEE value
[46]	2018	atmosphere, method, man, material, machine	reduce OEE
[47]	2019	machine, man, method, material, measurement	six big losses
[48]	2019	machine, man, method, material, environment	reduced speed losses
[48]	2019	machine, man, method, material, environment	rework losses
[48]	2019	machine, man, method, material, environment	breakdown losses
[49]	2019	method, material	process cycle efficiency
[50]	2020	machine, man, environment, material, method	idling and minor stoppage losses

3 Complexity of Overall Equipment Effectiveness

The complexity of the OEE indicator on assembly lines is best represented by a cause and effect diagram. The areas of man, environment, method, material, machine and measurement, fully cover the conditions that have to be fulfilled for the OEE indicator to be 100% (Fig. 2).

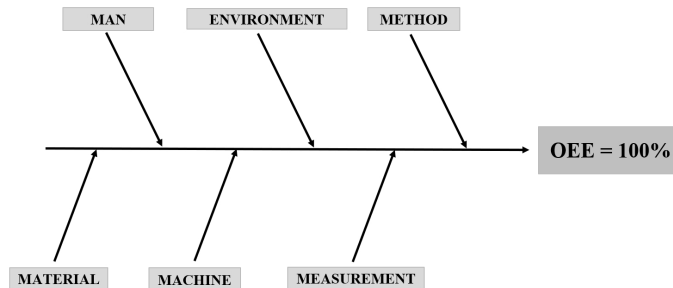


Figure 2
Elements of cause and effect diagram

Each of the six elements is described in detail below in case of the semi-automatic assembly lines.

3.1 Man as a Key Element

In case of hybrid lines, the human is a key factor, as a certain percentage of assembly operations are physically carried out by humans. In addition to work operations, machine set-up, quality control operations, some material handling and operational management are also performed by human beings. The human factor manifests itself in five major areas:

- **Qualification:** Typically determined by the operator's, setter's education, special knowledge for the assembly task, practical experience, internal and external training
- **Skills and abilities:** Workers and machine adjusters must have proper perceptions (eyesight, hearing) to fulfill the assembly and machine setting processes, another important factor is fine motor skills, stamina and communication skills (e. g. be able to indicate the problems properly)
- **Personality and character:** For right assembly operations punctuality, adequate speed, compliance, monotony tolerance and systematic, conscientious work is needed
- **Motivation:** Maximum efficiency can be achieved based on pre-defined goals, need the expectations of employee, crucial factor the rewards and condemnations, management must ensure the team spirit, company welfare and excellent work conditions
- **Organization:** The most critical factor is the available staff (right person in the right workplace), within the factory the continuous improvement activities are indispensable, manufacturing and assembly processes should be supported by the leaders, engineers and managers, scheduling and production planning are also significant elements.

Fig. 3 depicts the role of the Man factor in the cause and effect diagram.

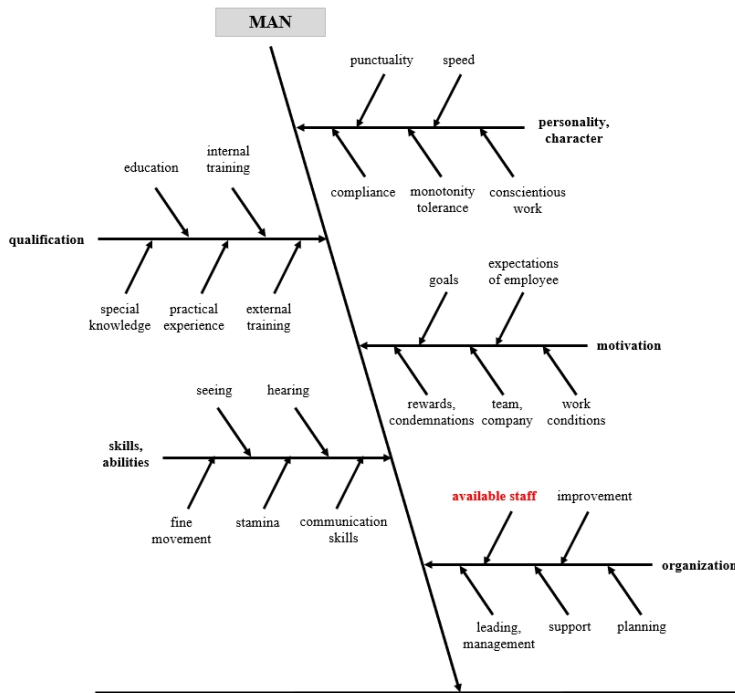


Figure 3
Role of the Man factor in cause and effect diagram

3.2 Environment of Assembly Operations

The manufacturing environment for semi-automatic assembly lines or hybrid lines is extremely complex. Several assembly operations take place simultaneously, the steps of the process are built on each other, and in the case of a one-piece material flow, it is essential to serve the production with raw materials and semi-finished products in time. Companies have to adapt to changing market needs (batch size, product variety) in a number of ways. This requires a thorough understanding of the following 5 key environmental factors:

- **Work environment:** The direct working environment of the assembler, which includes safety and health protection, ergonomic design of workstations, correct perception of the environment and automation of machinery
- **Production environment:** The correct execution of assembly workflows is ensured by technological complexity and concerns, the 5S design of the manufacturing environment, lossless assembly processes and visual support

- **Market environment:** The turbulent market environment includes, on the customer side, the intensity of orders, the state of competition in the market, the pull system, and on the supplier side, the production plan feasibility and, as main factors, the cycle time and cycle time feasibility of assembly operations
- **Company environment:** Within manufacturing companies the production team organization is important, as well as to define the appropriate shift schedule with necessary overtime, employees need to be motivated, committed and engaged
- **Worker environment:** Operator and setter social situation and social acceptance (be able and want to work in that position), easy plant and workplace availability, preferred benefit package.

Fig. 4 shows the role of the Environment factor in the fishbone diagram.

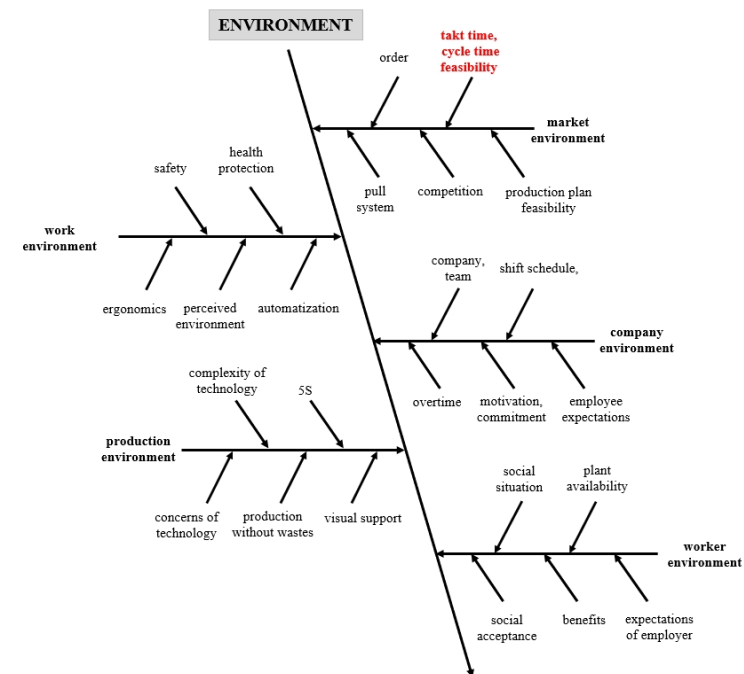


Figure 4
Environment factor in fishbone diagram

3.3 Methods to Achieve the Set Goals

The methods, especially the practical methods, show the way to achieve a high OEE percentage on a semi-automatic assembly line. There is no single method to achieve 100% efficiency, either in the short or long term. On the contrary, a combination of well-chosen procedures and processes can bring you closer to the desired result. In the case of OEE, the following five main groups of methods need to be examined:

- **Production technology:** The most important category is the properly designed assembly technology and processes, it pays attention to repair, rework checking, packaging processes with necessary automation
- **Measurement and control:** During the assembly operations, the quality of the product and the correctness of manufacturing processes must be constantly monitored, aided by the 100% inspection, SPC control, six sigma method, failure analysis, PDCA cycle, Pareto analysis, Poka-yoke and the check of prescriptive maintenance activities
- **Work process:** Relevant factor the predefined Standard Operational Procedures (SOP), assembly processes, material flow, applied best practices and the planned and realized cycle time
- **Lean methods:** Numerous Lean tools exist, but before using them we need to determine the goals of assembly process, the expectations by taking into account company characteristics, working conditions, team structure and reward- and motivation factors
- **Material and information flow:** Besides the workforce it is important to take into consideration the components and materials flow, besides planning, continuous development and support, the organization must also adapt to achieve loss-free assembly.

Fig. 5 shows the relationship of Methods in the Ishikawa diagram

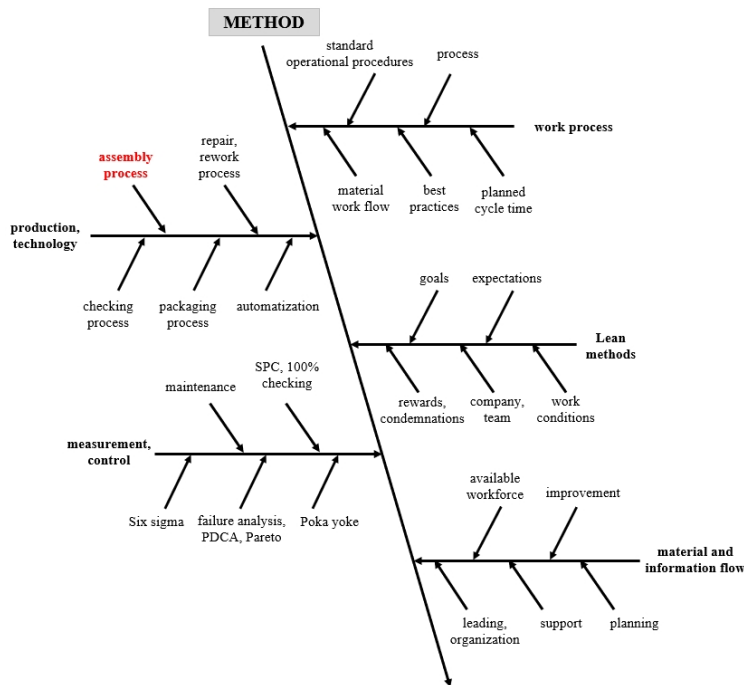


Figure 5
Relationship of Methods in the Ishikawa diagram

3.4 Material, Component, Part and Subassembly

Raw materials, auxiliary materials, semi-finished products, assemblies, sub-assemblies are essential for the operation of assembly processes. They must be available at the right time, in the right quantity, in the right order, in the right place and of the right quality. Any one of these missing will result in a significant OEE loss. A particular aspect is that the availability of components to be assembled can be taken into account in production planning and, if necessary, the production sequence can be modified to ensure continuous assembly. The following five main factors influence material complexity:

- **Material failure:** It is of paramount importance that the quality, surface and color of the materials to be incorporated, as well as the required quantity of materials, are available (problems can arise from incomplete or surplus materials during assembly)
- **Size error:** The materials used in the assembly must have the dimensions prescribed on the drawing, such as width, length, height, tolerances, defined shape and position

- **Quantitative error:** On the production lines, the right quantity of building materials must be available for assembly (not more, not less, not mixed, not interlocked)
- **Material handling:** During material handling processes, materials awaiting assembly must be protected from contamination and damage, stored at appropriate temperatures and they must be identifiable
- **Design failure:** During the design process focus should be placed on the possible function and comfort problems as well as, the ease of assembly, repair and general checking of the product.

Fig. 6 depicts the Material factor in the cause and effect diagram.

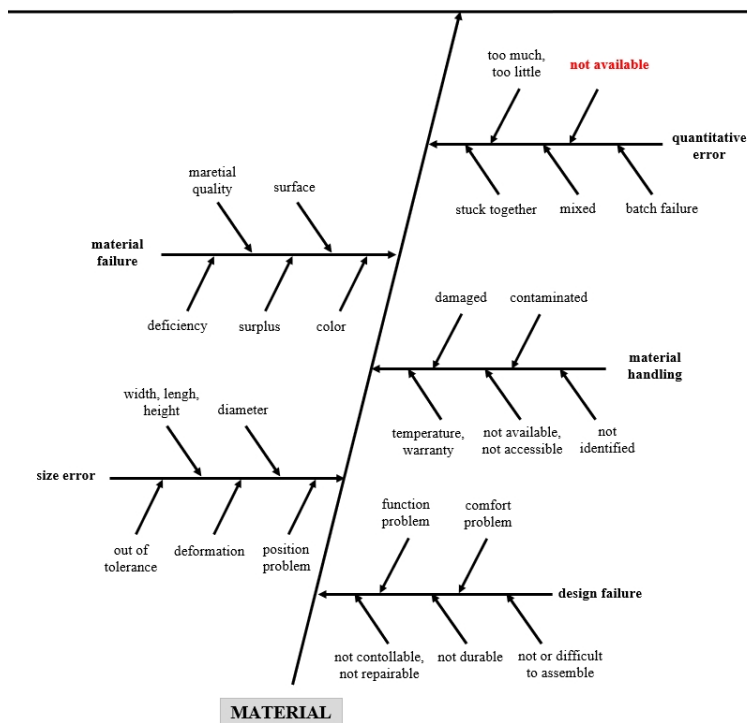


Figure 6
Role of the Material factor in cause and effect diagram

3.5 Machine, Tool and Workstation

Semi-automatic assembly lines consist of different workstations connected in series or in parallel, where mechanical and manual assembly operations are carried out. The continuous availability of modular assembly lines, machines,

equipment and tools used today is complex in several respects. The five main aspects are the following:

- **Maintenance:** A maintenance plan must be drawn up and its content must be carried out in a timely and appropriate manner, the necessary documentation (drawings, manuals) must be available, machinery and tools must be easily repairable and replaceable
- **Machine and tool adjustment:** Workstations and tools must be easy to set up based on the setup instructions provided, a fault log is an essential requirement, and quick changeover during product changeovers must be ensured (using SMED and OTED)
- **Stability:** The assembly line must be stable and continuously operational with low energy consumption, supported by a reliable PC and PLC network, the degree of machine capability and process capability should be high
- **Standardization:** It is advisable to build the assembly line from standard parts for which the spare part must be continuously provided, the complete assembly line must be connected to the Manufacturing Execution System (MES) so that the installed parts and key process parameters and values are digitally recorded and stored
- **Safety:** Machinery and equipment must be safe, safe and easy to use from a safety point of view, and ergonomically designed.

Fig. 7 shows the role of the Machine factor in the fishbone diagram.

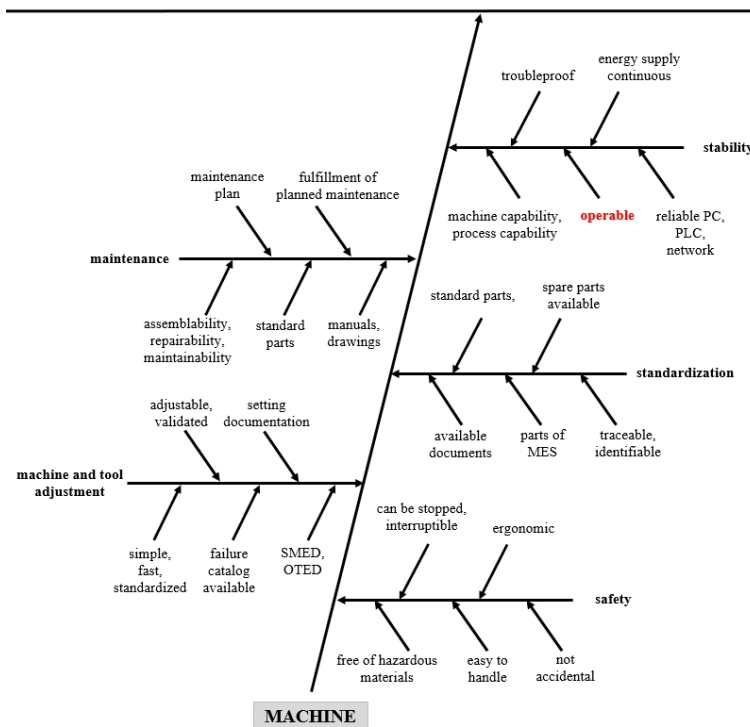


Figure 7
Machine factor in fishbone diagram

3.6 Measurement for Right Quality

The products ordered by the customer must be of the quality expected. Both the quality of the product and the quality of the processes must be measured and checked before and during production and assembly. Based on the results obtained, further interventions and corrections are possible. During measurement, the following 5 factors influence the OEE:

- **Material checking:** It is necessary to check the quantity, quality and function of the components and materials to be incorporated prior to assembly operations, preferably at the time of receipt of the goods, the traceability of materials (e.g. FIFO, batch) is also essential
- **Product control:** During assembly, the conformity of the product shall be checked and documented at the required frequency and in the required number of pieces in the defined condition and location with regard to its functional operation

- **Machine and tool checking:** Testing, checking, calibration and safety control by appropriate frequency essential at the machines and tools, in addition, the performance of maintenance should also be checked
- **Checking assembly process:** During assembly operations and type change the first and last assembled unit must be checked, in addition to these, simulation and poke yoke checks are also essential
- **Measuring instrument checking:** The measuring instruments and gauges used in production must be checked and documented at appropriate intervals for functionality, reliability and accuracy.

Fig. 8 shows the relationships of Measurement in the Ishikawa diagram

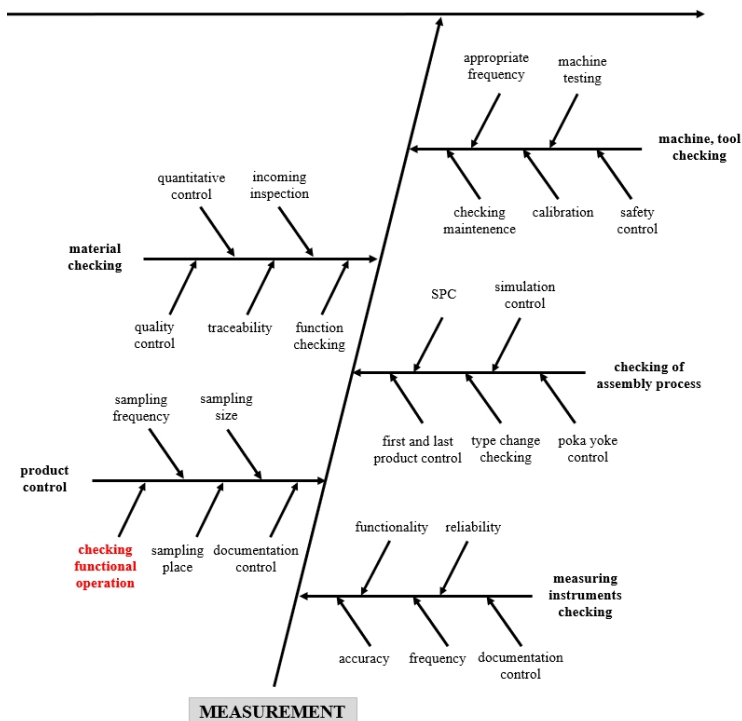


Figure 8

Relationships of Measurement in the Ishikawa diagram

In the cause and effect diagram, the most important factors for each branch have been highlighted in red, as follows:

- **Man:** Within organization, the available staff
- **Environment:** Within market environment, the Takt-time and Cycle-time feasibility

- **method:** Within production technology, the assembly process
- **material:** Within quantitative error, the not available material
- **machine:** Within stability, the operable machine
- **Measurement:** Within production control, the checking functional operation.

The authors are aware that the factors listed could be presented in much more detail, but for reasons of content, the article presents a kind of overview of how the OEE indicator can be influenced by a number of factors and how the interrelationships between factors lead to complexity on semi-automatic assembly lines.

Conclusions

In this work, the complexity of the Key Performance Indicator (KPI), used to measure the performance of a semi-automatic assembly line, has been presented. Based on a cause and effect diagram, the six main groups (man, environment, method, material, machine and measurement) were further broken down into five factors, within which, five factors were also identified. All the factors are necessary to a varying degrees, to achieve 100% OEE, but the indispensable factors are, available manpower, cycle time, cycle time feasibility, right assembly process, available material, operable machine and the checking functional operation. In the future, a further expansion of this article may apply weighting and ranking factors, presented in terms of their impact on the value of OEE.

References

- [1] H. P. Wiendahl, J. Reichardt, P. Nyhuis, Handbook Factory Planning and Design, Springer-Verlag, Berlin, 2015
- [2] H. ElMaraghy, W. ElMaraghy, Smart Adaptable Assembly Systems, Procedia CIRP 44 (2016) pp. 4-13
- [3] E. Permin, F. Bertelsmeier, M. Blum, J. Bützler, S. Haag, S. Kuz, D. Özdemir, Self-Optimizing Production Systems, Procedia CIRP 41 (2016) pp. 417-422
- [4] M. Glatt, J. C. Aurich, Physical Modeling of Material Flows in Cyber-Physical Production Systems, Procedia Manufacturing 28 (2019) pp. 10-17
- [5] C. Küber, E. Westkämper, B. Keller, H. F. Jacobi, Method for a Cross-Architecture Assembly Line Planning in the Automotive Industry with Focus on Modularized, Order Flexible, Economical and Adaptable Assembly Processes, Procedia CIRP 57 (2016) pp. 339-44
- [6] L. M. Dawood, Z. H. Abdullah, Szudy impact of Overall Equipment and Resource Effectiveness onto Cement Industry, Journal of University of Babylon, Engineering Sciences 26 (2018) pp. 187-198

-
- [7] B. Denkena, M. A. Dittrich, S. Wilmsmeier, Automated production data feedback for adaptive work planning and production control, *Procedia Manufacturing* 28 (2019) pp. 18-23
- [8] A. Fast-Berglund, U. Harlin, M. Akerman, Digitalisation of Meetings – From White-Boards to Smart-Boards, *Procedia CIRP* 41 (2016) pp. 1125-1130
- [9] S. Okwir, S. S. Nudurupati, M. Ginieis, J. Angelis, Performance Measurement and Management Systems: A Perspective from Complexity Theory, *International Journal of Management Reviews* 20, No. 3 (2018) pp. 731-754
- [10] S. Mantravadi, C. Moller, An Overview of Next-generation Manufacturing Execution Systems: How important is MES for Industry 4.0?, *Procedia Manufacturing* 30 (2019) pp. 588-595
- [11] H. P. Wiendahl, H. A. ElMaraghy, P. Nyhuis, M. F. Zäh, H. H. Wiendahl, N. Duffie, M. Brieke, Changeable manufacturing - Classification, design and operation, *CIRP Annals* 56/2 (2007) pp. 783-809
- [12] G. Agyei, I. Asamoah, A Selection of Drill Rigs using Overall Equipment Efficiency Approach, *Journal of Science and Technology Research* 1, (2019) pp. 41-52
- [13] L. C. Corrales, M. P. Lambán, M. E. H. Korner, J. Royo, Overall Equipment Effectiveness: Systematic Literature Review and Overview of Different Approaches, *Applied Sciences* 10 (2020) pp. 6469
- [14] M. Kurdve, U. Harlin, M. Hallin, C. Söderlund, M. Berglund, U. Florin, A. Landström, Designing Visual Management in Manufacturing from a User Perspective, *Procedia CIRP* 84 (2019) pp. 886-891
- [15] S. Nakajima, Introduction to TPM: Total Productive Maintenance, Productivity Press Cambridge, 1988
- [16] J. Dias, E. Nunes, S. Sousa, Productivity Improvement of Transmission Electron Microscopes - A Case Study, *Procedia Manufacturing* 51 (2020) pp. 1559-1566
- [17] M. S. J. Hossain, B. R. Sarker, Overall Equipment Effectiveness measures of engineering production system, Annual Meeting of the Decision Sciences Institute, Conference Paper, 2016
- [18] M. Braglia, M. Frosolini, F. Zammori, Overall equipment effectiveness of a manufacturing line (OEEML), *Journal of Manufacturing Technology Management*, 20 (2008) pp. 8-29
- [19] R. Oliveira, S. A. Taki, S. Sousa, M. A. Salimi, Global Process Effectiveness: When Overall Equipment Effectiveness Meets Adherence to Schedule, *Procedia Manufacturing* 38 (2019) pp. 1615-1622

- [20] J. Oliveira, J. C. Sa, A. Fernandes, Continuous Improvement through ‘Lean Tools’: An Application in a Mechanical Company, *Procedia Manufacturing* 13 (2017) pp. 1082-1089
- [21] G. R. Naik, V. A. Raikar, P. G. Naik, A Simulation Model for Overall Equipment Effectiveness of a Generic Production Line, *Journal of Mechanical and Civil Engineering* 12 (2015) pp. 52-63
- [22] P. S. Sisodiya, M. Patel, V. Bansod, A literature review on Overall Equipment Effectiveness, *International Journal of Research in Aeronautical and Mechanical Engineering* 2 (2014) pp. 35-42
- [23] K. Sowmya, N. Chetan, A review on Effective Utilization of Resources Using Overall Equipment Effectiveness by Reducing Six Big Losses, *International Journal of Scientific Research in Science, Engineering and Technology* 2 (2016) pp. 556-562
- [24] A. J. Gujar, N. M. Kambale, S. D. Maner, S. S. Joshi, S. G. Chandne, A. A. Chavare, A case study for Overall Equipment Effectiveness improved in manufacturing industry 6 (2019) pp. 4841-4844
- [25] J. Lee, E. Lapira, B. Bagheri, H. Kao, Recent advances and trends in predictive manufacturing system in big data environment, *Manufacturing Letters* 1 (2013) pp. 38-41
- [26] S. F. Fam, S. L. Loh, M. Haslinda, H. Yanto, L. M. S. Khoo, D. H. Y. Yong, Overall Equipment Effectiveness (OEE) Enhancement in Manufacture of Electronic Components and Boards, *Industrx through Total Productive Maintenance Practices, MATEC Web of Conferences* 150 (2018)
- [27] M. Subramaniyan, Production Data Analytics – To identify productivity potentials, Chalmers University of Technology, Gothenburg, Sweden, 2015
- [28] L. Hassani, G. Hashemzadeh, The impact of Overall Equipment Effectiveness on production losses in Moghan Cable and Wire manufacturing, *International Journal for Quality Research* 9 (2015) pp. 565-576
- [29] M. P. Rössler, E. Abele, Uncertainty in the analysis of the Overall Equipment Effectiveness on the shop floor, *IOP Conference Series: Materials Science and Engineering* 46 (2013)
- [30] A. S. Vairagkar, S. Sonawane, Improving Production Performance with Overall Equipment Effectiveness (OEE), *International Journal of Engineering Research and Technology* 4 (2015) pp. 700-704
- [31] K. Pradeep, S. Raviraj, L. R. R. Lewlyn, Overall Equipment Efficiency and Productivity of a News Paper Printing Machine of a Daily News Paper Company – A Case Study, *International Journal of Engineering Practical Research* 3 (2014) pp. 20-27

- [32] M. Sokovic, D. Pavletic, K. K. Pipan, Quality improvement methodologies – PDCA Cycle, RADAR Matrix, DMAIC and DFSS, *Journal of Achievements in Materials and Manufacturing Engineering* 43 (2010) pp. 476-483
- [33] K. E. Chong, K. C. Ng, G. G. G. Goh, Improving Overall Equipment Effectiveness (OEE) through Integration of Maintenance Failure Mode and Effect Analysis (Maintenance-FMEA) in a Semiconductor Manufacturer: A Case Study, In 2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, Singapore: IEEE (2015) pp. 1427-1431
- [34] K. Ishikawa, *Guide to Quality Control*, Tokyo (1968)
- [35] N. Ahmad, J. Hossen, S. M. Ali, Improvement of Overall Equipment Efficiency of ring frame through Total Productive Maintenance: a textil case study, *International Journal of Advanced Manufacturing Technology* 94 (2018) pp. 239-256
- [36] J. Hossen, N. Ahmad, S. M. Ali, An application of Pareto analysis and cause and effect diagram (CED) to examine stoppage losses: a textil case from Bangladesh, *The Journal of the Textile Institute* 108 (2017)
- [37] Gy. Czifra, P. Szabó, M. Mikva, J. Vanová, Lean principles application in the automotive industry, *Acta Polytechnica Hungarica* 16 (2019) pp. 43-62
- [38] N. Galaske, D. Strang, R. Anderl, Process Deviation in Cyber-Physical Production System, *Proceedings of the World Congress on Engineering and Computer Science*, San Francisco, USA (2015) pp. 1035-1040
- [39] A. Greeshma, A. V. Pooja, M. V. Machaiah, T. P. Govekar, A. Balakrishna, Improvement of Overall Equipment Effectiveness of Barrel Pre-Honing Machine Line, *International Journal of Research Engineering and Technology* 5 (2016) pp. 40-50
- [40] K. S. Arun, B. K. Kanchan, G. Prabha, D. Rajenthirajumar, Increasing an Overall Equipment Effectiveness Visibility and Analysing in a Manufacturing Industry, *International Journal of Manufacturing and Material Processing* 1 (2016) pp. 89-103
- [41] S. F. Fam, N. Ismail, H. Yanto, D. D. Prastyo, B. P. Lau, Lean Manufacturing and Overall Equipment Efficiency (OEE) in Paper Manufacturing and Paper Product Industry, *Journal of Advanced Manufacturing Technology*, Special Issue iDECON (2016) pp. 461-474
- [42] S. Raut, N. Raut, Implementation of TPM to enhance OEE in a medium scale industry, *International Research Journal of Engineering and Technology* 4 (2017) pp. 1035-1040
- [43] M. Kapuyanyika, K. Suthar, To Improve the Overall Equipment Effectiveness of Wheel Surface Machining Plant of Railway Using Total

- Productive Maintenance, *International Journal of Scientific Research in Science and Technology* 4 (2018) pp. 1860-1874
- [44] A. Gedefaye, M. Alehegn, H. Bereket, Balasundaram, TPM and RCM Implementation in Textile Company for Improvement of Overall Equipment Effectiveness, *International Journal of Advances in Scientific Research and Engineering* 4 (2018) pp. 129-136
- [45] A. A. U. Nugeroho, G. R. Prabandanu, R. Nuryadin, E. Rimawan, Effectiveness Analysis of Soehnel L1 Machine Using Overall Equipment Effectiveness (OEE) Method in PT PQR, *International Journal of Innovative Science and Research Technology*, 3, No. 9 (2018) pp. 296-300
- [46] S. Nallusamy, V. Kumar, V. Yadav, U. K. Praaad, S. K. Suman, Implementation of Total Productive Maintenance to Enhance the Overall Equipment Effectiveness in Medium Scale Industry, *International Journal of Mechanical and Production Engineering Research and Development* 8 (2018) pp. 1027-1038
- [47] D. Nusraningrum, L. Setyaningrum, Overall Equipment Effectiveness (OEE) Measurement Analysis for Optimizing Smelter Machinery, *International Journal of Business Marketing and Management* 4 (2019) pp. 70-78
- [48] D. Nusraningrum, E. G. Senjaya, Overall Equipment Effectiveness (OEE) Measurement Analysis on Gas Power Plant with Analysis of Six Big Losses, *International Journal of Business Marketing and Management* 4 (2019) pp. 19-27
- [49] A. Dalimunthe, Sukardi, I. Fahmi, Analysis of the Production Loss of the Automotive Company PT DNIA Using Value Stream Mapping and Overall Resource Effectiveness, *International Journal of Research and Review* 6 (2019) pp. 124-132
- [50] E. B. Meike, K. Hayu, Sunardiyanta, Analysis of Effectiveness Measurement of Stretch Blow Machine Using Overall Equipment Effectiveness (OEE) Method, *International Journal of Advances in Scientific Research and Engineering* 6 (2020) pp. 131-137

Industry 4.0 Narratives through the Eyes of SMEs in V4 Countries, Serbia and Bulgaria

Andrea Tick

Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary
tick.andrea@uni-obuda.hu

Abstract: In the third decade of the 21st Century, thanks to the technological developments and digitization, the spread of Industry 4.0 (I4.0) in production and manufacturing as well as in trade and service industry is unquestionable. The spread is inevitable not just among large, capital-strong companies but I4.0. is also penetrating into the life of SMEs. The present research was conducted among SMEs in V4 countries, Serbia and Bulgaria, and while it analyses which I4.0 technologies predict SMEs' familiarity with Industry 4.0, it also finds similarities with the spread of the relevant terms in the narratives in three corpora. The quantitative research uses regression models to analyze the spread of narratives and the behavior of SMEs and finds that four I4.0 technologies significantly contribute to the familiarity with I4.0 among SMEs in the participating countries, implying that raising awareness and training on special I4.0 technologies need to be strengthened among SMEs. Moreover, the research found that the familiarity of I4.0 terms among SMEs and the spread of these terms in the three corpora are partly in alignment, therefore as narratives boost the spread of the term I4.0 so SMEs get more aware and familiar with certain I4.0 technologies.

Keywords: Industry 4.0; Cloud computing; Big Data Analysis; IoT; 3D printing; Robotics; Ngram Viewer; SME; V4 countries

1 Introduction

The 21st Century has been continuously digitalized, which gradually helped companies to introduce new and innovative technologies, change business and production processes as well as exploit the benefits of Industry 4.0 (I4.0). The adoption of digital technologies represents one of the most significant international business developments of the past few years. I4.0 technologies were first introduced in capital-strong large companies but, with time, SMEs continuously get familiar with these technologies, slowly introducing and integrating them in their business operations despite the fact that such investments are capital intensive [1, 2]. Beyond the technological changes I4.0 brings socio-economic changes (impact on labor market, changes in social structure) as well. However, without the awareness of and familiarity with these technologies SMEs

are not in the position to make responsible decisions in case of such introductions and deployments.

Companies at the same time are continuously interested in using new technologies to adapt to the ongoingly changing business conditions, especially in the times of a pandemic when contactless digitalization helps SMEs to maintain their business operation [1, 3] and long-term competitiveness. The fourth industrial revolution poses a huge challenge for manufacturing companies [4], which affects the companies' technological systems, operational processes as well as their management systems [2]. On the other hand, companies including SMEs are also influenced by the spreading of I4.0 technologies and strive to be proactive in technology usage since it ensures their innovative profile, supports cost effectiveness and improves performance. They must use I4.0 technologies to catalyze the adoption of relevant I4.0 innovations to remain competitive in the global value chain [3]. The pandemic reinforced the importance of the deployment of digital technologies, including I4.0 technologies, and therefore the knowledge of I4.0 technologies makes the digital transformation easier for SMEs [1]. I4.0 has contributed to an increase in efficiency in supply chain management [5], but on the other hand, it might lead to a partial replacement of human labor [6] or might increase cybersecurity issues [7]. Cugno et al. [1] investigates what role I4.0 technologies play in the recovery of SME manufacturing activity to pre-COVID-19 levels and point out that such analyses might support managers to identify the optimal and most appropriate I4.0 technology to implement. According to [8, p. 254] digitalization and I4.0 might give "a key stimulus for innovation in various areas of business" and it can become the driving force in industries. SMEs need to be aware of the digital transformation used in I4.0 and should be able to proact and react properly.

The present research investigates how aware and familiar SMEs are with I4.0 and its elements and draws a parallel with the spread of these terms — that is the usage and occurrence of these terms — in the American, British and German corpora. The research reveals that familiarity with certain I4.0 technologies ensures SMEs to be aware of I4.0 and explores whether these are the same technologies that occur in relation with I4.0 in the narratives. The paper focuses on SMEs in the V4 countries, Serbia and Bulgaria, since these countries share similar economic environmental conditions and are clustered in the same group by digitalization maturity [9]. The research is quantitative in its nature and concludes that out of the eight I4.0 technologies surveyed four technologies significantly contribute to the acquaintance of I4.0 among SMEs, the occurrence of three of them strongly correlate and strongly boost the spread of I4.0 in the corpora while Cloud Computing services, Supply Chain Management and Virtual Reality behave differently among SMEs and in the narratives. The research concludes that SMEs' awareness needs to be raised about I4.0 technologies, promotion and training need to focus on the technologies that SMEs do not associate with I4.0.

The paper is organized as follows: after the definition of the terms, it presents the research model, the research questions, the research methodology and data collection methods. Following, it gives the demographic profile and the responses of the SMEs. Then the behavior of SMEs and the spread of the narratives are analyzed, and the two results are compared. Finally, the paper draws conclusions on the hypothesis and research questions, gives recommendations, discusses the limitations and future possibilities of the research.

2 Industry 4.0

The concept of Industry 4.0 (I4.0), defined by the German Industry–Science Research Alliance [10], has exponentially spread in the narratives (Figure 3) since it was defined in 2011. However, definitions vary across industry and academic research [11]. It can be stated that I4.0 is present at all levels in the management hierarchy, from the production of smart products through process management to the strategic decision level at top management.

I4.0 is based on two pillars, one being digitization while the other incorporates the exponential technologies. Digitization (‘binary conversion’) and digitalization or even digital transformation are defined and used differently in recent literature [12]. In Clerk’s definition, digitalisation is centred on digital information [13]. However, the term digitalization can be understood from both a technical and a business perspective [14]. In business terms digitalization defines newly created business models and processes [15], while in technical sense it refers to the digitization of processes, contents and objects that were previously physical or analogue (Csedő et al., 2019). „In corporate terms digitalisation means turning interactions, communications, business functions and business models into (more) digital ones which often boils down to a mix of digital and physical as in omnichannel customer service, integrated marketing or smart manufacturing with a mix of autonomous, semi-autonomous and manual operations” [16]. This paper uses digitization in its technical sense and considers specific exponential technologies in I4.0.

2.1 Industry 4.0 Technologies

In order to raise awareness of and familiarity with I4.0 among SMEs the knowledge about exponential technologies needs to be raised. According to [17] the sudden proliferation of Internet of Things (IoT) and Big Data caused a mass of disorganized knowledge; however, these technologies are key drivers of business re-engineering. I4.0 entails “the increasing digitization of the entire *supply chain*, which makes it possible to connect actors, objects and systems based on real-time data exchange” [2, p. 120]. Communication technologies and digitization during the 4th industrial revolution first triggered machine-to-human (M2H) then machine-to-machine

(M2M) communication, and the exponential development of *artificial intelligence* (AI) with Web 3.0 and 4.0 opened a new avenue to automated production, *robotization* and resulted in the emerge of sensors and *Internet of Things* (IoT). In order to gather, collect and process data *Cloud computing* created the background for *big data analysis* and provides extensive storing and computing capacities and capabilities [18]. M2H and M2M contribute to Cyber-Physical systems that are capable of creating a digital representation of the physical world, and, as such, the interconnection and communication integrating artificial intelligence allows for *Virtual Reality* (VR), *Augmented Reality* (AR) to be implemented for business processes not just in digital devices [19]. According to Rübmann et al. [20] nine pillars can be identified in I4.0, namely Autonomous robots, Simulations, Horizontal and vertical system integration, IoT, Cybersecurity, The Cloud, Additive Manufacturing, Augmented Reality and Big Data Analytics. The present survey was not aiming to deal with cybersecurity, however, included Artificial Intelligence and Virtual Reality to integrate two essential elements for digital twin and business analytics possibilities. The research used the following pillars: Cloud computing, Big Data Analysis, 3D printing and Robotics, IoT, AI, VR, AR and Supply Chain Management (SCM).

These terms, referred to as narratives here, can follow various patterns, from the pattern of a pandemic through a hype curve to a product life cycle, i.e., they go viral or popular, have their own birth, virulent period, decline and death. Immediately as a new technological innovation is introduced and is accepted to a great extent, a new narrative emerges and becomes viral while the older one declines and gets forgotten [21].

The originality of this study lies in enriching the literature on the topic of I4.0 awareness among SMEs and its relation to the spread of I4.0 technologies in the narratives. To this end it provides an understanding how the awareness of I4.0 and its technologies can be raised, and which technologies can be expected as familiar technology among SMEs, and to what extent SME owners and managers are aware and familiar with the terms. Furthermore, the study contributes to the literature in including SMEs from the V4 countries, Serbia and Bulgaria. It uses data from two different sources to test the relationship between the spreading of and familiarity with I4.0 and its technologies.

3 Research Methods and Data

The aim of the research is to explore how familiarity with I4.0 can be increased among SMEs in the V4 countries, Serbia and Bulgaria, and whether it can be increased by raising the familiarity with certain I4.0 technologies. It strives to give a good prediction for familiarity, and it explores to what extent I4.0 technologies influence the spread of I4.0 narratives. Furthermore, it compares whether SMEs'

awareness of I4.0 follows the trends in the narratives. Finally, the research aims to find which Industry 4.0 technologies should be more promoted among SMEs in order to familiarize these small- and medium-sized companies with I4.0 thus helping SME managers both to digitalize more and boost business performance and efficiency and to invest in I4.0 technologies to recover from COVID-19.

The research model is based on I4.0 and its selected pillars, and is presented in Figure 1, the familiarity with Industry 4.0 technologies determines and predicts the familiarity with I4.0 among SMEs in the V4 countries, Serbia and Bulgaria.

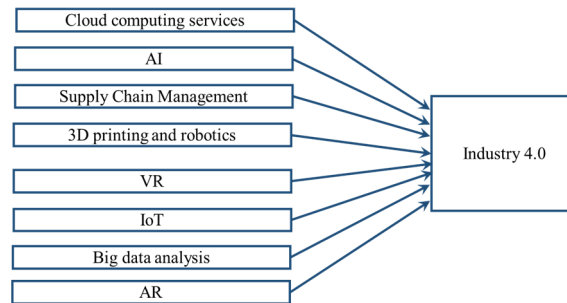


Figure 1

The proposed research model (developed by author)

As no similar analysis could be found in the literature the research contributes to the field of interest by proposing and formulating three research questions, namely:

RQ1: Do SMEs in the V4 countries, Serbia and Bulgaria associate I4.0 technologies, namely Cloud computing, Big Data Analysis, 3D printing and Robotics, IoT, AI, VR, AR and SCM with Industry 4.0 and can the familiarity with these technologies predict the familiarity with I4.0 at SMEs?

RQ2: Do the terms I4.0 and I4.0 technologies spread similarly and are the occurrences highly correlated in the selected corpora? Does the spread of I4.0 technologies in the narratives significantly influence the spread of the term Industry 4.0?

RQ3: Do the SMEs in the V4 countries, Serbia and Bulgaria follow the trend of the spreading of the terms in the selected corpora regarding the familiarity with I4.0 and its technologies? Do they identify the same terms as I4.0 technologies as the narratives suggest?

3.1 Research Method and Data Collection among SMEs

The research used self-administered questionnaires, anonymity was ensured, and responding SMEs gave their consent to the survey. Responses were collected during the pandemic between September and November 2021. The quantitative analysis was conducted by the statistical program SPSS V25, Rapid Miner and MS Excel.

Descriptive analysis, Chi2 tests, correlations, regression procedures, and data mining modeling were used to explore which I4.0 technologies are determining the model. Binary logistics regression was applied to predict the probability of the familiarity with I4.0 given certain predictor variables (I4.0 technologies) [22].

3.2 Research Method and Data Collection for Spreading of the Terms Related to I4.0

The Google Ngram Viewer was used to collect data on the spread of narratives from 1950 to 2019 (the latest date available). Ngram Viewer provides a good visual representation of the frequency of terms in various corpora ranging from English (British and American separately) through French, German, Italian, Spanish, etc. to even providing Chinese corpus. It is an online search engine that charts the relative frequencies of any set of search words and phrases, using a yearly count of n-grams found in millions of books, printed between 1950 and 2019, in the Google Books corpus. The corpus enables the quantitative analysis of cultural, linguistic as well as economic or business trends [23, 24]. The source is limited to the pool of Google Books but depicts well the popularity, development and proliferation of these phrases.

Quantitative analyses were conducted in MS Excel, correlation and regression analyses were carried out to see the relationships and influences of the terms on I4.0 and its spreading in the corpora. Finally, the results of the two analyses are compared and conclusions are drawn about the behavior of SMEs compared to the spread of narratives in the corpora.

4 Familiarity with I4.0 and its Elements in the Narratives

4.1 Spread of Industry 4.0 Terms

As soon as the technological innovations are announced they appear in the narratives and, as Schiller [21] states, they go viral and follow the spread of viruses. The life of technological innovations at the same time follows the shape of the Gartner's Hype cycle [25] as well — Technology Trigger, Peak of Inflated Expectations, Trough of Disillusionment, Slope of Enlightenment and Plateau of Productivity — and as the technologies get older and more mature they are approaching the plateau of production, they are more widely used in industry and are more widespread in publications. According to Gartner Research [25] and Kenn, et al. [26] Engineering and Business Maturity and the Hype Cycle of technologies converge and run align in the phase of the plateau of production.

The Google Ngram Viewer is one of databases that allow researchers to see the spread of narratives. The usage of the terms — I4.0, Cloud computing, Big Data Analysis, 3D printing and robotics, IoT, AI, VR, AR and SCM — was analyzed between 1950 and 2019 (the latest date available). As Figure 2 displays, the usage and spread of the listed terms show similar trends in the American, English and German Corpora.

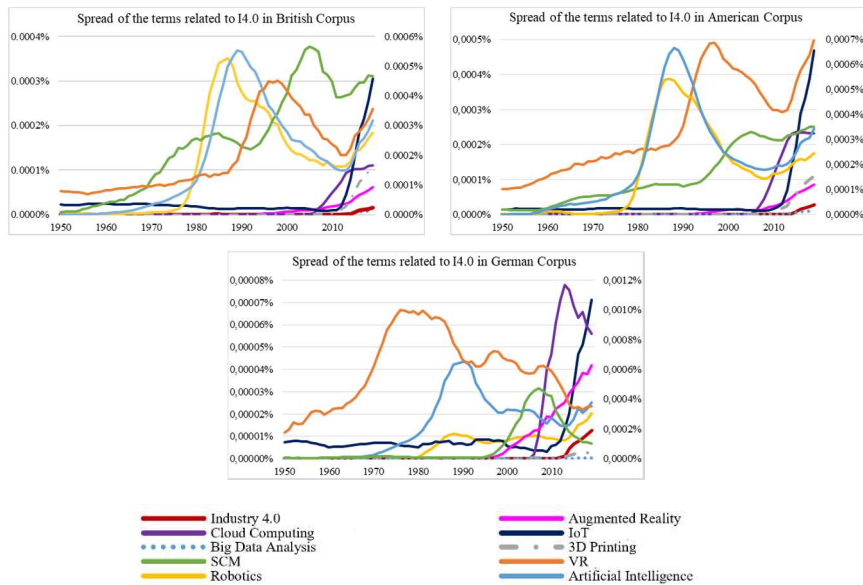


Figure 2

Spread of the terms related to Industry 4.0 in the British, American and German corpora (developed by author)

Applying the terminology of the Hype cycle, I4.0 is in the phase of the exponential and rapid growth together with IoT, 3D printing, Big Data Analysis, implying them being in the phase of Technology Trigger. Since no slowdown can be noticed in the spreading, these terms probably have not reached their peak of Inflated Expectation phase yet. Robotics has already reached the plateau of production in the British and American corpora while it is in its second Technology Trigger phase in the German Corpus. After the peak era in the 90s' AI is also in the plateau of production with a second awakening in the British corpus. The spread of VR shows similar trends in the American and British corpora with a 5-year time shift, while it is in the phase of Trough of Disillusionment in the German Corpus.

Slowdown in the American and British corpora and downturn in the German corpus imply that Cloud Computing has already reached its Peak of Inflated Expectation. AR has been also exponentially growing in the corpora, with a short phase of Disillusionment in the American corpus in the first decade in the century, but since

then the term has been virulent and infectious in all three corpora not having reached its peak yet. The term SCM behave differently in all three corpora. While it is still virulent in the American corpus (phase of Technology Trigger), it has its second epidemic wave (Scope of Enlightenment) in the British corpus, and it is not virulent at all in the German corpus (Trough of Disillusionment).

The correlation of the occurrences of the terms was also analyzed to support the research questions. Based on the Pearson's r correlation coefficients (Table 1), the spread of the term I4.0 is in strong correlation with the spread of the I4.0 technologies, such as AR, Cloud Computing, IoT, Big Data Analysis, 3D printing in the narratives, however, the spread of the terms AI, Robotics, SCM and VR moderately or weakly contribute to the spread — and familiarity — of I4.0 in all three corpora.

Table 1
Correlation of I4.0 elements with I4.0 in the American, British and German corpora

Industry 4.0 / Corpus	American	British	German
<i>Augmented Reality (AR)</i>	0.850	0.874	0.831
<i>Cloud Computing</i>	0.727	0.766	0.693
<i>IoT</i>	0.986	0.991	0.989
<i>Big Data Analysis</i>	0.937	0.962	0.994
<i>3D Printing</i>	0.960	0.983	0.974
<i>Robotics</i>	0.061	0.135	0.602
<i>Artificial Intelligence (AI)</i>	0.118	0.147	0.153
<i>SCM</i>	0.430	0.330	0.088
<i>Virtual Reality (VR)</i>	0.377	0.245	-0.301
<i>Multiple R²</i>	<i>99.43%</i>	<i>99.69%</i>	<i>99.73%</i>

Moreover, VR is in negative correlation with I4.0 in the German corpora ($r_D=-0.301$) implying that VR is less associated with I4.0. The term AI is progressing toward the plateau of production; however, having integrated semantic analytics and machine learning it started its second wave around the 2010s', but is presumably not directly linked to I4.0 [27]. Robotics is also in weak correlation with I4.0 in the American and British corpora ($r_{USA}=0.061$, $r_{Br}=0.135$) while it contributes strongly to the familiarity of I4.0 in the German corpus ($r_D=0.602$). The spread of the term is different in the three corpora, while it behaves similarly in the British and American corpora, it shows a continuous growing pattern with a rapid increase from around 2015 in the German corpus. This might explain the strong correlation there. I4.0 technologies contribute with different strength to the spread of the term I4.0, which supposes that the weaker the correlation, the technology is less associated with I4.0. SCM is in weak correlation with I4.0 in the German corpus ($r_D=0.088$) and relatively weakly supports I4.0 in the American and British ($r_{USA}=0.43$, $r_{Br}=0.33$) corpora. The German corpus gives strong correlation for each element except AI and SCM. VR is negatively correlated with I4.0 in the German corpus, although the correlation is relatively weak in all three corpora.

Considering only I4.0 in all three corpora between 2000 and 2019, the spread is exponential, the steepest being in the American corpus (Figure 3).

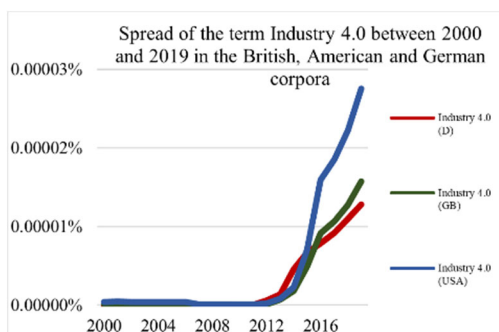


Figure 3

Spread of the terms 'Industry 4.0' in the British, American and German corpora (developed by author)

The significance of the elements was also tested in the regression analysis. Different technologies proved to be significant in the different corpora (Table 2).

Table 2

I4.0 technology significances in regression to I4.0 as dependent variable

I4.0 elements	Significance level in regression model (p values)		
	<i>American</i>	<i>British</i>	<i>German</i>
<i>Augmented Reality</i>	0.1596	0.7872	0.2508
<i>Cloud Computing</i>	0.0060	0.0107	0.0144
<i>IoT</i>	0.0987	0.0928	0.0000
<i>Big Data Analysis</i>	0.0000	0.0000	0.0000
<i>3D Printing</i>	0.0000	0.0001	0.6972
<i>Robotics</i>	0.9569	0.0200	0.0510
<i>Artificial Intelligence</i>	0.9332	0.4324	0.1219
<i>Virtual Reality</i>	0.0830	0.0470	0.9589
<i>SCM</i>	0.2952	0.5800	0.6459

Dependent variable: Industry 4.0

As Table 2 shows the usage of the term Cloud Computing and Big Data analysis contribute significantly to the spread of I4.0 in all three corpora, while the spread of the term 3D Printing significantly boosts the spread of I4.0 in the American and British corpora. Robotics and VR are significant in the British corpus. In each case the p value is less than 5% (Table 2). Surprisingly, the spread of IoT boosts the spread of I4.0 solely in the German Corpus at $p=0.05$, and Robotics is significant only at $p=0.10$ in the German corpus while IoT is significant at $p=0.10$ in the American and British corpora. Should the spread of viruses be considered, the significant elements in the German corpus are presumed to be more influential in the familiarity with Industry 4.0 among SMEs in V4 countries, Serbia and Bulgaria.

Consequently, based on the spread of narratives, the correlation and regression between I4.0 and I4.0 technologies, it is hypothesized that:

H1₁: The familiarity with Cloud Computing, Big Data Analysis, 3D printing, Robotics, IoT and VR contribute positively to the familiarity with I4.0 among SMEs in the V4 countries, Serbia and Bulgaria, they associate these technologies with I4.0 and the familiarity with I4.0 among SMEs can be predicted with high probability.

H1₂: AR, AI and SCM are not considered elements of I4.0 among SMEs in the V4 countries, Serbia and Bulgaria. SMEs do not associate these technologies with I4.0.

Namely, the SMEs that are familiar with the above technologies are more probable to be familiar with the term I4.0 and use it in their daily business operations. The following section presents the SME responses and the results of the research based on the survey among SMEs in the V4 countries, Serbia and Bulgaria.

5 Familiarity with I4.0 and its Elements among SMEs

5.1 Demographic Profile of SMEs

A total number of 635 responses were collected from the V4 countries, Serbia and Bulgaria. After filtering large companies 535 valid SME responses were analyzed. In the research Hungary represents 20.56% of the SMEs, Slovakia gives 17.01% while the other countries take around 15% of the responses. The country distribution is not significantly different, they are equally represented in the sample ($\text{Chi}^2=6.492$, $p=0.261$).

Table 3 presents the demographic profile of the responding business professionals and SMEs.

Table 3
SME Demographic Profile

Personal characteristics	n=535	Distribution of respondents (%)	Business characteristics	n=535	Distribution of respondents (%)
<i>Age</i>			<i>SME size (number of employees)</i>		
18-30	117	21.87	Micro	243	45.42
31-45	193	36.07	Small	139	25.98
46-60	180	33.64	Medium-sized	153	28.60
> 61	45	8.41			
<i>Gender</i>			<i>The dominating sector of the company</i>		
Male	326	60.93	Production	161	30.09

Female	204	38.13	Services	95	17.76
No wish to answer	5	0.93	Trade	279	52.15
<i>Position</i>			<i>Company age (years)</i>		
The owner	192	35.96	Up to 2 years	50	9.35
Senior manager	90	16.85	3-5	52	9.72
Manager	108	20.22	6-10	104	19.44
Employee	144	26.97	11-20	132	24.67
			>21	197	36.82

One fifth of the respondents are under 30 while over two thirds of the respondents are aged between 31 and 60. Sixty percent of the respondents are male (60.93%) and 38.13% of them are female in the sample. In terms of their position, almost an equal number of owners and managers responded, 35.96% and 37.07%, respectively, while 26.97% of the respondents were employees. In terms of business characteristics, the largest proportion is that of micro enterprises (45.42%), Small enterprises give a quarter of the sample (25.98%) and medium-sized enterprises made up 28.6% of the sample. More than 60% of the enterprises surveyed are more than 11 years old while 9.35% and 9.72% are less than 2 years old or are between 3 and 5 years. The remaining 20% are between 6 and 10 years old. More than half of the enterprises in the sample are belong to the Trade sector, one third to the Production sector and 17.76% to the Services sector.

Based on the distributions, micro enterprises ($\text{Chi}^2=35.723$, $p=0.000$), more mature enterprises ($\text{Chi}^2=140.262$, $p=0.000$) and businesses in the services sector ($\text{Chi}^2=97.450$, $p=0.000$) are more represented in the sample. At the same time, owners and managers are also overrepresented ($\text{Chi}^2=9.843$, $p=0.007$), which fits the analysis well since the introduction of I4.0 technologies and I4.0 depends on the management of an enterprise to a great extent.

5.2 Country Comparison on Familiarity with I4.0

Figure 4 shows that SMEs in the participating countries are differently familiar with I4.0. In total, 52.9% of SMEs are not familiar with the term I4.0, less than half (47.1%) of them know the term.

Over two thirds of the SMEs in the Czech Republic are familiar with I4.0, 61% in Slovakia, while half of the Serbian SMEs know the term. Hungary is the fourth with 40% [28], and less than 40% of SMEs in Poland and Bulgaria are familiar with I4.0.

There is a significant difference between the countries in terms of familiarity with I4.0 ($\text{Chi}^2 = 30.346$, $p=0.000$, while Cramer's $V=0.24$, $p=0.000$). As Table 4 shows, SMEs in the Czech Republic are significantly more familiar with the term I4.0 than in Hungary (B), Poland (D) and Bulgaria (F), while Slovakian SMEs do not differ significantly from the Hungarian (B), Serbian (E) and the Czech SMEs (A).

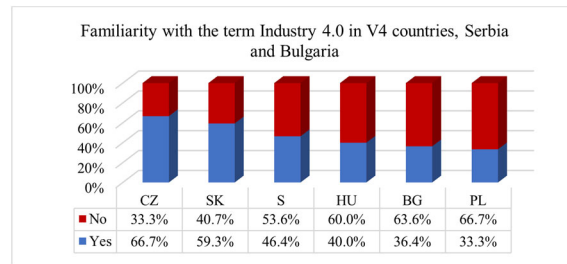


Figure 4

Familiarity with I4.0 in V4 countries, Serbia and Bulgaria (developed by author)

Table 4

Significant differences of I4.0 familiarity between SMEs by country

Are you familiar with the term INDUSTRY 4.0 (%)							
Country	CZ (A)	H (B)	SK (C)	PL (D)	S (E)	BG (F)	Total
<i>Yes</i>	66.7 B D F	40	59.3 D F	33.3	46.4	36.4	47.1
<i>No</i>	33.3	60 A	40.7	66.7 A C	53.6	63.6 A C	52.9
<i>Total</i>	100	100	100	100	100	100	100

Results are based on two-sided tests. For each significant pair, the key of the category with the smaller column proportion appears in the category with the larger column proportion. Significance level for upper case letters (A, B, C): 0.05

5.3 I4.0 Technologies to Determine Familiarity with I4.0 among SMEs

The familiarity with I4.0 technologies was also surveyed on a Likert scale ranging from 1—‘never heard about it’ to 5—‘have heard and use it in everyday business operations’. Further response options were 2—‘have heard but never used’, 3—‘have heard and do plan to use it’ and 4—‘have heard and use it occasionally’. With these statements the research strives to explore why SMEs in these countries show a low-level of familiarity with I4.0 on average and seeks to find a cause-and-effect relationship between the familiarity of I4.0 and its pillars. Table 5 presents that SMEs are familiar with cloud computing services (Mean≈3), half of the SMEs have heard about it and plan to use it in their business processes (Median=3) but based on the Median and Mode values the majority of SMEs have heard about the term but have never used it.

The worst case in these countries is the unfamiliarity with Big Data Analysis, as most SMEs most have not heard of and never used this possibility in I4.0 (Mode=1). Based on the descriptive results the familiarity with the technologies is low, most of the responding SMEs have not heard about the technology, or have heard but never used them.

Table 5
SME Familiarity with the Elements of I4.0

I4.0 elements	n	Mean	Median	Mode	SD	IQR
<i>Cloud computing services</i>	530	2.93	3	2	1.440	2
<i>AI</i>	531	2.51	2	2	1.155	1
<i>Supply Chain Management</i>	530	2.50	2	2	1.152	1
<i>3D printing and robotics</i>	530	2.48	2	2	1.165	1
<i>VR</i>	529	2.46	2	2	1.116	1
<i>IoT</i>	526	2.44	2	2	1.277	2
<i>Big data analysis</i>	531	2.38	2	1	1.231	2
<i>AR</i>	528	2.23	2	2	1.110	2

Supposedly, as hypothesised earlier, an increase in the awareness of the individual technologies could boost familiarity with I4.0 and consequently might lead to a better understanding and higher rate of usage of these technologies, leading to a positive contribution to digitalisation and business recovery after COVID-19.

6 Contribution of Industry 4.0 Technologies to the Familiarity with I4.0 among SMEs

The research conducted among SMEs found that in average over 50% of SMEs are not familiar with the term I4.0, however, in Poland and Slovakia the familiarity is over 59%. According to the aim of the research the familiarity with an awareness of I4.0 and its elements need to be boosted, so it is to be investigated what leads to the familiarity with I4.0 and how it can be changed to the positive. Which elements contribute positively, and which hinders the spreading of I4.0?

All the 535 responses were used for the analysis. The analysis did not differentiate between the countries in order to get a general view in the V4 region, Serbia and Bulgaria. The data were cleaned, meaning that all records with no response for the analyzed questions, and all records with unengaged responses were deleted. The missing values were replaced by the Median due to the Likert scale used for rating. Finally, a total number of 436 responses remained for analysis using correlation, logistic regression with enter and with the stepwise Wald method.

At first, correlation was checked to see whether there is a relationship between the familiarities with the technologies (Table 6). While the I4.0 technologies are relatively weakly correlated with I4.0 ($0.202 < r < 0.374$), some of them are in strong correlation pairwise (e.g $r=0.709$ in AR–VR relation).

Figure 5 displays the weights of the elements in the model, showing that familiarity with Big Data Analysis, IoT and 3D printing and Robotics would rather determine the SMEs' familiarity with I4.0.

Table 6
Correlation of the elements of I4.0

	<i>I4.0</i>	<i>Cloud Computing Services</i>	<i>Big Data Analysis</i>	<i>3D Printing and Robotics</i>	<i>IoT</i>	<i>VR</i>	<i>AR</i>	<i>SCM</i>
<i>Cloud Computing Services</i>	0.202							
<i>Big Data Analysis</i>	0.374	0.53						
<i>3D Printing and Robotics</i>	0.309	0.295	0.409					
<i>IoT</i>	0.345	0.383	0.535	0.453				
<i>VR</i>	0.294	0.311	0.466	0.574	0.509			
<i>AR</i>	0.332	0.386	0.568	0.523	0.603	0.709		
<i>SCM</i>	0.299	0.286	0.438	0.409	0.408	0.453	0.404	
<i>AI</i>	0.235	0.321	0.489	0.505	0.459	0.622	0.595	0.454

Each correlation is significant at the 0.01 level (2-tailed).

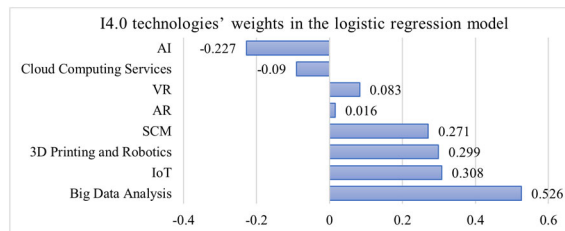


Figure 5

I4.0 elements' weights on the logistic regression model

Surprisingly Cloud Computing services has a negative weight (-0.09), implying that SMEs do not associate Cloud Computing services with I4.0. It is, despite the fact that these are the most widespread among SMEs in these countries; 58.4% of the SMEs in the survey have heard about it, plan to use it or use the technology. AI has the lowest weight in the model (-0.227), while its familiarity is not outstandingly low among SMEs (41.74%). SCM is the second well known term and technology, however, less than half of the participating SMEs, 47.25% of them, have heard about the technology and plan to use it or use it.

Since more than half of the SMEs in the V4 countries, Serbia and Bulgaria marked that they were not familiar with the term I4.0 logistics regression was used to predict the knowledge of which pillars of I4.0 used in the research could contribute to the better familiarity and knowledge of I4.0, i.e. which elements are significant for SMEs to be acquainted with and be promoted more. The sample contained independent observations, and no multicollinearity problem occurred as tolerance values ranged between 0.374 and 0.7, while VIF values ranged between 1.43 and 2.68 for the predictors [29].

With all the elements entered in the model, it classified 68.8% of the responses well, increasing considerably from the 50.7% in the sample (with a precision of 70.3%), while the Wald method resulted in an accuracy of 69% (with a precision of 70.6%), i.e. the stepwise method has slightly improved on the model. Based on the Hosmer and Lemeshow Test both methods resulted in a model that fits the original data well at $p=0.01$ (Enter method: $\text{Chi}^2=17.491$, $p=0.025$ and Wald method: $\text{Chi}^2=17.851$, $p=0.022$). According to researchers [30] the conventional significance level $p=0.01$ can be used with large samples (over 300) if alpha is fixed since the probability of Type II error decreases.

Both the Enter and the Wald methods gave a medium effect size. Nagelkerke's Pseudo R^2 being 0.262 and 0.253, respectively, indicating that the non-significant elements added some explanation why SMEs are familiar or not familiar with the term I4.0. The Chi2 test of Model Coefficient proved to be significant (-2log likelihood decreased significantly, $p=0.000$) so the use of the I4.0 technologies as independent variables is justified (Table 7).

Table 7
Model Summary and Pseudo R^2

Method	Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
Enter	1	509,070 ^a	0.196	0.262
Wald	4	512,563 ^b	0.190	0.253

Table 8 presents how the familiarity with I4.0 technologies used in the survey contribute to the familiarity with I4.0 among the SMEs in the V4 countries, Serbia and Bulgaria.

Table 8
Logistic Regression model (Enter Method) – contribution of I4.0 elements to the familiarity with I4.0

I4.0 elements	B	S.E.	Wald	df	Sig.	Exp(B)
Cloud computing services (x_1)	-0.063	0.090	0.491	1	0.483	0.939
Big Data analysis (x_2)	0.428	0.125	11.716	1	0.001	1.533
3D printing and Robotics (x_3)	0.261	0.123	4.543	1	0.033	1.298
IoT (x_4)	0.240	0.111	4.673	1	0.031	1.271
Virtual Reality (x_5)	0.077	0.161	0.227	1	0.634	1.080
Augmented Reality (x_6)	0.152	0.169	0.809	1	0.368	1.164
Supply Chain Management (x_7)	0.239	0.115	4.361	1	0.037	1.270
Artificial Intelligence (x_8)	-0.201	0.136	2.182	1	0.140	0.818
Constant	-2.732	0.385	50.316	1	0.000	0.065

According to the Enter method, when SMEs are familiar with *Big Data analysis*, *3D printing and robotics*, *IoT* and *SCM* they are predicted to be familiar with Industry 4.0 as these elements are significant in the model. However, the other elements, namely *Cloud Computing services*, *VR*, *AR*, and *AI* have proved to be insignificant in the model. The Exp(B) value being larger than 1 for the significant

elements, i.e Big Data analysis improves the prediction by 53.3%, 3D Printing and Robotics by 29.8%, IoT by 27.1% and SCM by 27%. Two of the non-significant elements, VR and AR also increase the prediction by 8% and 16.4%, respectively. Surprisingly, two of the non-significant elements (Cloud Computing services (-6.1%) and AI (-18.2%)) seem to contrast with the familiarity with I4.0 among the participating SMEs. These two technologies are not associated with I4.0 among the responding SMEs.

Based on the coefficients, the logistics regression function is the following using all the technologies in the model,

$$\log\left(\frac{p}{1-p}\right) = -2.732 - 0.063x_1 + 0.428x_2 + 0.261x_3 + 0.240x_4 + 0.077x_5 + 0.152x_6 + 0.239x_7 - 0.201x_8 \quad (1)$$

while the probability of SMEs' familiarity with I4.0 is given by the following equation:

$$P(I4.0) = \frac{1}{1 + e^{-(-2.732 - 0.063x_1 + 0.428x_2 + 0.261x_3 + 0.240x_4 + 0.077x_5 + 0.152x_6 + 0.239x_7 - 0.201x_8)}} \quad (2)$$

The Wald method leaves the four previously significant elements in the model (Table 9), namely *Big Data Analysis*, *3D printing* and *Robotics*, *IoT* and *SCM*.

Table 9

Logistic Regression model (Wald Method) – contribution of I4.0 elements to the familiarity with I4.0

<i>I4.0 elements</i>	<i>B</i>	<i>S.E.</i>	<i>Wald</i>	<i>df</i>	<i>Sig.</i>	<i>Exp(B)</i>
Big Data analysis (x_1)	0.401	0.109	13.447	1	0.000	1.493
3D printing and Robotics (x_2)	0.260	0.111	5.469	1	0.019	1.297
IoT (x_3)	0.255	0.103	6.123	1	0.013	1.291
Supply Chain Management (x_4)	0.218	0.111	3.849	1	0.050	1.244
Constant	-2.828	0.357	62.586	1	0.000	0.059

However, in this model the contribution to the familiarity with I4.0 among SMEs are different, the Exp(B) values show that Big Data analysis improves the prediction by 49.3%, 3D Printing and Robotics by 29.7%, IoT by 29.1% and SCM by 24.4%. This could be explained by the correlations between the significant and non-significant elements of I4.0 (see Table 6).

Based on the coefficient values the logistic regression function is the following using the four significant technologies,

$$\log\left(\frac{p}{1-p}\right) = -2.828 + 0.401x_1 + 0.260x_2 + 0.255x_3 + 0.218x_4 \quad (2)$$

while the probability of SMEs' familiarity with I4.0 is given by the following equation with the four significant technologies:

$$P(I4.0) = \frac{1}{1 + e^{-(-2.828 + 0.401x_1 + 0.260x_2 + 0.255x_3 + 0.218x_4)}} \quad (2)$$

The four non-significant elements in this model were Cloud Computing Services, VR, AI and AR, their elimination resulted in eliminating the strong correlations between the elements in the original correlation matrix.

The following section will compare the results and will discuss the findings. Finally, the hypotheses will be evaluated and the research questions answered.

Conclusions

The research aim was to explore and predict the familiarity with I4.0 by the I4.0 technologies among SMEs in the V4 countries, Serbia and Bulgaria and compare the findings with the spreading if these narratives in various corpora. The research results show that there are similarities but also discrepancies in the list of I4.0 technologies that contribute positively to and can predict well the familiarity with I4.0 among SMEs and the technologies that spread similarly as I4.0 in the narratives. The results imply that certain I4.0 technologies are not associated with Industry 4.0 among SMEs and do not occur together with Industry 4.0 in the narratives.

Based on the results the first hypothesis, according to which

H1₁: The familiarity with Cloud Computing, Big Data Analysis, 3D printing, Robotics, IoT and VR contribute positively to the familiarity with I4.0 among SMEs in the V4 countries, Serbia and Bulgaria, they associate these technologies with I4.0 and the familiarity with I4.0 can be predicted among SMEs with high probability, is partially confirmed. Big Data Analysis, 3D printing and Robotics, IoT and SCM significantly predict the familiarity with I4.0 among SMEs, meaning that the higher the familiarity with these terms the higher the probability of SMEs being familiar with the term I4.0 and use the technology in their daily business operations. Cloud Computing services, VR, AR and AI do not predict significantly the familiarity with I4.0 among SMEs. Cloud Computing services contribute negatively implying that SMEs do not associate the technology with I4.0.

The hypothesis stating that

H1₂: AR, AI and SCM are not considered elements of I4.0 among SMEs in the V4 countries, Serbia and Bulgaria. SMEs do not associate these technologies with I4.0, can also be partially accepted, since SCM proved to be a significant predictor in the model while AR and AI are non-significant technologies when used for predicting I4.0 familiarity among SMEs in the V4 countries, Serbia and Bulgaria. Furthermore, AI proved to negatively contribute to the prediction, implying that SMEs do not associate the technology with I4.0. SCM, however, proved to be the fourth significant predictor that improves the familiarity with I4.0 by 27% among the participating SMEs. The results for both hypotheses align with the findings of [1, 8, 31, 32].

The partial acceptance of the above two hypotheses gives answers to the first Research Question, as expect Cloud Computing services and AI, the selected I4.0 technologies contribute positively to the familiarity with I4.0 among the responding

SMEs and four of them significantly predict the familiarity with I4.0. SMEs associate these technologies with Industry 4.0, except Cloud Computing services and AI, despite the fact that Cloud computing is the most frequently used services.

The present research covered the spread of these narratives in three corpora and found that the occurrences of AR, Cloud Computing, IoT, Big Data analysis, and 3D printing are highly correlated with the occurrences of I4.0 in all three corpora, while Robotics, AI, SCM and VR show strong correlation in certain corpora. VR is negatively correlated with I4.0 in the German corpus, implying no association with I4.0. On the other hand, apart from Cloud Computing and Big Data analysis, the other terms do not significantly spread the same way as I4.0 in all three corpora. The spreading and usage of the terms 3D printing and Robotics are similar with that of I4.0 in two corpora, while IoT and VR spread similarly as I4.0 in only one corpus. Consequently, responding Research Question 2, AR, AI and SCM do not significantly influence the spread of the term I4.0 in the narratives.

Finally, Research Question 3 seeks similarities in the spreading of the terms and SMEs' familiarity with I4.0 and its technologies. Based on the results, the research question can be partially answered. AI and AR are not good predictors of I4.0 in either the narratives or among SMEs, they were insignificant in both models (Table 10), implying that SMEs do not associate these technologies with I4.0.

Table 10
Comparison of significance of I4.0 technologies among SMEs and in the narratives

<i>I4.0 technologies</i>	<i>Significant among SMEs</i>	<i>Significant in the narratives</i>
<i>Big data analysis</i>	+	+
<i>3D printing and robotics</i>	+	+
<i>IoT</i>	+	+*
<i>Supply Chain Management</i>	+	—
<i>Cloud computing services</i>	—	+
<i>VR</i>	—	+**
<i>AI</i>	—	—
<i>AR</i>	—	—

Significance level is $p < 0.05$ if not marked otherwise

*Significant in American and British corpora at $p < 0.1$, significant in German corpus at $p < 0.05$

**Significant in American corpora at $p < 0.1$, significant in British corpus at $p < 0.05$, not significant in German corpus

Big Data Analysis, 3D printing and Robotics, as well as IoT proved to be significant in both models, the familiarity of these terms predicts well the familiarity with I4.0 among SMEs while they spread similarly in the narratives. SMEs associate these terms with I4.0, i.e. if they are familiar with these terms, they are predicted to be familiar with I4.0. SCM, Cloud Computing and VR behave differently, and while SCM is a positive contributor to the familiarity with I4.0 the term does not occur together with I4.0 in a significant volume. The same applies to Cloud Computing and VR but in a reverse mode, they occur together with I4.0 in the narratives but do not predict the familiarity with I4.0 among the participating SMEs in the V4 countries, Serbia and Bulgaria.

Consequently, if SMEs are to be strengthened to be digitalized and use Industry 4.0 technologies, the technologies that proved to be insignificant should be popularized, promoted and introduced so that SMEs, their owners and managers would learn about these technologies and would introduce them in their business practices to a greater extent. Without familiarity with I4.0 technologies it is hard for SMEs to digitalize and improve on the integration of these technologies. Therefore, the digitalization of the sector and the spreading of I4.0 solutions could be improved and would help SMEs to increase their competitiveness, efficiency and business performance. The results align with the findings in [4, 6] as well.

The research has its limitations, since the sampling method did not allow us to have a fully representative sample, however, the sample size was large enough to make it possible to draw conclusions on the behavior of SMEs. The researchers are planning to gather more data and develop further research models to investigate the digitalization level of SMEs that would further support the use of I4.0 technologies at SMEs in these countries.

Acknowledgement

This paper was supported by the International Visegrad Fund, project number 22110036, titled "Possibilities and barriers for Industry 4.0 implementation in SMEs in V4 countries and Serbia".

References

- [1] M. Cugno, R. Castagnoli, G. Büchi and M. Pini, "Industry 4.0 and production recovery in the covid era," *Technovation*, vol. 114, p. 102443, 2022.
- [2] D. Horváth and R. Z. Szabó, "Driving forces and barriers of Industry 4.0: Do multinational and small and medium-sized companies have equal opportunities?," *Technological Forecasting & Social Change*, vol. 146, pp. 119-132, 2019.
- [3] N. Abu Hasan, M. Abd Rahim, S. H. Ahmad and M. Meliza, "Digitization of Business for Small And Medium-Sized Enterprises (SMEs)," *Environment-Behaviour Proceedings Journal*, vol. 7, no. 19, pp. 11-16, 2022.
- [4] J. Bleicher and H. Stanley, "Digitization as a catalyst for business model innovation a three step approach to facilitating economic success," *Journal Business Management*, vol. 4, no. 2, pp. 62-71, 2018.
- [5] A. Spieske and H. Brikel, "Improving supply chain resilience through industry 4.0: a systematic literature review under the impressions of the COVID-19 pandemic," *Computers and Industrial Engineering*, vol. 158, p. 107452, 2021.

- [6] C. Acioli, A. Scavarda and A. Reis, "Applying Industry 4.0 technologies in the COVID-19 sustainable chains," *International Journal of Productivity and Performance Management*, vol. 70, no. 5, pp. 998-1016, 2021.
- [7] N. Melluso, S. Fareri, H. Fantoni, A. Bonaccorsi, F. Chiarello, E. Coli, V. Giordano, P. Manfredi and S. Manafi, "Lights and shadows of COVID-19, Technology and Industry 4.0," *arXiv Preprint ArXiv*, 2020.
- [8] A. Zaušková, A. Kusá, M. Kubovics, Š. Simona and R. Miklenčičová, "Awareness of Industry 4.0 and its tools across the V4," *Serbian Journal of Management*, vol. 1, no. 17, pp. 253-264, 2022.
- [9] Z. Bánhidi, M. Tokmergenova and I. Dobos, "International benchmarking and methodological framework for the development of the digital economy," *Information Society*, vol. XXII, no. 1, pp. 9-28, 2022.
- [10] D. Buhr, Social Innovation Policy for Industry 4.0., Friedrich-Ebert-Stiftung, 2017.
- [11] D. Fettermann, C. G. Sá Cavalcante, T. D. de Almeida and G. L. Tortorella, "How does Industry 4.0 contribute to operations management?," *Journal of Industrial and Production Engineering*, vol. 35, no. 4, pp. 255-268, 2018.
- [12] A. Tick, R. Saáry and J. Kárpáti-Daróci, "The effect of digitalisation on sustainable operation of SMEs – the case of Hungary," in *Possibilities and barriers for Industry 4.0 implementation in SMEs in V4 countries and Serbia*, Bor, University of Belgrade, 2022, pp. 121-150.
- [13] J. Clerck, "Digitization, digitalization and digital transformation: The differences.," 25 July 2017. [Online]. Available: [https://www.i-scoop.eu/digital-transformation/digitization-digitalization-digital-transformation-disruption/..](https://www.i-scoop.eu/digital-transformation/digitization-digitalization-digital-transformation-disruption/) [Accessed 10 May 2022].
- [14] R. A. Şerban, "The Impact of Big Data, Sustainability, and Digitalization on Company Performance," *Studies in Business and Economics*, vol. 12, no. 3, pp. 181-189, 2017.
- [15] Á. Gubán and Á. Sándor, "Opportunities to measure the Digital Maturity of SMEs," *Budapest Management Review*, vol. 52, no. 3, pp. 13-28, 2021.
- [16] R. Saáry, J. Kárpáti-Daróczi and A. Tick, "Profit or less waste? Digitainability in SMEs - a comparison of Hungarian and Slovakian SMEs," *Serbian Journal of Management*, vol. 17, no. 1, pp. 33-49, 2022.

- [17] A. Sestino, M. I. Prete, L. Piper and G. Guido, "Internet of Things and Big Data as enablers for business digitalization strategies," *Technovation*, vol. 98, no. C, 2020.
- [18] T. Atobishi, M. Bahna, K. Takács-György and C. Fogarassy, "Factors Affecting the Decision of Adoption Cloud Computing Technology: The Case of Jordanian Business Organizations," *Acta Polytechnica Hungarica*, vol. 18, no. 5, pp. 131-154, 2021.
- [19] P. Baranyi, T. Csapó, T. Budai and G. Wersényi, "Introducing the Concept of Internet of Digital Reality - Part I," *Acta Polytechnica Hungarica*, vol. 18, no. 7, pp. 225-240, 2021.
- [20] M. Rüßmann, M. Lorenz, P. Gerbert, M. Waldner, J. Justus, P. Engel and M. Harnisch, "Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries," Boston Consultation Group, 2015.
- [21] R. J. Schiller, *Narrative Economics, How stories go viral and drive major economic events*, Princeton, NJ: Princeton University Press, 2020.
- [22] A. Field, *Discovering Statistics with IBM SPSS Statistics*, Newbury Park, CA: Sage, 2013.
- [23] Y. Lin, J. Michel, E. L. Aiden, J. Orwant, W. Brockman and S. Petrov, "Syntactic annotations for the Google books Ngram corpus," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, 1998.
- [24] A. Tick and J. Beke, "Online, Digital or Distance? – Spread of Narratives in ICT-supported Education," *Journal of Higher Education Theory and Practice*, vol. 21, no. 6, pp. 15-31, 2021b.
- [25] Gartner, "Gartner Research," 2018. [Online]. Available: <https://www.gartner.com/en/documents/3887767/understanding-gartner-s-hype-cycles>. [Accessed 9 January 2021].
- [26] J. Kenn, M. Raskino and B. Burtin, "Understanding Gartner's Hype Cycles," Gartner Inc., 2017.
- [27] I. Milošević, S. Arsić, M. Glogovac, A. Rakić and J. Ruso, "Industry 4.0: Limitation or benefit for success?," *Serbian Journal of Management*, vol. 17, no. 1, pp. 85-98, 2022.
- [28] A. Tick, J. Kárpáti-Daróci and R. Saáry, "'To familiarise or not to familiarise' - industry 4.0 implementation in SMEs in Hungary," in *Possibilities and*

barriers for Industry 4.0 implementation in SMEs in V4 countries and Serbia, M. Trumić, Ed., Bor, University of Belgrade, 2022, pp. 35-61.

- [29] J. F. J. Hair, R. E. Anderson, R. L. Tatham and W. C. Black, *Multivariate Data Analysis*, 3 ed., New York: Macmillan, 1995.
- [30] J. H. Kim and P. I. Ji, "Significance Testing in Empirical Finance: A Critical Review and Assessment," *Journal of Empirical Finance*, vol. 34, no. C, pp. 1-14, 2015.
- [31] I. Milošević, "The effects of familiarity of Industry 4.0 technologies on behaviour intention of SMEs in Serbia," in *Possibilities and barriers for Industry 4.0 implementation in SMEs in V4 countries and Serbia*, I. Mihajlović, Ed., Bor, University of Belgrade, 2022, pp. 181-206.
- [32] S. Arsić, "Industry 4.0 technologies: Results of an International Study in SMEs," in *Possibilities and barriers for Industry 4.0 implementation in SMEs in V4 countries and Serbia*, I. Mihajlović, Ed., Bor, University of Belgrade, 2022, pp. 62-83.

DDoS Attack Intrusion Detection System Based on Hybridization of CNN and LSTM

Ahmet Sardar Ahmed Issa¹, Zafer Albayrak²

¹ Department of Computer Engineering, Karabuk University, Karabuk, Turkey, 1928126532@ogrenci.karabuk.edu.tr

² Department of Computer Engineering, University of Applied Sciences, Sakarya, Turkey, zaferalbayrak@subu.edu.tr

Abstract: A distributed denial-of-service (DDoS) attack is one of the most pernicious threats to network security. DDoS attacks are considered one of the most common attacks among all network attacks. These attacks cause servers to fail, causing users to be inconvenienced when requesting service from those servers. Because of that, there was a need for a powerful technique to detect DDoS attacks. Deep learning and machine learning are effective methods that researchers have used to detect DDoS attacks. So, in this study, a novel deep learning classification method was proposed by mixing two common deep learning algorithms, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). The NSL-KDD dataset was used to test the model. This method architecture consists of seven layers to achieve higher performance compared with traditional CNN and LSTM. The proposed model achieved the highest accuracy of 99.20% compared with previous work.

Keywords: DDoS attacks, Deep learning, CNN, LSTM, NSL-KDD

1 Introduction

Currently, networks are very important for everyone because they present many features and one of the most important is the resource sharing. A network is defined as connecting two or more nodes, regardless of which nodes may be a computer, server, mobile phone, etc. The merging of computer networks in worldwide has formed the important technology is Internet that is indispensable. Today, the Internet is becoming highly vulnerable to many forms of cyberattacks. The most dangerous kind of cyber-attack is distributed denial of service (DDoS) attack [1]. In a DDoS and Denial of Service (DoS) scenario, the attacker tries to flood the host's service, making the host unavailable to legitimate users [2]. Generally, DoS attack is initiated from a single infected device or virtual machines utilizing an Internet connection whereas DDoS attacks are initiated from many different infected devices or virtual machines to overload the target systems [3]. Even if an organization has

implemented a typical security system, it will be virtually impossible to protect against a DDoS attack because of the large number of attacks in the same time and the attack is improved very fast [4]. This is largely due to the fact that DDoS attacks try to simulate normal traffic but have increased exponentially. A DDoS attack targeted GitHub [5], NETSCOUT Arbor [6], and Amazon platform [7]. These are some of the biggest DDoS attacks in the world in recent years. This has led to huge losses in industry and government globally due to DDoS attacks in recent years [8]. These problems are caused by the devices interacting with remote applications, which allows malicious agent to control the devices. The main reasons for the increase in DDoS attacks are that implementing DDoS attacks is easy and simple, does not require a great deal of technological understanding on the part of the attacker, and there were many platforms and software that could be used to coordinate the attack [9]. In general, the attackers use many devices called botnet in the DDoS attacks quickly [10].

Figure 1 shows how the attacker controls the system by connecting to the control server [11]. An efficient server with abundant resources like memory, processing power, and bandwidth is called a control server. In addition, the handlers of Botnets, also known as Agents, are the ones who receive commands from attackers. All of the attacker's commands go to the victims through these botnets. Even if malware is already installed on the compromised computer, the owner doesn't know whether it is part of a Botnet. Proxy servers are commonly used by attackers to distribute malware, execute DDoS attacks, and carry out other attacks on their victims [12]. DDoS attacks can be separated into two types. They are the application layer and the network layer [13], or they can be divided into three types [14]. At the first, volume-based attacks include UDP floods and other spoofed-packet floods. Secondly, protocol attacks cover SYN floods, Smurf DDoS, Ping of Death, fragmented packet attacks, and different types of DDoS. Lastly, application layer attacks include some advanced techniques such as SIDDOS, HTTP GET/POST floods. Security hackers are daily developing new techniques for evading defensive measures and evading detection. Therefore, daily improvement intrusion detection systems (IDS) are needed [15]. IDS is the system that can recognize a new DDoS speedily and without the need for human assistance. To increase the adaptability and accuracy of an IDS, an IDS-based machine learning has been used over the past few decades [16]. In addition, these systems are hampered by their essential reliance on previous information, their slowness, and their failure to learn from vast volumes of data. Their ability to learn isn't always powerful, either [17]. Deep learning models have recently been deployed to recognize detecting troubles, considerably increasing their chances of success [18].

In ML, deep learning (DL) is a new field that has emerged recently, the concept of which came from neural networks that mimic the human brain [19]. It has achieved successes in many areas such as speech recognition, image processing, language translation, and the IDS field [20]. Deep learning-based IDS has been found to be more effective at recognizing than traditional machine learning in several recent studies.

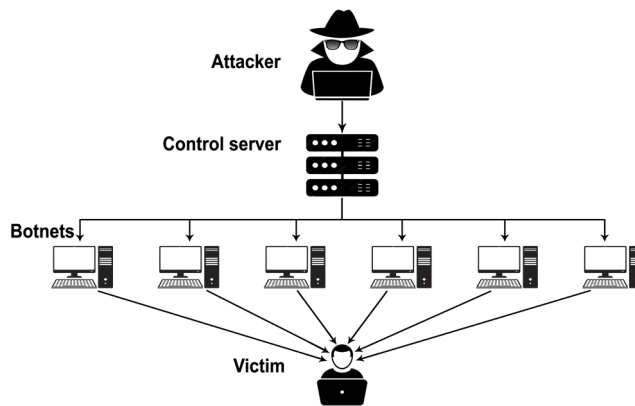


Figure 1
DDoS attack

Although deep learning algorithms analyze deeper and faster network data, none of these algorithms individually can reflect the correlation of features between multidimensional features. Another issue is that training datasets with false labels aren't taken into consideration [21].

In this paper, to solve the problems we discussed before, we proposed a new DL model that combines the Convolutional Neural Network (CNN) layers for feature extraction from input automatically [22] with the Long Short Term Memory neural network (LSTM) for predicting sequence [23]. In the proposed model design has seven layers to achieve high performance compared with each CNN and LSTM individually. The performances of in the proposed model, CNN, and LSTM were compared according to four metrics. These four metrics are accuracy, precision, recall, and F1 score. The model achieved the best accuracy among other state-of-the-arts applied to the same dataset, the NSL-KDD dataset. Other sections of the paper are arranged as follows. Sect. 2 deliberates about and concludes the related work. Sect. 3 concludes by discussing the NSL-KDD dataset and the methods used in this paper. Sect. 4 provides information on the evaluation criteria being used. Sect. 5 contains information about the experiments and the paper results. Finally, Sect. 6 is the paper's conclusion and future works.

2 Related Work

Recently, machine learning and deep learning algorithms have had great success in predicting DDoS attacks. In 2017, a feature selection approach by authors in [24] is utilized to facilitate successful intrusion detection system with machine learning. This method is the combination between DDoS Characteristic Features (DCF) and Consistency Subset Evaluation (CSE). ANN and black hole optimization approach is proposed by Kushwah and Ali [25] as a model in cloud computing for detecting

DoS attacks. Researcher in [26] proposed the Dendritic Cell Algorithm (DCA), an AIS-based algorithm for identifying most frequent denial of service attack and distributed DoS attacks that impact network communication to analyze the suggested detection method. In 2018, the researchers in [27] suggested a method based on genetic algorithm (GA) to identify DDoS attacks in cloud platform. This approach was to optimize Bernoulli Naïve Bayes BNB classifier using genetic algorithm. The H2O data mining tool was used in implementing algorithms, and a comparison of the algorithms' accuracy in DDoS attacks detection was performed [28]. Entropy estimation, co-clustering, information gain ratio (IGR) for features selection, and the Extra-Trees ensemble classifying algorithm are utilized to identify DDoS attacks; called Semi-supervised approach [29]. Network traffic data entropy is estimated and analyzed over time-based sliding windows. The second step the co-clustering algorithm divide network traffic time to three clusters when the network entropy reaches its limits. The third step is features selection represented by IGR and lastly classification algorithm is Extra-Trees ensemble. In 2019, Anjum and Shreedhara in [30] proposed an approach to improve the performance compared to the supervised and unsupervised techniques for DDoS attack detection. They proposed Semi-Supervised Machine learning technique is presented which is the combination of both supervised and unsupervised techniques. Researcher in [31] have claimed that neural networks (NN) are a good choice for DDoS detection. To develop the neural network model, the Deduct or modelling environment was employed. A single-layer perceptron for this NN model was comprised of 35 neurons (or nodes) that are (11 input neurons, 23 hidden and only one output node). A contingency table was used to evaluate the accuracy of the developed model. According to researchers in [32], they suggested to classify the incoming request as a DDoS attack and a legitimate request. A hybrid method for selecting features and classifying it is being presented. What is interesting about the work is that it relies on an available thresholding methodology with the technique of classifying, based on varied network traffic situations. This new method using the algorithm combination of Mean Absolute Deviation (MAD) thresholding and random forest (RF) classification algorithm proved to be most effective. Azizi and Hosseini in [33] have suggested a hybrid framework for DDoS detection. Processes are classified into two groups based on the outcomes. Because each group completed its own work, the speed with which work can be organized is increased as a result of this technique. Random forest appears to produce better results in both datasets under consideration (the NSL-KDD dataset and other modern dataset), however, in a particular case, any other of the algorithms may perform superior. The researcher in [15] suggested a network IDS (NIDS) that is capable of detecting a DDoS attack using ensemble classifiers and a reduced feature dataset.

The researchers in [21] addressed the major obstacles hindering the development of IoT intrusion detection systems in 2020. A unique CNN model was suggested, which uses a feature fusion method and a loss function based on cross entropy which utilizes multilayer convolution. Their solution is more advanced than current deep

learning methods, which are mostly focused on normal network intrusion problems. DDoS attacks in cloud computing can be detected and reduced using artificial immune systems (AIS) described by Prathyusha and Kannayaram [34] in 2020. According to authors in [35], a recommended architecture for DDoS classification is the auto encoder (AE) and the deep neural network (DNN) architectures developed in 2020. Initially, a naïve artificial intelligent and DNN model is generated, and hyperparameters values are randomly being used to create the model. An upgraded model is created from the baseline by enhancing it with additional algorithmic improvements. In 2020, Bagyalakshmi and Samundeeswari [36] proposed two approaches which are the filter method represented by Learning Vector Quantization (LVQ) and the dimensionality reduction method defined by Principal Component Analysis (PCA). Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM) are used to classify DDoS attacks, and these algorithms use the selected features out from each method.

3 Methodology

Deep learning is a new part of machine learning, but it has some key differences: DL needs a large amount of data to recognize the data excellently. Also, in DL, the features extracted are automatically [37]. Moreover, DL does not need to break problems down into sub-problems to solve them and gather the end result like ML, so DL directly solves the problem. Furthermore, DL takes a long time to train data in the training phase, but in the testing phase it is very quick. For these reasons, it can be summarized that deep learning has better performance than machine learning, especially with large datasets. Therefore, in the present study, two methods of deep learning were used, CNN and LSTM, and they were combined together to extract a novel method that gives better results. Figure 2 demonstrates the model of the methodology proposed in this work. In the following subsections, the dataset will be introduced as the first step. Secondly, the preprocessing technique will be implemented on the entire suggested dataset. Thirdly, the CNN and LSTM will be introduced individually. Then, the proposed model, which consists of CNN and LSTM, will be explained. Finally, in the last subsection is the learning functions and parameters.

3.1 Dataset

The NSL-KDD dataset was used to test our suggested model. Over time, the KDD'99 dataset has been refined to be more useful for algorithm performance evaluations by removing or reassigning records from classes that were previously duplicated. The NSL-KDD dataset consists of 41 features per record [38]. The NSL-KDD dataset consist of 148514 rows. In this study, the data will be divided into a training and test set. The training set is 80% and becomes 118811 rows, while the test set is 20% that becomes 29703 rows.

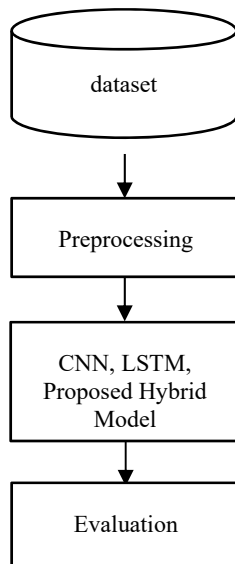


Figure 2

The proposed methodology model

3.2 Preprocessing

Data preprocessing is an important and necessary phase in the machine learning and data mining processes that involves manipulating or removing data before it is utilized for performance improvement. When dealing with a large dataset, preprocessing can be utilized to deal with multiple issues at once. Preprocessing techniques should be utilized to extract redundant data or unwanted data. Therefore, the task of preprocessing is to make the data suitable for processing in the training phase [39]. One of the preprocessing approaches used in this paper is standardizing features, which means eliminating the mean and dividing them all by the standard deviation. That is calculated as.

$$z = (x - \mu) / \delta \quad (1)$$

where μ represent the mean of the training samples. 0 will represent the mean if *with_mean = False*. Also, δ represents the standard deviation of the training samples. But it will be 1 if *with_std = False*. Each feature is separately centered and scaled by calculating the necessary statistics from the training set examples. By using a transform, the mean and standard deviation are stored to be used in the testing set.

3.3 Convolutional Neural Network (CNN)

This type of deep neural network, known as a "convolutional neural network," has been commonly utilized in a variety of fields due to its high performance [40].

CNNs are the most accurate multilayer neural networks; they use the same feedforward and backpropagation as other NNs' algorithms, but their architecture is unique. CNNs have the following architecture: the input layer comes first, followed by the several hidden layers, and finally the output layer [41]. Where the hidden layers are generally comprised of convolutional, pooling like maxpooling, and fully connected layers. Also, convolution process and sampling process are the two basic operations in the CNN algorithm. The convolution process applies filters to the original data or feature map that is created from the original data and then adds bias. The convolution process is conceptualized as a one-dimensional process with a specified input $I(t)$ and a kernel $K(a)$. The process to calculate the convolution may be summarized as follows.

$$s(t) = \sum_a I(t + a) \cdot k(a) \quad (2)$$

The core of the process is that the kernel is a considerably smaller collection of multiple points of data than the data input, but when the input is equal to the kernel, the convolution process output is greater. Moving along the network, using a technique called sampling to lessen their dependency on the precise placement of elements. Max-pooling seems to be the most widely used pooling method, and hence, it is mostly found in this layer. A technique of selecting the biggest element inside small region in the certain pooling region is known as "max-pooling". when the stride is set to two, the max-pooling layer output will be halved [42]. In the present study, CNN was comprised of five layers. Firstly, the data comes from the NSL-KDD dataset and it is preprocessed. This layer is called the input layer. After that comes the convolutional layer, which is one dimension (Conv1D). With the parameters: filter equals 10, kernel_size equals 3, and stride equals 1. Also, the activation function is a Rectified Linear Unit (ReLU) function, which will be explained afterward. The next layer is the max pooling layer, which has one dimension, and the pooling size is equal to 2. Before data was moved to a last layer, the flatten layer flattened it because the pooling size was greater than one. Softmax is the activation function utilized with the last layer (a fully connected layer). The CNN parameters are tabulated in Table 1.

Table 1
CNN parameter setting

<i>Algorithm</i>	<i>Initializer</i>	<i>Activation Function</i>	<i>Optimizer</i>	<i>Epochs</i>
CNN and LSTM	glorot_uniform	Relu, Softmax	Adam	500

In the table above, the term "activation function" refers to $f:R \rightarrow R$ [43]. There are many different activation functions but for these non-linear functions, the non-linear activation functions are necessary. A non-linear activation function with a finite number of possible values was published in the literature in the past. Activation functions such the Rectified Linear Unit ReLU function and Softmax function are often employed, especially because they are the most prevalent. Generally, in the output layer, the softmax function and Cross Entropy loss function are combined and utilized for multi-classification activities. The Softmax layer standardizes

outputs of the preceding layer in order to be one. The preceding layer model's units represent the un-normalized score that the input belongs to a specific class. This layer has normalized by the Softmax, therefore the output value indicates the likelihood of each class [43]. The ReLU function will return 0 as an output if the input is less than 0, while it will return the same input number if the input is higher than 0.

$$\text{softmax}(x) = \frac{e^{x_1}}{\sum_{c=1}^n e^{x_c}} \quad (3)$$

$$\text{ReLU}(x) = \max(x, 0) \quad (4)$$

ReLU functions are mathematically a lot simpler because both forward and backward passes through a ReLU are simple statements. There is an enormous benefit in situations when a network has a large number of neurons because the training and assessment duration may be considerably reduced [43].

3.4 Long Short Term Memory Neural Network (LSTM)

LSTMs are a common kind of recurrent neural network (RNN) built primarily for the purpose of learning long-term reliance. An RNN and an LSTM network are both neural networks with the same structure. There is a major distinction between LSTM and RNN's basic unit since LSTM has a memory block built in. The LSTM memory blocks are called cells that are responsible for remembering things. Also, the cells are controlled by three techniques called gates: the Forget gate, the Input gate, and the Output gate. A forget gate is in charge of erasing unwanted data from the cell state. Where adding information to the cell's state is a responsibility of the input gate. At the same time, extracting valuable info from the current cell state and displaying it as an output, it is managed from the output gate side. A complete overview of LSTM is shown in Figure 3.

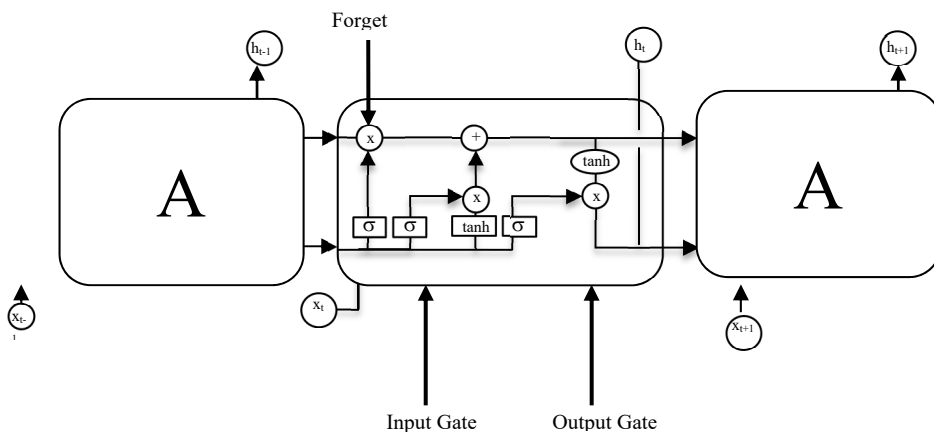


Figure 3
LSTM with its gates [44]

In the current study, LSTM was made up of three layers. Initially, the input layer is the same as CNN, which is the preprocessing layer followed by an LSTM layer. The LSTM layer has an activation function named ReLU of 41 units, and the initializer is `glorot_uniform`. Finally, it is a fully connected layer. Softmax is the activation function used for this layer, as CNN. The LSTM parameters are tabulated in Table 1.

3.5 Proposed Model

Proposed model is a hybrid method that combines CNN and LSTM into a single model that consists of seven layers. The present study combined CNN with LSTM in order to indicate the high quality of detecting DDoS attacks. Figure 4 illustrates the overall architecture of the suggested propose model. The figure includes seven layers. As it is mentioned in the following paragraph:

Initially, the input layer is the same as the first layer in CNN and LSTM, which is the preprocessing data followed by the convolutional layer, which is one dimension (Conv1D). With the parameters: filter equals 10, kernal_size equals 3, stride equals 1, and the activation function is a ReLU function.

The next layer is the max pooling layer, which has one dimension, and the pooling size is equal to 2. The second layer is repeated in the fourth layer, and the third layer is repeated in the fifth layer. Moreover, the next layer, the LSTM layer, has the same activation function as the second and fourth layers. The last layer in the proposed model, like the CNN and LSTM output layers, is a fully connected layer with softmax activation function.

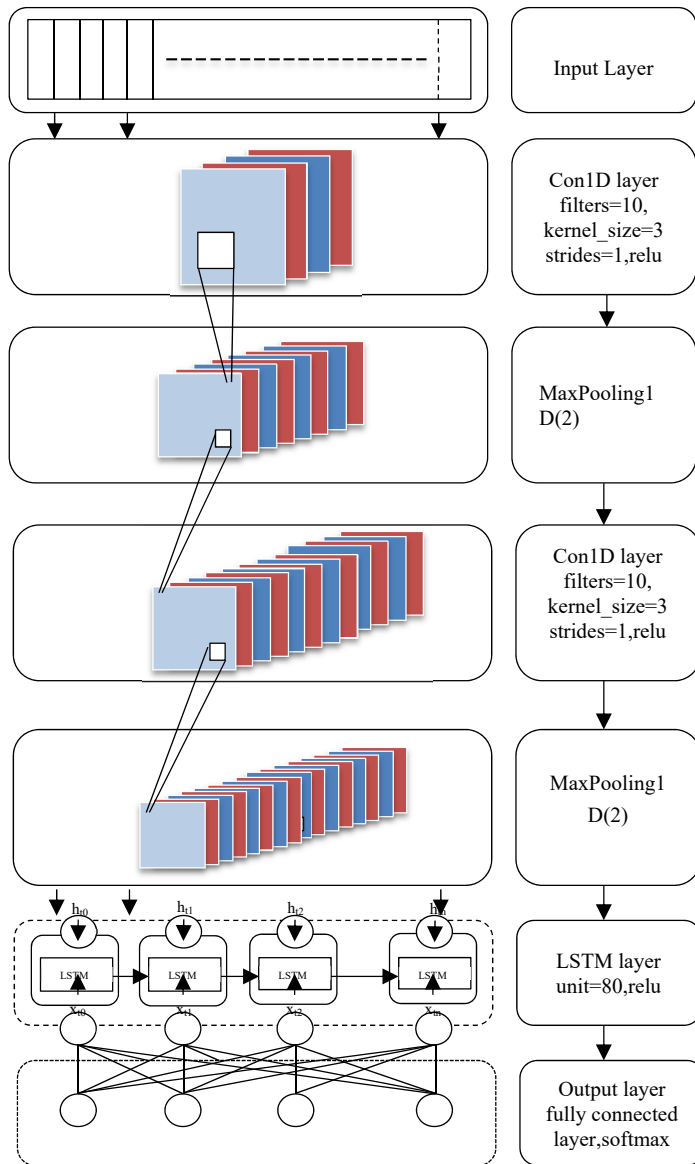


Figure 4

The General Structure of the Proposed Model

3.6 Learning

The `glorot_uniform` initializer was used as the `kernel_initializer` to initialize the weights for all CNN, LSTM, and in the proposed model methods [45]. This `glorot_uniform` function is useful for obtaining samples from a distribution of uniform within the bounds of two limitations. The limitation is the square root of six divided by $(fan-in + fan-out)$. In the same time, the number of weight tensor input units is represented by $fan-in$ and the number output weight tensor output unit is represented by $fan-out$. The weights were updated in the training phase, and in the same phase, the backpropagation technique was used. The Sparse Categorical Cross-entropy is a loss function that is utilized to compute the error, where the error is the difference between the predicted value $f(x_i, \theta)$ and the actual value y :

$$SCCE = -\sum_{i=1}^n y_i \log f(x_i, \theta) + (1 - y_i) \log(1 - f(x_i, \theta)) \quad (5)$$

The error will move backward across the network while the weights wait for themselves to become current. All intermediate nodes between layers are, therefore, linked, and they all will contribute their error values to forward propagation as it passes through them. The propagation mechanisms, both forward and backward, wrapped the entirety of the network [43]. In the current paper, a stochastic gradient descent optimizer known as Adaptive Moment Estimation (ADAM) [46] was employed for weight updating, with a learning rate of 0.0001. Learning rate is an important hyperparameter to minimize loss function because it controls the weight update. The learning rate must be right, not tiny or huge because the tiny learning rate makes the processing in the training phase slow, and at the same time, being too high can cause unwanted divergent action in the loss function. During processing in the training phase, the networks went through 500 epochs of repetition. Where one epoch refers to one pass forward and one pass backward of all the data in the training phase or a comprehensive training cycle of all the data. Also, the size of the batch is equal to 32.

4 Evaluation Criteria

In the present study, the evaluation criterias were applied on NSL-KDD dataset testing. The evaluation of results composed of four criterias, which were Accuracy, Precision, Recall and F1 score. The results of the present study were classified according to normality and abnormality. In each result, there were four expectations, namely: True Positive (TP) is the correct recognition of DDoS attacks; True Negative (TN) is the correct recognition of normal records; False Positive (FP) identified DDoS attacks incorrectly; and False Negative (FN) recognizes normal records incorrectly.

Accuracy: indicates the correct predicts from all predications.

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN} \right) \quad (6)$$

Precision (P): is a measure of a system's ability to distinguish between an assault and what is considered normal [47].

$$Precision = \left(\frac{TP}{TP+FP} \right) \quad (7)$$

Recall or true positive rate: represent the number of predicted DDoS attacks in real DDoS attacks [48].

$$Recall = \left(\frac{TP}{TP+FN} \right) \quad (8)$$

F1 score: The F1 score can be defined as a harmonic average of recall and precision, and the F1 score result is between the worst 0 and the best 1 [49].

$$F1 \text{ score} = \left(\frac{2TP}{2TP+FP+FN} \right) \quad (9)$$

5 Experiment and Results

In the current study, the experiments were formed by Python language. Python is an efficient high-level and object-oriented programming language. A wide range of machine learning, artificial intelligence and computation libraries are available by Python, such as: NumPy, SciPy, Scikit Learn, Keras, Theano and many others [50]. The Keras library which provided by Python, was used to create and train suggested models, and it was executed on TensorFlow's framework. TensorFlow is a free and open-source framework that may be used for high-performance numerical computing. The TensorFlow is a flexible and extensible architecture that makes it possible to run computation easily on many platforms (Tensor Processing Unit, Graphics Processing Unit, Central Processing Unit), on desktops, in data centers, on mobiles, and many other devices.

In the present study, five experiments were conducted for each of the upcoming methods: CNN, LSTM, and in the proposed model to obtain comprehensive results. The mean, median, and standard deviation (SD) of accuracy, precision, recall, and F1 score for each of the aforementioned methods were indicated in order to be able to make a comparison between them, as it is shown in Table 2, Table 3, and Table 4. Table 2 illustrates the suggested CNN's performance for each fold. Shown in the fourth fold the accuracy was the highest at 97.83%.

While the precision rate was the highest in the fifth fold 98.23%. Furthermore, recall was considered as the highest rate in the first fold with the percentage of 97.92%. Moreover, in the fifth fold, F1 score was demonstrated as the highest rate by 98.00%. The mean of accuracy, precision, recall, and F1 score was 1, 2, 3, and 4 respectively. Table 3 represents the suggested LSTM's performance for each

iteration. As it is obvious in the middle table the results of the fourth fold were the highest ones among all of the folds. The accuracy, precision, recall, and F1 score in mentioned fold were 98.97%, 84.19%, 84.39%, and 84.28% respectively. Moreover, the mean of every five iterations of LSTM method for each metric (accuracy, precision, recall, and F1 score) were 97.25%, 79.55%, 78.64%, 78.65% respectively.

Table 2
The suggested CNN's performance for each fold

<i>Fold</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1 score</i>
1	97.67	97.94	97.92	97.92
2	97.80	93.77	83.67	83.72
3	97.74	83.80	83.65	83.72
4	97.83	84.16	83.55	83.85
5	97.75	98.23	97.78	98.00
Mean	97.76	91.58	89.31	89.44
Median	97.75	93.77	83.67	83.85
Standard deviation	0.061	7.160	7.793	7.776

Table 3
The suggested LSTM's performance for each fold

<i>Fold</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1 score</i>
1	97.55	83.12	81.30	82.17
2	98.57	83.87	83.66	83.75
3	98.23	79.19	83.91	81.10
4	98.97	84.19	84.39	84.28
5	92.93	67.38	59.96	61.95
Mean	97.25	79.55	78.64	78.65
Median	98.23	83.12	83.66	82.17
Standard deviation	2.471	7.092	10.513	9.421

Table 4 demonstrates the performance of the proposed model for every five iterations. As it is mentioned the second fold achieved the highest metrics. In the second fold as it is seen, accuracy, precision, recall, and F1 score were 99.31%, 99.18%, 99.18%, 99.18% respectively. Furthermore, the mean of accuracy was 99.20%, while the mean of precision was 91.94%. Also the mean of recall was 93.37%, and the final mean metric was 92.41%. The current study was conducted to indicate that using the hybrid method, which consisted of CNN and LSTM, obtained better results than using them separately.

As it is clear in terms of comparison and Figure 5, proposed model was much improved than others in terms of the four metrics. Also, the mean, max, and min of every metric of proposed model were more elevated than CNN and LSTM methods, but proposed model terms of SD only recall was better than the others.

Table 4
The suggested in the Proposed Model's performance for each fold

<i>Fold</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1 score</i>
1	99.21	92.01	99.10	94.36
2	99.31	99.18	99.18	99.18
3	99.11	99.03	98.99	99.01
4	99.19	84.75	84.78	84.77
5	99.20	84.71	84.79	84.75
Mean	99.20	91.94	93.37	92.41
Median	99.20	92.01	98.99	94.36
Standard deviation	0.071	7.188	7.835	7.250

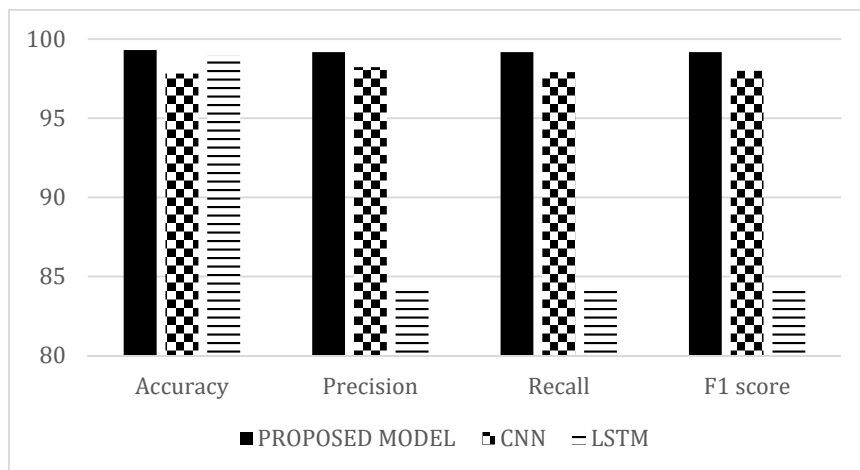


Figure 5

The performance comparison between CNN, LSTM, and the proposed model based on mean

Table 5

The comparison of proposed model with many state-of-the-art approaches in term of accuracy

<i>No</i>	<i>Name</i>	<i>Year</i>	<i>Accuracy (%)</i>	<i>Technique</i>
1	Our proposed model	current	99.20	Proposed Hybrid Model
2	Yusof et al. [24]	2017	91.7	DCF + CSE
3	Kushwah and Ali [25]	2017	96.3	ANN + black hole optimization algorithm
4	Igbe et al. [26]	2017	98.6	DCA
5	Derakhsh et al. [27]	2018	82.44	GA
6	Hoon et al. [28]	2018	93.26	DRF
7	Idhammad et al. [29]	2018	98.23	semi-supervised

8	Anjum and Shreedhara [30]	2019	93.26	semi-supervised
9	Mukhametzhanov et al. [31]	2019	97.94	NN
10	Verma et al. [32]	2019	98.23	MAD+RF
11	Hosseini and Azizi [33]	2019	98.9	hybrid technique
12	Das et al. [15]	2019	99.1	Ensemble technique
13	Ma et al. [21]	2020	92.99	CNN
14	P.-K.-Y.[34]	2020	96.7	AIS
15	Bhardwaj et al. [35]	2020	98.43	AE+DNN
16	B. and S. [36]	2020	98.74	LVQ+DT

Table 5 demonstrates the comparison of proposed model with many state-of-the-art approaches in terms of accuracy. As shown in the table, there are no hybrid techniques of two deep learning algorithms in the previous work on the NSL-KDD dataset but there are many good techniques such as: ensemble technique, hybrid technique, semi-supervised technique and others. By comparing the present study with them, the present study achieved the highest result and the accuracy rate was 99.20%.

From the results of the experiments, it is seen that the hybridization of two deep learning technologies, CNN and LSTM, leads to excellent results in detecting DDoS attacks depending on their architecture. In addition to that, the functions and parameters used in the learning have a magical effect to make the proposed model more accurate. This hybridization that relies on CNN as a feature extractor and LSTM as a predictor has a better accuracy when compared to each one individually. Moreover, from the comparison of proposed model and previous work of the same dataset, the NSL-KDD dataset, it is found that the current method has the best accuracy in detecting DDoS attacks. It has become apparent for the researcher that the usage of proposed model was greater than the usage of DL, and traditional ML algorithms.

Conclusion

The results obtained in the present study indicated that the proposed model has higher performance than CNN and LSTM in terms of accuracy, precision, recall, and F1 score. Also, the mean of the four metrics' accuracy, precision, recall, and F1 score rate are 99.20%, 91.94%, 93.37%, and 92.41%, respectively. Moreover, the DDoS detection in the NSL-KDD dataset achieved the highest accuracy among other previous studies. The findings of the current study indicate that the proposed model is better than using CNN and LSTM separately on this dataset. The present study can contribute to making DDoS attack detection more accurate. For future work, the present study suggests that proposed model be implemented in various sectors, not only for attack detection. Furthermore, we propose improving the architecture used from serial to parallel and introducing voting technology to it.

References

- [1] Bharot, N. et al.: *DDoS Attack Detection and Clustering of Attacked and Non-attacked VMs Using SOM in Cloud Network*. In: International Conference on Advances in Computing and Data Sciences. Springer, 2019, pp. 369-378
- [2] Baykara, M., Das, R.: A Novel Hybrid Approach for Detection of Web-Based Attacks in Intrusion Detection Systems. *International Journal of Computer Networks and Applications*, 4(2), 2017, pp. 62-76
- [3] ISSA, Ahmed Sardar Ahmed, and Zafer ALBAYRAK. "CLSTMNet: A Deep Learning Model for Intrusion Detection." *Journal of Physics: Conference Series*. Vol. 1973, No. 1, IOP Publishing, 2021
- [4] Özalp, A. N et al.: Layer-based examination of cyber-attacks in IoT. *International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2022, pp. 1-10
- [5] Hyder, H. K., Lung, C. H.: Closed-Loop DDoS Mitigation System in Software Defined Networks. *DSC 2018 - 2018 IEEE Conf. Dependable Secur. Comput.*, 2019, pp. 1-6
- [6] Musotto, R., Wall, D. S.: More Amazon than Mafia: analysing a DDoS stresser service as organised cybercrime. *Trends Organ. Crime*, 2020
- [7] Chen, W. et al.: *Intrusion Detection for Modern DDoS Attacks Classification Based on Convolutional Neural Networks*. In: Studies in Computational Intelligence. Springer, Cham, 2021, pp. 45-60
- [8] Alabadi, Montdher, and Yuksel Celik. "Anomaly detection for cyber-security based on convolution neural network: A survey." *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, 2020
- [9] Lima Filho, F. S. De et al.: Smart Detection: An Online Approach for DoS/DDoS Attack Detection Using Machine Learning. *Secur. Commun. Networks*, 2019
- [10] Atasever, S. et al.: Siber Terör ve DDoS. *Süleyman Demirel University Journal of Natural and Applied Sciences*, 23(1), 2019, pp. 238-244
- [11] Tuan, T. A. et al.: Performance evaluation of Botnet DDoS attack detection using machine learning. *Evol. Intell.*, 13 (2), 2020, pp. 283-294
- [12] Beitollahi, H. et al.: ConnectionScore: A Statistical Technique to Resist Application-layer DDoS Attacks. *J. Ambient Intell. Humaniz. Comput.*, 5 (3), 2014, pp. 425-442
- [13] Ajeetha, G., Madhu Priya, G.: Machine Learning Based DDoS Attack Detection. *2019 Innov. Power Adv. Comput. Technol. i-PACT 2019*, 1, 2019, pp. 1-5

- [14] Yusof, M. A. M. et al.: Detection and Defense Algorithms of Different Types of DDoS Attacks Using Machine Learning. *Lect. Notes Electr. Eng.*, 488, 2018, pp. 370-379
- [15] Das, S. et al.: DDoS Intrusion Detection Through Machine Learning Ensemble. *Proc. - Companion 19th IEEE Int. Conf. Softw. Qual. Reliab. Secur. QRS-C 2019*, 2019, pp. 471-477
- [16] Naveen Bindra, Manu Sood: Detecting DDoS Attacks Using Machine Learning Techniques and Contemporary Intrusion Detection Dataset. *Autom. Control Comput. Sci.*, 53 (5), 2019, pp. 419-428
- [17] Otoum, Y. et al.: DL-IDS: a deep learning-based intrusion detection framework for securing IoT. *Trans. Emerg. Telecommun. Technol.*, (September 2020), 2019
- [18] Obaid, K. B. et al.: Deep Learning Models Based on Image Classification: A Review. *Int. J. Sci. Bus.*, 4 (11), 2020, pp. 75-81
- [19] Aytaç, T. et al.: Detection DDOS Attacks Using Machine Learning Methods. *Electrica*, 20(2), 2020, pp. 159-167
- [20] Yuan, X. et al.: Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Trans. neural networks Learn. Syst.*, 30 (9), 2019, pp. 2805-2824
- [21] Ma, L. et al.: A Deep Learning-Based DDoS Detection Framework for Internet of Things. *IEEE Int. Conf. Commun.*, 2020-June, 2020
- [22] Tasdelen, A., Sen, B.: A hybrid CNN-LSTM model for pre-miRNA classification. *Sci. Rep.*, 11 (1), 2021, pp. 1-9
- [23] Donahue, J. et al.: *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 2625-2634
- [24] Yusof, A. R. A. et al.: Adaptive feature selection for denial of services (DoS) attack. *2017 IEEE Conf. Appl. Inf. Netw. Secur. AINS 2017*, 2018-Janua, 2017, pp. 81-84
- [25] Kushwah, G. S., Ali, S. T.: Detecting DDoS attacks in cloud computing using ANN and black hole optimization. *2nd Int. Conf. Telecommun. Networks, TEL-NET 2017*, 2017, pp. 1-5
- [26] Igbe, O. et al.: Denial of service attack detection using dendritic cell algorithm. *2017 IEEE 8th Annu. Ubiquitous Comput. Electron. Mob. Commun. Conf. UEMCON 2017*, 2018-Janua (October), 2017, pp. 294-299
- [27] Derakhsh, A. M. et al.: Using Genetic Algorithm to Improve Bernoulli Naïve Bayes Algorithm in Order to Detect DDoS Attacks in Cloud Computing Platform. *Int. J. Sci. Eng. Investig.*, 7 (March), 2018

- [28] Hoon, K. S. et al.: Critical review of machine learning approaches to apply big data analytics in DDoS forensics. *2018 Int. Conf. Comput. Commun. Informatics, ICCCI 2018*, (1), 2018, pp. 2-6
- [29] Idhammad, M. et al.: Semi-supervised machine learning approach for DDoS detection. *Appl. Intell.*, 48 (10), 2018, pp. 3193-3208
- [30] Anjum, M., Shreedhara, K. S.: Performance Analysis of Semi-Supervised Machine Learning Approach for DDoS Detection. *Int. J. Innov. Res. Technol.*, 6 (2), 2019, pp. 144-147
- [31] Mukhametzyanov, F. et al.: The neural network model of DDoS attacks identification for information management. *Int. J. Supply Chain Manag.*, 8 (5), 2019, pp. 214-218
- [32] Verma, P. et al.: An Adaptive Threshold-Based Attribute Selection to Classify Requests Under DDoS Attack in Cloud-Based Systems. *Arab. J. Sci. Eng.*, 45 (4), 2019, pp. 2813-2834
- [33] Hosseini, S., Azizi, M.: The hybrid technique for DDoS detection with supervised learning algorithms. *Comput. Networks*, 158, 2019, pp. 35-45
- [34] Prathyusha, D. J., Kannayaram, G.: A cognitive mechanism for mitigating DDoS attacks using the artificial immune system in a cloud environment. *Evol. Intell.*, (0123456789), 2020, pp. 1-12
- [35] Bhardwaj, A. et al.: Hyperband tuned deep neural network with well posed stacked sparse autoencoder for detection of ddos attacks in cloud. *IEEE Access*, 8, 2020, pp. 181916-181929
- [36] Bagyalakshmi, C., Samundeeswari, E. S.: DDoS attack classification on cloud environment using machine learning techniques with different feature selection methods. *Int. J. Adv. Trends Comput. Sci. Eng.*, 9 (5), 2020, pp. 7301-7308
- [37] Barik, K. et al.: Applied Artificial Intelligence Cybersecurity Deep: Approaches, Attacks Dataset, and Comparative Study. *Applied Artificial Intelligence*, 36(1), 2022, pp. 1-24
- [38] Nandi, S. et al.: Detection of DDoS Attack and Classification Using a Hybrid Approach. *ISEA-ISAP 2020 - Proc. 3rd ISEA Int. Conf. Secur. Priv. 2020*, 2020, pp. 41-47
- [39] Zulkepli, F. S. et al.: Data pre-processing techniques for publication performance analysis. *Lect. Notes Data Eng. Commun. Technol.*, 5, 2018, pp. 59-65
- [40] Kesenek, Y., Özçelik, İ., & Kaya, E. (2022) A new document classification algorithm against malicious data leakage attacks. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 37(3), 2022, pp. 1639-1654

-
- [41] Özyurt, F.: UC-Merced Image Classification with CNN Feature Reduction Using Wavelet Entropy Optimized with Genetic Algorithm. *International Information and Engineering Technology Association*, 37(3), 2020, pp. 347-353
- [42] Zeng, H. et al.: Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics*, 32 (12), 2016, pp. i121–i127
- [43] Ketkar, N., Santana, E.: *Deep Learning with Python*. Springer, 2017
- [44] Gudikandula, P.: *Recurrent Neural Networks and LSTM explained | by purnasai gudikandula | Medium*. no date
- [45] Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. *Proc. Thirteen. Int. Conf. Artif. Intell. Stat.*, 2010, pp. 249-256
- [46] Kingma, D. P., Ba, J. L.: Adam: A method for stochastic optimization. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, 2015, pp. 1-15
- [47] Bhuyan, M. H. et al.: Network Anomaly Detection: Methods, Systems and Tools. *IEEE Commun. Surv. Tutorials*, 16 (1), 2014
- [48] Macías, S. G. et al.: ORACLE: Collaboration of Data and Control Planes to Detect DDoS Attacks. 2020
- [49] De, V. et al.: Detection of reduction-of-quality DDoS attacks using Fuzzy Logic and machine learning algorithms. *Comput. Networks*, 186, 2021
- [50] Pedregosa, F. et al.: Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12, 2011, pp. 2825-2830

Monitoring and Control of Energy Consumption Systems, using Neural Networks

**Olga Shvets¹, Marta Seebauer², Assel Naizabayeva¹,
Alibi Toleugazin¹**

¹D.Serikbayev East-Kazakhstan Technical University,
Serikbayev st., 19, 070003 Ust-Kamenogorsk, Kazakhstan,
oshvets@ektu.kz, anaizabaeva@ektu.kz, toleugazin.a@edu.ektu.kz

²Óbuda University, Bécsi út 96/B, 1034 Budapest, Hungary
seebauer.marta@amk.uni-obuda.hu

Abstract: Industries, cities, towns and households, around the world, need reliable, affordable and sustainable energy to meet their electricity demand. Renewable energy can make a significant contribution to the development of this area and satisfy this need of the population, both in private households and in the field of industry, transport and supply of entire settlements. This study examines the design of an intelligent energy management system for a residential building. The smart home energy management system must use new infrastructure based on modern technologies such as DSE (Deep Sea Electronics) controller, smart devices, advanced communications, electrothermal models of critical components and advanced optimization models. The main advantage of this energy management system is that it will allow real-time control and monitoring of a home that includes all the components connected to it (for example, a distribution transformer and household appliances). The control system should work without changing the customer's lifestyle. The article discusses topical issues of energy saving in accordance with the development program of the Republic of Kazakhstan until 2050, analyzes the trends in energy saving policies, in different countries. It is developed C# software, for monitoring and control.

Keywords: smart house; energy saving; neuronet; fuzzy logic; energy management

1 Modern Approaches to Energy Saving Policy: Trends in the World and in the Republic of Kazakhstan

1.1 Trends in Energy Saving Policy

The global response to energy security challenges is essentially a growth model that is based on the principles of energy efficiency and environmental sustainability.

Over the past two decades, the main focus has been on integrating energy efficiency with environmental policies, especially in relation to global climate change. Almost all national and regional energy efficiency strategies are directly linked to climate change policies. The global potential for energy savings is enormous. According to the International Energy Agency [1], successful implementation of energy efficiency measures would reduce greenhouse gas emissions by 80%, while significantly improving the security of supply. The International Energy Agency estimates that only improving the energy efficiency of electrical appliances through the use of the best available technologies, as part of a policy aimed at reducing the end-user costs of using electrical appliances, will save up to 1000 TWh by 2030, as compared to the current situation. The production of cars with lower fuel consumption will sharply reduce the demand for fuel resources. In rapidly growing developing economies, the transport sector is projected to account for 43% of energy demand by 2025, up from nearly 35% in 2008.

China, India, Brazil and other countries, where over the past two decades there has been a rapid economic growth and demand for energy, in the face of rising prices for hydrocarbons, are also beginning to switch to energy conservation policies. One of the most important recent trends is the improvement of energy-saving and energy-efficient technologies in construction. The potential for energy savings is high – the IEA estimates that buildings and appliances could account for one quarter of the potential CO₂ emissions reductions up to 2050 [1-4]. Energy saving in the transport sector is also a priority area. Increasing the share of new and renewable energy sources in developed countries is also integrated into energy efficiency policies. The ongoing development of new technologies makes the development of renewable energy sources such as solar energy, hydropower and biomass more affordable and efficient. The main limiting condition is the economic factor – as long as they are still expensive. However, continuous scientific and technical progress in the use of new and renewable energy sources (NRES) and the constant rise in the cost of traditional energy resources, primarily 6 liquid hydrocarbons, expand the scope of NRES mainly in areas without centralized energy supply.

There are very clear differences in approaches to energy saving in different countries associated with the peculiarities of the national mentality, cultural preferences and prevailing stereotypes of behavior. However, an important common feature of developed countries is the concentration of policy on achieving energy savings at the stage of energy use.

Let's consider approaches to energy saving in different countries.

1.2 United States of America

The US economy is 2.5 times more energy efficient than the Kazakhstani economy. According to some experts, 9 times less energy is spent on industrial production in America than in Kazakhstan. At present, the level of energy consumed in the country for the production of goods and services in the amount of \$1, has decreased by more than 50% compared to 1970. The American achievements in energy efficiency are the result of years of energy conservation efforts. A feature of the US energy efficiency policy is the very widespread use of various measures of financial incentives and the evasion of the adoption of all kinds of codes and regulations.

As part of the Vision 2025 initiative, more than a half of all states have adopted their own energy efficiency programs and have established building codes that require new buildings to be energy efficient.

The main goals and directions for improving energy efficiency in the United States include the following key points:

- 1) Reduce US dependence on oil imports
- 2) Develop and introduce energy-saving technologies for public buildings, residential buildings, transport, energy, and industry.

The Energy Efficiency and Renewable Energy Authority has been established within the US Department of Energy with the following key objectives to support these goals [2] [3]:

- Strengthening the energy security of the United States
- Improving the quality of the environment
- Ensuring the economic viability of public-private partnerships, whose activities are aimed at increasing the efficiency and productivity of labor
- Introduction of environmentally friendly, reliable and affordable 12 energy technologies; introduction of alternative energy sources into everyday life, ensuring a higher quality of life.

There are federal programs in the United States for promotion energy conservation and ways to improve energy efficiency.

1.3 Japan

By setting a 30% increase in energy efficiency by 2030 over 2006, the Japanese government is committed to ensuring a modern energy supply / demand structure in a market with the high prices expected by the government in the medium to long term. Japan has pledged to provide funding in the amount of 1.6 trillion. yen to create a so-called "low-carbon society" – a society with low CO₂ emissions, including 3770 billion yen to replace old cars with new, more fuel efficient cars and 295 billion yen to help purchase energy-efficient household appliances. The stimulus package in Japan also includes the allocation of resources to subsidize businesses that introduce energy-efficient hardware and equipment, and to improve small and medium-sized enterprises by conducting energy diagnostics and investing in innovative energy-saving technologies [2].

Approaches to energy saving in Japan implies the introduction of 3 fundamentals into various spheres of society: solar energy, electric cars, energy-saving household appliances. The specific goal is to double the share of renewable sources in energy consumption and achieve the highest indicator in the world – 20%.

Germany, the United Kingdom and the United States are also implementing eco-driving programs based on the experience of Japan.

1.4 The European Union

The European Union is a major driving force in promoting energy efficiency strategies and combating global climate change, and its regulatory impact, extends far beyond its member states. Not all EU Member States give the same attention to energy efficiency, but there is now a requirement for some basic policy. A number of countries far exceed this minimum [2-4].

A number of countries have integrated renewable energy and energy efficiency policies, where this combination is often referred to as a sustainable energy strategy. Such measures have been in place for a long time, and the resulting benefits are undeniable. Since the 1990s, the EU's energy efficiency policy has been closely linked to tackling climate change and has also integrated many aspects of renewable energy and the improvement of technologies for the use of all fossil fuels.

Germany does not have a specific general energy conservation law, but there is a Federal Cogeneration Act and an Energy Saving Ordinance (to introduce a low energy housing standard). Much of the legal framework is based on the transposition of the EU Energy Efficiency Directives into national legislation. An important feature of the organization of energy saving in the country is the preferential financing of energy saving measures by banks and large corporations, and not by the state.

The energy saving management system provides for the delegation of basic functions to the regional and local levels. Energy efficiency issues are closely linked to climate change mitigation activities. When purchasing computers and other electronic devices, the administrative institutions of Berlin should opt for the products that consume the least amount of electricity. Germany is one of the recognized world leaders in energy efficiency in buildings.

Germany and the UK are leading the way in implementing building certification. Only in Germany there are energy efficiency requirements that ensure the optimal level of minimum costs over a 30-year life cycle of buildings.

There is no general energy efficiency law in the UK. Much of the legal framework is based on the transposition of the EU Energy Efficiency Directives into national legislation.

According to the National Energy Efficiency Action Plan, the state policy priority is to consistently promote energy efficiency in business, the public sector and in households. Achievement of targets for reduction of carbon dioxide emissions according to the plan to reduce emissions of carbon dioxide 1980-2050 suggests that total energy consumption in 2050 should not exceed 2011 levels.

The main goals of the UK in the field of improving energy efficiency and the transition to a "low carbon" economic model: development of a distributed power generation system, including "low carbon" heat generation; more active development of communal systems, including combined heat and power generation systems; active participation in the European carbon trading system; increasing the share of using renewable energy sources; support and development of alternative fuels for transport [2] [4].

A system of national, regional and local funds and agencies to support energy efficiency has been developed.

1.5 Kazakhstan

With the adoption by Kazakhstan of the "Strategy "Kazakhstan 2050" and the Concept of transition to a "green" economy, the country has chosen a fundamentally new way of development of society. According to the Concept, the key role will be played by the focus of state policy on reducing environmental impact, resource conservation and achieving a high level of quality of life of the population. Energy efficiency is one of the central points in a gradual transition to a green economy. At present, in terms of the energy intensity of GDP, Kazakhstan is among the countries with the highest values. According to the experts of the Charter, significant opportunities for improving energy efficiency in industry, energy, housing and communal services and transport are concentrated in Kazakhstan [5] [6].

Energy accounts for about 47% of the total consumption of primary energy resources. At the same time, in the energy sector, there is a high proportion of wear and tear of generating and power grid equipment, which, as a result, leads to low efficiency of power generation and a relatively high amount of losses in power grids. In the industrial sector, a high level of energy consumption is primarily due to the activities of such energy-intensive sectors of the economy as oil and gas, metallurgy and mining. At the same time, the technical condition of the equipment and the problem of reducing the workload of enterprises significantly affect the efficiency of the industry. A number of legislative restrictions adopted in terms of energy consumption in industry have not yet yielded positive results. An analysis of the approved norms of energy consumption in industry showed their inapplicability to the working conditions of some enterprises, especially in terms of the mining and metallurgical complex and coal mining. In terms of housing and communal services, most of the existing housing stock consists of apartment buildings with central heating based on boiler houses or CHP plants. With the current state of infrastructure, district heating networks are characterized by low efficiency and significant heat losses. On average, residential buildings in Kazakhstan consume three times more energy per unit area than in the Nordic countries. The high level of heat loss is mainly associated with outdated equipment, as well as the lack of proper repair. The transport sector accounts for up to 17% of the total consumption of the country's primary energy resources, while the technical condition of a part of the vehicle fleet and the quality of the fuel used have a significant impact on specific fuel consumption and emissions of harmful substances. The transition to new fuel quality standards, the introduction of modern navigation and information systems will improve the energy efficiency of the transport sector and increase the throughput of the transport system.

2 Review and Simulation

2.1 Literature Review

As smart homes have become a very active and well-established research topic, many publications on this topic can be found. This field is developing rapidly and is attracting synergy of several areas of science to improve the quality of life for people.

Richard Harper, who has researched the field of smart technology for private homes, wrote that the way a home is built or environmental considerations will not make it a smart home. But “what makes it smart, is the interactive technology it contains” that can help realize “the dream of a home that can actively help its inhabitants” [7-9].

Research on smart homes has mainly focused on hardware solutions for a long time. Currently, the term mainly refers to the integration of information technology into residential buildings. Safety, healthcare, energy efficiency and improving the comfort of residents are the main research topics in the field of smart homes. Interconnection between devices and advanced control of lighting, entertainment and multimedia devices to improve comfort is proposed in several publications [7] [10-13]. In addition, in the field of security, remote information and intervention systems have been developed to enhance control inside the house when the resident is absent. Another application of security systems mentioned in the literature is presence modeling [14] [15]. However, there is still room for more research in other areas.

Many approaches to energy conservation in buildings using smart technologies can be found in the literature.

Most of the approaches to energy conservation in smart buildings described in the literature aim to reduce the energy consumption of heating, ventilation and air conditioning (HVAC) devices such as a home heating system [16], air conditioning [17], or both [18] [19]. Others do not directly address the reduction in consumption of such devices, but provide improved monitoring and control [8]. Most of these projects use a wide variety of sensors to measure humidity and temperature and process data with a fuzzy controller [20] [21] for power distribution. Others, who also sought to “minimize household energy waste,” identified two more areas for incorporating energy management functions: “lighting and appliances” [22] [23]. When simulating various scenarios using synthetic data, they found that the potential for energy savings in private homes is nearly 30 percent. They examined the presence sensors to turn on the lighting devices, that detected the person using infrared heat detection.

The new approach suggests that the controlled home is equipped with some energy efficient devices such as rooftop solar PV panels, energy storage, and controlled / uncontrolled appliances. There are many benefits to implementing this new approach. The main one is that it allows real-time control and monitoring of all the various components connected to a customer/utility-owned home. Thus, the burden of integrating new loads is reduced and further increases the integration of renewable energy sources into the distribution system. In addition, it allows you to implement various home optimization functions to maximize the benefits for both utilities and consumers. All this without any inconvenience to the end user and without overloading/overheating the distribution transformer.

Any system is influenced by external factors. The energy management system includes the use of alternative energy, for example, as in this work – the operation of solar panels. The efficiency of solar panels is directly dependent on the direction of the sun's rays. For best efficiency, the sun's rays should be directed perpendicular to the surface of the module. The illumination of the surface of solar panels with this arrangement will tend to maximum. The system for controlling

the maximum illumination during the day must periodically change the position of the solar panels to maintain a right angle between the direction of the rays and its plane, i.e. ensure that solar panels rotate during the day to maximize the flow of solar radiation.

2.2 Data Processing

As a basis for calculating the power generated by the solar panel, we take the monthly electricity consumption. In order to calculate the required power for our solar panel, we need to know the monthly electricity consumption. We can determine the required amount of electricity consumed in kilowatts per hour, by looking at an electric meter (Figure 1).

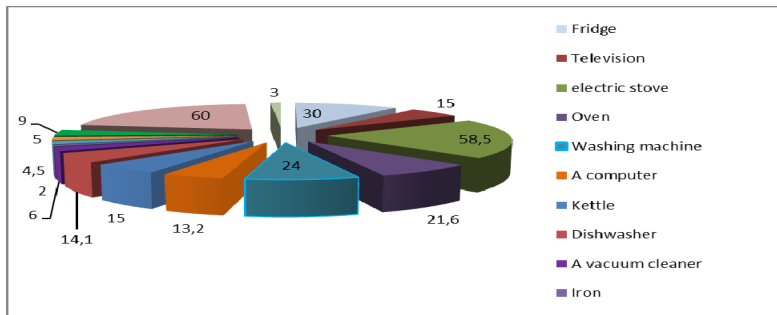


Figure 1

Electricity consumption by various devices

If the costs are, for example, 281 kWh, then the solar battery should generate about 10 kWh of electricity per day. Based on this, it can be calculated that to obtain 10 kWh of energy under ideal conditions, an array of panels with a capacity of at least 1 kW, the number of 15 panels, will be required. In the calculations, it should be borne in mind that solar panels generate electricity only during daylight hours, and their performance depends on both the angle of the sun above the horizon and weather conditions. On average, up to 70% of the total amount of energy is generated from 9am to 4pm, and in the presence of even slight cloudiness or haze, the power of the panels decreases 2-3 times. If the sky is covered with continuous clouds, then at best we can get 5-7% of the maximum capacity of the heliosystem.

After calculating how much energy the solar panel produces in one day, the annual output of the solar panel can be determined (Figure 2) [21].

For example, consider the average daily insolation by months from one of the meteorological servers for Ust-Kamenogorsk. The data are indicated taking into account atmospheric phenomena and are averaged over several years.

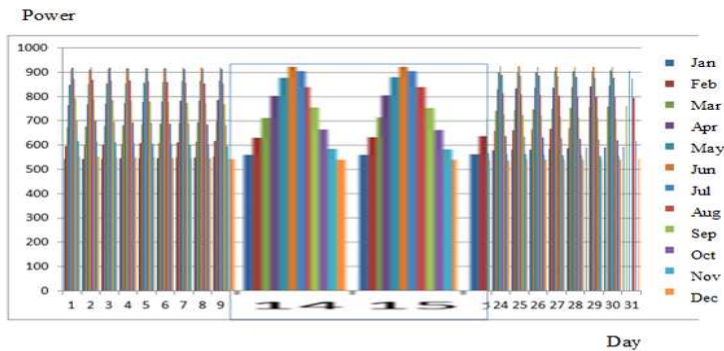


Figure 2
Annual power of the solar panel (2019)

The unit of measurement of insolation in the table is $\text{kWh} / \text{m}^2 / \text{day}$ (kilowatt-hours per square meter per day).

The angle of inclination of the plane, degrees in relation to the ground (0 degrees – insolation on the horizontal plane and 90 degrees – insolation on the vertical plane, etc.), with the plane oriented to the South.

As we can see, the most unfavorable month for this region is December, the daily average insolation on the horizontal surface of the earth is $0.5 \text{ kWh} / \text{m}^2 / \text{day}$ and on the vertical – $1.22 \text{ kWh} / \text{m}^2 / \text{day}$. With an angle of inclination of the plane relative to the ground of 70 degrees, the insolation will be $1.26 \text{ kWh} / \text{m}^2 / \text{day}$, the optimal angle for December is 74 degrees. The most favorable month is June and the insolation on the horizontal surface will be $5.27 \text{ kWh} / \text{m}^2 / \text{day}$, the optimal tilt angle for June is 11 degrees [21].

The angle of inclination of the solar panel, with year-round use in a system that consumes on average the same power regardless of the season, must coincide with the optimal angle of inclination of the most unfavorable month in terms of the amount of solar radiation.

The optimal tilt angle for December in Ust-Kamenogorsk is 74 degrees, so it is worth installing a solar panel, since in other months the insolation is noticeably greater, and as a result, the generation of electricity will be more than enough. Moreover, in winter at tilt angles of 70-90 degrees, precipitation in the form of snow will not accumulate on the solar panel. If the task is to obtain maximum power from solar panels throughout the year, then it is required to constantly orient the solar panel as perpendicular to the sun as possible. Average daily insolation by months is presented in Table 1.

Table 1
Average daily insolation by months

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Average annual insolation kW * h / m ² / day
0°	0.75	1.56	2.81	3.87	5.13	5.27	5.14	4.30	2.63	1.49	0.81	0.50	2.86
40°	1.51	2.55	3.78	4.34	5.12	4.97	5.00	4.57	3.22	2.20	1.46	1.08	3.32
55°	1.66	2.70	3.82	4.16	4.70	4.51	4.53	4.31	3.17	2.27	1.58	1.20	3.22
70°	1.72	2.71	3.67	3.79	4.18	3.95	4.00	3.85	2.97	2.24	1.62	1.26	3.00
90°	1.65	2.50	3.19	3.07	3.21	2.99	3.05	3.08	2.51	2.02	1.53	1.22	2.50
Optimal angle	72.0	63.0	50.0	34.0	20.0	11.0	16.0	27.0	43.0	58.0	69.0	74.0	44.6

The choice of a suitable neural network for modeling depends on the specific task, as well as on the type of data and their volume. There are many classifications of networks, but for solving problems typical for the electricity market, it is best to use a multilayer perceptron (the problem of predicting energy consumption) and Kohonen networks (the problem of constructing a client profile of electricity consumption).

In the real world, there are many parameters that affect power consumption and determine the dimension of the vector of input signals X , and not all of them have the same effect on power consumption. For example, it can be assumed that the electrical load in the forecast period depends on the following parameters (predictors): load in the last week; day of the week; number of working days; the duration of daylight hours; air temperature; cloudy; the end of the month; customer equipment maintenance schedule; duration of the heating period; client type; branch of the economy. How can we single out the most significant among the many parameters?

Most of the significant parameters for forecasting consumption relate to the so-called cyclical parameters: daily, weekly dependencies; monthly, quarterly, annual; weekends / working days, etc. And another significant group of parameters is determined by functional characteristics: meteorological conditions; client type; branch of the economy; characteristics of premises, etc. In addition, today it is customary to single out the factors of the market environment

(www.gkhprofi.ru/news.php?id=69) that affect consumption: volumes and prices of the “day ahead market”; volumes and prices of the “balancing market”; market supply and demand, etc.

2.3 Simulation

Let's consider a simplified model for solving problems typical for the electric power industry.

Description of the model. A multilayer perceptron is a network of several layers of neurons connected in series. At the lowest level of the hierarchy is the input layer of sensory elements x_1, \dots, x_L , whose task is only to receive and distribute the input information over the network. Further, there is one or (less often) several hidden layers (Figure 3).

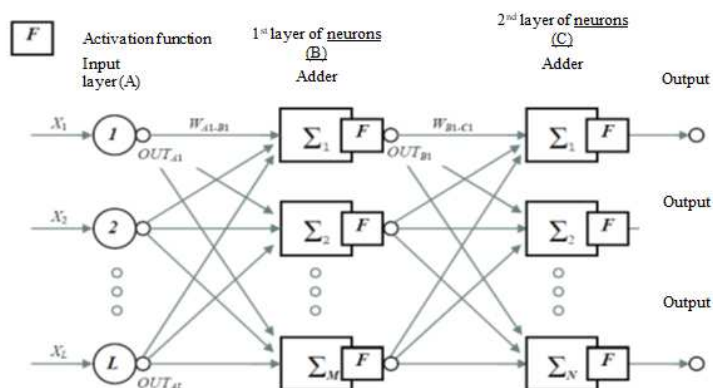


Figure 3

The structure of the neural network

Each neuron on the hidden layer has several inputs connected to the outputs of the neurons of the previous layer or directly to the input sensors x_1, \dots, x_L , and one output. The outputs of the neurons of the last, output layer describe the result processed by the network. The neurons of each layer are not connected with each other and only interact with the neurons of the previous and subsequent layers. Each neuron sums up the signals coming to it from the neurons of the previous hierarchy level with weights, and then, using the activation function, transforms the summation result. The activation function $F(\xi)$ provides the nonlinearity required for the convergence of the learning process. The neuron output signal is given by the expression:

$$OUT = F(\xi) = \frac{1}{1 + \exp(-\xi)} \quad (1)$$

where $\xi = \sum_{k=1}^L x_k w_k$; x_k — neuron inputs; w_k — synaptic weights of inputs; L — is the number of neuron inputs. The sigmoid is selected as the activation function. The structure of the neural network used in this work is shown in Figure 3.

The following designations are used: - each layer of the neural network has its own letter (for example, the letter A corresponds to the input layer, the output to C; - the neurons of each layer are numbered with Arabic numerals; - W_{A1-B1} - synaptic weight between neurons A1 and B1; - OUT_{A1} - output neuron A1. Before using a neural network to obtain a forecast, it must be trained, that is, to determine the synaptic weights. The deterministic method of back propagation of the error is often used to train perceptrons [15] [20]. This method assumes a priori knowledge of the set of required outputs of the neurons of the last layer networks, called target, for a given set of inputs of the initial (zero) layer. For brevity, these sets of inputs and outputs will be called vectors. During training, it is assumed that for each input vector there is a target vector that parallels it, specifying the required output. These vectors are called a training pair. The network learns on many pairs. The optical weights are initialized to random numbers ranging from 0 to 0.1. The learning process consists in calculating the output vector of the network and correcting the weight matrices for each training pair at each iteration according to the formulas below [6] [19]. The formula for correcting the weights for the output layer is $w_{p-k}(i+1) = w_{p-k}(i) + \eta \delta_k OUT_p$, where i is the number of the current training iteration;

$$\delta_k = OUT_k(1 - OUT_k)(T_k - OUT_k) \quad (2)$$

- w_{p-k} : is the value of the synaptic weight connecting the neuron p of the hidden layer with the neuron k of the output layer
- η : coefficient of "learning rate", which allows you to control the average value of the change in the weights
- OUT_p : output of neuron p of the hidden layer
- T_k : is the target value of the output of the neuron k of the output layer
- OUT_k : output of neuron k of the output layer

The formula for correcting the weights for the hidden layer is written as:

$w_{p-q}(i+1) = w_{p-q}(i) + \eta \delta_q OUT_p$, where i is the number of the current training iteration;

$$\delta_q = OUT_q(1 - OUT_q)(T_q - OUT_q) \quad (3)$$

$$\delta_q = OUT_q(1 - OUT_q) \sum_{k=1}^N \delta_k w_{q-k}; w_{p-q} \quad (4)$$

w_{p-q} — is the value of the synaptic weight connecting neuron p of the previous (in this case, input) layer with neuron q of the hidden layer

η — is the coefficient of "learning rate", which allows to control the average value of the change in the weights

OUT_p — output of neuron p of the previous (in this case, input) layer

OUT_q — output of neuron q of the hidden layer

N — is the number of neurons of the next (in this case, output) layer. The iteration includes enumerating all training pairs and summing the mean square errors of predictions over all network outputs of all training pairs $E(i)$. The learning process ends when the difference between the total errors of the current and previous iterations $E(i) - E(i-1)$ is less than a given threshold.

In this paper, we propose a model consisting of two similar perceptrons, one of which is applicable for predicting the hourly load profile of a working day, the other for a weekend or holiday. In the proposed model, the learning processes for workdays and weekends (or holidays) were separated, i.e., four matrices of synaptic weights were calculated (two for each perceptron). When training one neural network, the actual data of only working days were selected, and the resulting weight matrices were used only when predicting the load profile of working days. The second neural network was used to predict the load profile of weekends (or holidays). In the case under consideration, the output layer, which is an hourly load profile for a day, contains 24 neurons, the hidden layer contains 15 neurons, and the input layer contains 30 neurons. The input vector contains the characteristics of electricity consumption for the previous day (3 neurons), the average daily air temperatures of the previous day and the forecast of the air temperature of the forecast day (2 neurons). When training the perceptron, the known actual temperature was used as a temperature prediction. To account for seasonality, 12 neurons are used by the number of months in a year and three by the number of decades L . A. Delegodin 69 per month, and to account for the type of day - 10 neurons (seven days of the week, holidays, pre-holiday and post-holiday days). The number 10 is supplied to three of these inputs, corresponding to the month, decade in the month and the type of the predicted day, and the number 1 is supplied to the other inputs. the value of the maximum and minimum hourly costs for the previous day, in the calculation of which it is taken into account whether the forecast day is a working day or a day off (or a holiday). The learning process for each network includes two stages. First, the network is trained on the entire set of training pairs (a time interval of up to 1.5 years, depending on the actual data available in the database), and then it is retrained on a minimum time interval (a month preceding the predicted day). The training pair is the input vector and the target vector (known actual hourly load profile for the day represented by the input vector). The values of the input and target vectors are normalized, that is, they are converted to relative values in the range from 0 to 1. The forecast accuracy based on the use of artificial intelligence methods depends on the available input data that determine the network architecture, the degree of data reliability and the required forecast period. For short-term forecasting of the enterprise load, the necessary initial data are statistical reporting data on daily power consumption. For high reliability of the data used at the enterprise under study, a high-precision multifunctional automated system for monitoring and accounting for consumption should be initially implemented. Such a system was created, for example, in 2005 at the Novosibirsk Scientific Center within the framework of the Energy Saving SB RAS program [18] [24] [25]. The system

contains a central server with an integrated database and 26 peripheral data collection centers with their own local databases in the institutes of the center. An important result of the creation of this system is the receipt of a large array of sufficiently detailed information about the flows of various energy resources and the technological parameters of these resources.

Confidence intervals were calculated using a sample of 100 items. The average forecast error is 0.65%. Profiles of daily actual and predicted loads of consumption of active electricity for two decades of April: 1 - actual load; 2 - the predicted load is 0.87, the value of the confidence interval is 0.27 kWh, the minimum is 0.01 kWh, the average error of the hourly forecast is 1.87%. The average forecast error is 0.65%. Let's consider in detail the construction of a fragment of a neural network. The LVQ (Learning Vector Quantization) network – learning vector quantization – or the input vector classification network, which is a development of Kohonen's self-organizing map (SOM) [7] [14], was chosen as a neural network (NS) (Figure 4).

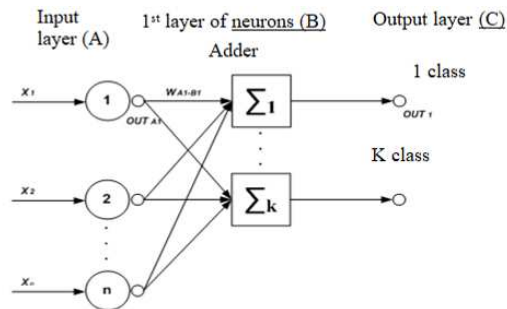


Figure 4

Block diagram of a neural network

Let's consider the description of the main actions of the developed algorithm.

The neural network analysis 1...k of the states of the energy consumption system of the house is carried out simultaneously by parallel analysis of the registered signal by each neural network, as a result of the analysis, an array of data is obtained from the outputs of the neural networks. Let's make an amendment: we are talking about the energy consumption system – energy supply. Further in the text, we will use one of the components of the system for brevity.

Then the post-processing of the neural network results is carried out, and the data from the outputs of the neural networks is sent to logical blocks designed to identify the corresponding state of each block of the energy consumption system of the house, and then to the priority encoder designed to highlight the corresponding state of the energy consumption of the house as a whole. Then the output of the neural network analysis result is carried out. Logic blocks are based on decision rules for each state of the home's energy system.

The construction of decision rules for analyzing the outputs of neural networks is based on the fact that deviations at each of the MI localizations are not manifested in all leads. To construct the decision rules for choosing one of the k states of the energy consumption system of a house, Table 2 is compiled. Table 2 presents combinations of the presence and absence of signs of excess power consumption in various localizations [15] [20].

Table 2
Household Energy System - Possible Scenarios

Scenario	Appliances consuming electricity														
	Ref	TV	Est	Ov	WsM	PC	Pot	DsM	Vac	Ir	McW	MCo	bulb	Ht	Led
Sc 1	+	-	+	-	+	-	-	+	+	+	-	-	-	-	+
Sc 2	+	-	-	-	+	+	-	+	-	-	-	-	+	+	-
Sc 3	+	+	-	-	+	+	+	+	+	+	+	+	-	-	+
Sc 4	+	+	-	+	-	+	+	+	-	-	+	+	-	+	+
Sc 5	+	-	+	-	-	+	+	+	-	-	-	-	+	-	-
Sc 6	+	+	-	+	-	-	-	+	+	+	+	-	-	-	-
Sc 7	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Sc 8	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+

Logic blocks are used to analyze data from the outputs of each NS. One logical block is used for each state of the home energy system. The inputs of the block receive the data obtained as a result of the NSA of those leads, which, in accordance with Table 1, "signal" the presence of excess power consumption and the data obtained as a result of the neural network analysis of the leads that do not "signal" the presence of a critical excess, i.e. correspond to the norm.

Decision rules for eight situations are constructed in accordance with Table 2:

- 1) The energy generated by solar panels is sufficient, you can use the energy storage mode:

$$F_1 = \overline{Ref} \& \overline{TV} \& \overline{Est} \& \overline{Ov} \& \overline{WsM} \& \overline{PC} \& \overline{Pot} \& \overline{DsM} \& \overline{Vac} \& \overline{Ir} \& \overline{McW} \& \overline{MCo} \& \overline{bulb} \& \overline{Ht} \& \overline{Led}$$

- 2) The energy generated by solar panels is sufficient:

$$F_2 = \overline{Ref} \& \overline{TV} \& \overline{Est} \& \overline{Ov} \& \overline{WsM} \& \overline{PC} \& \overline{Pot} \& \overline{DsM} \& \overline{Vac} \& \overline{Ir} \& \overline{McW} \& \overline{MCo} \& \overline{bulb} \& \overline{Ht} \& \overline{Led}$$

- 3) The energy generated by solar panels is enough for the operation of devices, connect an alternative source or battery for insurance:

$$F_3 = \overline{Ref} \& \overline{TV} \& \overline{Est} \& \overline{Ov} \& \overline{WsM} \& \overline{PC} \& \overline{Pot} \& \overline{DsM} \& \overline{Vac} \& \overline{Ir} \& \overline{McW} \& \overline{MCo} \& \overline{bulb} \& \overline{Ht} \& \overline{Led}$$

- 4) The energy generated by the solar panels is enough for the operation of

the included devices, for insurance, connect an alternative source or battery:

$$F_4 = \overline{Ref} \& \overline{TV} \& \overline{Est} \& \overline{Ov} \& \overline{WsM} \& \overline{PC} \& \overline{Pot} \& \overline{DsM} \& \overline{Vac} \& \overline{Ir} \& \overline{McW} \& \overline{MCo} \& \overline{bulb} \& \overline{Ht} \& \overline{Led}$$

- 5) The energy generated by the solar panels is enough for the operation of the included devices:

$$F_5 = \overline{Ref} \& \overline{TV} \& \overline{Est} \& \overline{Ov} \& \overline{WsM} \& \overline{PC} \& \overline{Pot} \& \overline{DsM} \& \overline{Vac} \& \overline{Ir} \& \overline{McW} \& \overline{MCo} \& \overline{bulb} \& \overline{Ht} \& \overline{Led}$$

- 6) The energy generated by solar panels is sufficient, you can use the energy storage mode:

$$F_6 = \overline{Ref} \& \overline{TV} \& \overline{Est} \& \overline{Ov} \& \overline{WsM} \& \overline{PC} \& \overline{Pot} \& \overline{DsM} \& \overline{Vac} \& \overline{Ir} \& \overline{McW} \& \overline{MCo} \& \overline{bulb} \& \overline{Ht} \& \overline{Led}$$

- 7) The energy generated by solar panels is barely enough to operate the devices, connect an alternative source or battery for insurance:

$$F_7 = \overline{Ref} \& \overline{TV} \& \overline{Est} \& \overline{Ov} \& \overline{WsM} \& \overline{PC} \& \overline{Pot} \& \overline{DsM} \& \overline{Vac} \& \overline{Ir} \& \overline{McW} \& \overline{MCo} \& \overline{bulb} \& \overline{Ht} \& \overline{Led}$$

- 8) The energy generated by solar panels is surplus, you can use the energy storage mode:

$$F_8 = \overline{Ref} \& \overline{TV} \& \overline{Est} \& \overline{Ov} \& \overline{WsM} \& \overline{PC} \& \overline{Pot} \& \overline{DsM} \& \overline{Vac} \& \overline{Ir} \& \overline{McW} \& \overline{MCo} \& \overline{bulb} \& \overline{Ht} \& \overline{Led},$$

Here,

$$\overline{Ref}, \overline{TV}, \overline{Est}, \overline{Ov}, \overline{WsM}, \overline{PC}, \overline{Pot}, \overline{DsM}, \overline{Vac}, \overline{Ir}, \overline{McW}, \overline{MCo}, \overline{bulb}, \overline{Ht}, \overline{Led}, \overline{TV}, \overline{Est}, \overline{Ov}, \overline{WsM}, \overline{PC}, \overline{Pot}, \overline{DsM}, \overline{Vac}, \overline{Ir}, \overline{McW}, \overline{MCo}, \overline{bulb}$$

are data from the outputs of the neural network.

At the stage of analyzing the outputs of neural networks (1 ... k)., Based on the decision rules (1 ÷ 8), a decision is made on the state of the power supply of the house. With the help of a priority encoder, one of the particular solutions obtained as a result of the analysis of the value of k by neural networks is selected. At the output of the priority encoder, a code of the input line number is formed, to which a positive input signal comes (a signal of a logical unit from the output of one of the k neural networks participating in the analysis). When several input signals arrive simultaneously, an output code is generated that corresponds to the input with the highest number, i.e. the higher inputs have priority over the lower ones. Therefore, the scrambler has priority. The result of performing the work of this stage is the number of the conclusion on the state of the energy consumption system. Then the resulting number is assigned a verbal description of the conclusion on the state of energy consumption, which is reported to the user.

The influence of the following parameters was investigated to assess the quality of training: the redundancy of the training sample, the amount of noise, the magnitude of the shift, the number of neurons in the hidden layer, the number of learning epochs. One of the reasons for the identified drawback is the problem of "dead" neurons, the essence of which is that in the process of learning the neural network, in reality, the weights of only a limited number of neurons are updated.

As a result of the research carried out, the following numerical indicators of the quality of teaching NS LVQ were obtained: specificity: 80%; sensitivity: 70%; generalization error: 0.25; learning error: 0.08. After training the neural network, the structure of the constructed fuzzy model will contain 252 rules. The two response surfaces of the system, obtained as a result of training, are shown in the Figure 5.

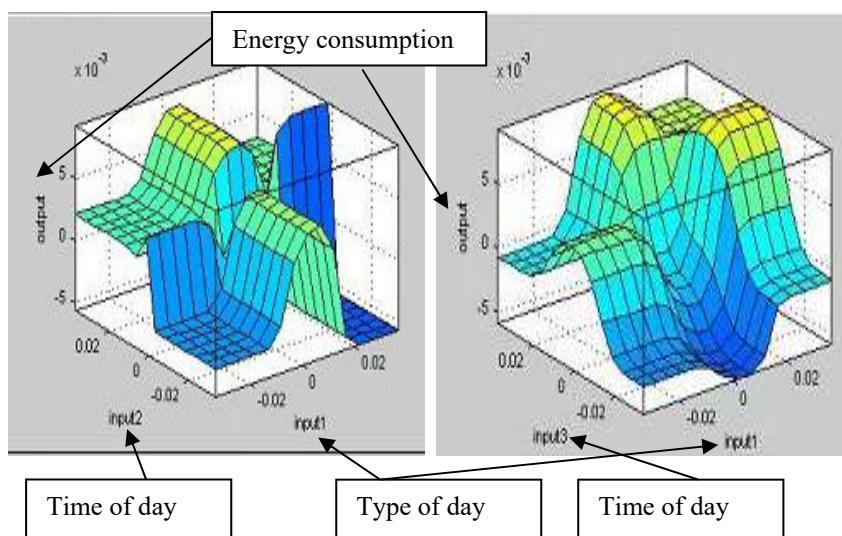


Figure 5

System response surfaces obtained as a result of training

The defuzzification stage allows to get a non-fuzzy value for each of the output variables, using the results of the accumulation of all output linguistic variables. The resulting surface allows to analyze the dependence of the values of the output variable on individual input variables. Combinations of input variables are set in accordance with their placement on the axes of the coordinate system. On the graph on the left, the dependence of energy consumption on the time of day and type of day, on the right, on the season and type of day.

The results obtained indicate the possibility of using the proposed approach to predict the electrical load. Neural networks are a suitable tool for solving energy consumption forecasting problems, alternative to traditional statistical methods.

Further improvement of the forecast accuracy is possible due to more accurate and fine tuning of the network structure and changing the number of input parameters.

The software is implemented in the C# environment in accordance with the developed schemes and requirements. Certificate of authorship for software #16772 on 14.04.2021 “Residential building intelligent energy management system ”Smarthouse”.

Development of the experimental installation and software will be discussed in details on AIS2022 and in the next publications.

Conclusions

The modern advanced approaches to the energy saving policy of such developed countries as the USA, Japan, Great Britain and some countries of the European Union are considered. Energy efficiency development trends in the world and the Republic of Kazakhstan.

The analysis of studies on smart home automation systems is carried out, their methods and their potential in the field of issuing recommendations for energy saving are discussed. We suppose that unpredicted situations, such as a pandemic, for example, will not significantly affect the accuracy of the system.

After analyzing the resulting fuzzy inference surface, we can conclude that it corresponds to expert ideas in the subject area under consideration. So, for example, it can be said that with the approach of the winter months and the simultaneous increase in load in the morning and evening hours, the situation becomes more complicated and more energy is required from the AC network. Accordingly, during the night hours in the summer, the battery power from solar panels is sufficient. Thus, the power of the selected solar panels is sufficient for the warm season at night and daytime, provided there are working days without the use of an alternating current network. At the same time, it is impossible to refuse to connect to the general AC network during peak loads in the morning and evening, as well as, in the cold season due to the connection of heating.

Acknowledgement

This work was supported by East-Kazakhstan technical University.

References

- [1] World Energy Outlook 2020, IEA (2020) IEA, Paris <https://www.iea.org/reports/world-energy-outlook-2020>
- [2] BP Global Statistical Review of World Energy, 2021, <http://www.bp.com/>
- [3] DNV's Energy Transition Outlook is an independent, model-based forecast of the world's most likely energy future through to 2050, 2021, <https://eto.dnv.com/2021#ETO2021-top>
- [4] European policies in energy saving, 2019, <http://europa.eu!/Tj97Qn>

- [5] Law of the Republic of Kazakhstan dated July 4, 2009 No. 165-IV "On support for the use of renewable energy sources" Article 1. Basic concepts used in this Law
- [6] Message from the President of the Republic of Kazakhstan Kassym-Zhomart Tokayev, Nur-Sultan, President's Address «Unity of the people and systemic reforms are a solid foundation for the nation's prosperity» September 2, 2021. Available at: <https://primeminister.kz/en/addresses/01092021>
- [7] Asaithambi S., Venkatraman S., Venkatraman R.: Big Data and Personalisation for Non-Intrusive Smart Home Automation, Big Data and Cognitive Computing, 2021, 5. 6. 10.3390/bdcc5010006
- [8] Györök G. Interactive monitoring of Electronic Circuits with Embedded Microcontroller, 19th IEEE World Symposium on Applied Machine Intelligence and Informatics, SAMI 2021; Slovakia; 2021, pp. 223-228
- [9] Harizaj M. and Ndreu A.: Living in 'Smart Cities and Green World', SCRD, Vol. 6, No. 3, pp. 27-40, Jun. 2022
- [10] Beszédes B., Széll K., Györök G. Redundant photo-voltaic power cell in a highly reliable system, Electronics (Switzerland), V. 10, Issue 11, 1 June 2021, No 1253
- [11] Björkskog C.: Human Computer Interaction in Smart Homes. Helsinki, Finland. Available at: <http://www.hiit.fi/~oulasvir/58307110/smarthomes.pdf> [Accessed March 20, 2021]
- [12] Jakkula, V., Youngblood, G. & Cook, D.: Identification of lifestyle behavior patterns with prediction of the happiness of an inhabitant in a smart home. ... Approaches to Beauty and Happiness. Available at: <http://www.aaai.org/Papers/Workshops/2006/WS-06-04/WS06-04-005.pdf> [Accessed March 12, 2021]
- [13] Nowak M. & Urbaniak A.: Utilization of intelligent control algorithms for thermal comfort optimization and energy saving, 2011 12th IEEE Control Conference, pp. 270-274, Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5945862
- [14] Clinckx N.: Smart Home: Hope or hype?, January 2013, pp. 1-20
- [15] Dementiev A.: «Smart» house in XXI century. Moscow, Publishing solutions, 2017 – 174 p
- [16] Villar J., Cal E. de la & Sedano, J.: A fuzzy logic based efficient energy saving approach for domestic heating systems. Integrated Computer-Aided Engineering, 15, 2009, pp. 1-9, Available at: <http://iospress.metapress.com/index/L74647166223125U.pdf> [Accessed March 12, 2021]

- [17] He Y.: Energy saving of central air-conditioning and control system: Case study: Nanchang Hongkelong Supermarket. Available at: <http://theseus17-kk.lib.helsinki.fi/handle/10024/21077> [Accessed March 4, 2021]
- [18] Olaru LM, Gellert A, Fiore U, Palmieri F.: Electricity production and consumption modeling through fuzzy logic. *Int J Intell Syst.* 2022;1-17. doi:10.1002/int.22942
- [19] Industry Trends. Building the intelligent business platforms of tomorrow Industry Trends, <https://atos.net/content/mini-sites/look-out2020/assets/> (Accessed 12.04.2021)
- [20] Inji, E., Attia, I. & Hamdy, P.: Energy Saving Through Smart Home., (2), 2011, pp. 223-227
- [21] Shvets O., Seebauer M., Naizabayeva A., Toleugazin A.: Autonomous power supply systems optimization for energy efficiency increasing. 15th International Symposium on Applied Informatics and Related Areas organized in the frame of Hungarian Science Festival 2020 by Óbuda University, 12.11.2020, pp. 128-132
- [22] Dab K, Agbossou K, Henao N, Dubé Y, Kelouwani S, Hosseini SS.: A compositional kernel based Gaussian process approach to day-ahead residential load forecasting. *Energy Build.* 2022;254:111459
- [23] Zhou F, Wang Z, Zhong T, Trajcevski G, Khokhar A.: HydroFlow: towards probabilistic electricity demand prediction using variational autoregressive models and normalizing flows. *Int J Intell Syst.* Forthcoming 2022
- [24] Today in energy, Energy Information Agency, USA, Sept. 2021, <https://www.eia.gov/todayinenergy/>
- [25] Albert J. R.: Design and Investigation of Solar PV Fed Single-Source Voltage-Lift Multilevel Inverter Using Intelligent Controllers. *J Control Autom Electr Syst* 33, 1537-1562 (2022) <https://doi.org/10.1007/s40313-021-00892-w>

Application of the QFD Technique Method in Logistics Strategy

Eszter Sós, Péter Földesi

Széchenyi István University, Department of Logistics and Forwarding,
Egyetem tér 1, 9026 Győr, Hungary, sos.eszter@sze.hu; foldesi@sze.hu

Abstract: One of the most significant corporate challenges today is to meet customer expectations. In order for customer satisfaction to be achieved, it is necessary to review the entire corporate system and related processes and coordinate the various corporate strategies. In the recent past, it was widely regarded as sufficient by managers to develop the right marketing strategy in order to sell a product. However, there is currently a discussion as to whether a marketing logistics strategy as a well-designed logistics environment is needed to sell the product and thereby gain customer satisfaction. In this article, we present the Quality Function Deployment (QFD) technique, an effective tool for transforming consumer needs into technical, quality characteristics. The method of QFD technique can also be successfully applied in the field of logistics. Utilizing it ensures the possibility of examining the impact of the sub-areas and processes of marketing and logistics services on the basis of customer needs. In addition, visual control can be used to illustrate that whilst two products require the same logistics strategy, lead times already cause significant differences in the interaction of logistics processes and technological parameters. The analysis also highlights the shortcomings of the logistics environment, thereby supporting the decision-making of the company management in both marketing and logistics strategy planning.

Keywords: logistics; marketing; logistics strategy; 6R; Quality Function Deployment (QFD); visual control

1 Introduction

The life and decision-making structure of companies have undergone significant changes in recent decades. Due to the competitive nature of the market, meeting customer expectations at the highest possible level has come to the forefront, and marketing has also gained significant ground.

As a result, expert groups specializing in marketing strategy are constantly formed to deal with the analysis of various effects on consumer decisions and define advertising strategies and campaigns for the sale of goods.

When defining a marketing strategy, it is necessary to define the logistics strategy at the same time, as customer service requires a well-established logistics system. In order to map the logistics environment required for a given product, the Quality Function Deployment (QFD) technique method was used [1]. After the presentation of the QFD technique, the applicability of this method in the field of logistics is described.

The adequacy of logistics processes depends to a large extent on choosing the right processes for the right logistics strategy. The appropriate logistics solution depends on what strategic decisions are made, such as Push or Pull, centralized or decentralized, single-stage or multi-stage in distribution, single-channel, or multi-channel.

Different logistics strategies can be well supported by different methods, for which the QFD technique is an excellent analogy [2]. But it is not only beneficial to plan things but also to recognize the real difference between each logistics strategy and possibly to explore new strategic considerations.

Therefore, in this article, as a working hypothesis, we also included the issue of the obvious short and long lead times in the study, which resulted in an interesting finding.

For our analysis, we constructed a relationship matrix that forms the core of the QFD technique for four different products: examples with short and long lead times in Push and Pull systems. At the intersections of the matrix, we use the values according to the relevance of the given relationship.

From the values obtained here, we show through visualization what differences are caused by different logistics strategies and nonidentical lead times.

With our research, we would like to explain that, taking into account the specifics of a given product, the same system is implemented through different logistics processes.

After evaluating the relationship matrix, it can be clearly decided what logistics strategy the product requires, and this can also be used to develop the company's marketing logistics strategy [3].

2 The Impact of Marketing on Logistics Strategy

To raise the issue, we first describe, through an example, what happens when a marketing strategy is not aligned with the logistics system.

The case in point is when a product suddenly gains popularity as a result of advertisements, but the associated supply chain is insufficiently prepared for this increased demand.

The customer enters the store with the intention of purchasing the product that has caught everyone's attention as a result of the marketing strategy. After serving the first few customers, the store runs out of stock, meaning only a portion of the increased demand is met, causing in turn dissatisfaction in the customers. The customer either has to then wait for a new supply of the product or look for another point of sale where the product is in stock, but perhaps not for the same price.

Such an example clearly shows that for a product to succeed, the needs for logistics services (6R) must also be met. Indeed, meeting customer expectations not only means that the right product is available of the right quality and at the right cost, but also in the right place, at the right time and in the right quantity, since each product is sold together with the associated logistics service.

In order to analyze the current state of the logistics system and its possible shortcomings, it is necessary to examine the following areas:

- Check product availability and stock available at the depot.
- Consideration should be given to whether storage capacity is available when purchasing larger stocks, or whether customers can be served by intermittent delivery.
- It is necessary to map whether there is a supplier who ensures regular, scheduled delivery and is able to respond appropriately to suddenly increased demands.

If we examine a product in the light of the entire supply chain, then an analysis of the production parameters is also necessary, as the capacity of the factory and the availability of the raw material also significantly influence the satisfaction of customer needs. [4].

Based on the examination of the elements of the supply chain, it can be suggested that it is not enough to only discuss marketing or logistics. In the case of a well-functioning company, it becomes necessary to define a marketing logistics strategy, since the product is sold together with the related services. Therefore, the efficiency of a company's logistics system is reflected in the quality of customer service [5].

In order to quantify customer needs, it is necessary to choose a technique, the application of which not only deals with one sub-process but also connects customer expectations in a targeted way with various company activities and technical parameters.

One of today's forgotten quality techniques is the Quality Function Deployment (QFD) technique, the use of which has declined significantly in Europe, but is still used in the Far East as a key element of engineering design [6].

3 Application of QFD Technique

The QFD philosophy originated in Japan, and its concept was developed by Professor Yoji Akao in 1966.

According to Akao, QFD is a technique for developing a quality product for customer satisfaction and translating consumer requirements into design goals throughout the manufacturing process, ensuring the realization of a quality product at the design stage, and extending quality control to the stage where there is not a finished product, just a concept or plan [7].

Shortly after the introduction of this technology significant cost reductions were seen, and its application spread rapidly among manufacturing companies.

The QFD technique is basically a design process that takes a qualitative approach to new products. Design, development, and implementation are also driven by customer needs and values [8].

It aims to meet the highest possible level of consumer needs through the design of manufacture and production planning processes developed by engineers, taking into account customer expectations. It can also be used as a documentation tool, as it provides an overview of each step of the design [9].

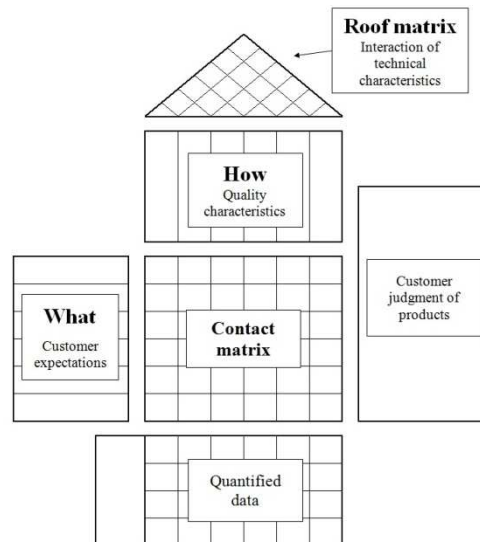


Figure 1

Structure of the House of Quality (HOQ) with the functions of each area [13]

A practical formalism of transforming consumer needs into technical, quality features is the House of Quality (HOQ) [10], as shown in Figure 1. It works as an aid to examine the expectations of the market, the technical factors influencing

satisfaction, the quality of competitor products, and it also helps in determining the technical parameters.

The quality house matrix of the QFD technique can be used effectively to design logistics strategies [11]. The needs of the customer can be examined both in the design of the new logistics environment and in relation to the existing logistics system.

This clarifies the logistics strategy applicable to the introduction of a given product, thus avoiding a potentially erroneous corporate decision. [12].

As described in the previous chapter, QFD is a systematic design preparation method that seeks to more fully meet customer needs, a technique used by manufacturers and service providers to gather requirements, expectations, and customer purchasing decision factors [14].

It is essential to realize customer needs thoroughly and to translate these needs into the professional, technical language of product design so that customer expectations are met [15].

To apply the method, a marketing and technical survey must first be prepared. In the first round, the customer needs and the priorities set for the product are collected. The survey is prepared on the customer side using the interview and questionnaire technique, and on the corporate side using the brainstorming method.

The parameters of the competing products and their suitability for the customer must be examined, including niche markets. If available, product compliance can also be supported by analyzing previous product development experience. Part of the preparation of the survey is also the cost estimate of the production process.

The practical tool for the implementation of the method is the previously mentioned quality house, in which data is processed by matrix technology.

The data needed to complete the matrix is collected and made available by different disciplines.

The task of marketers is to map the needs of the customer, by asking the "What does the customer want?" question, and filling the answers in the rows of the matrix.

The data required for the columns of the matrix is compiled by the product or service designers, based on the technical, quality characteristics, and product parameters suitable for satisfying the customer's expectations, by answering the "How is it implemented?" question.

At the intersections of the rows and columns, symbols or numbers are placed according to the relevance of the given relationship, which expresses the correlation between customer needs and the characteristics taken into account during the design.

Quantified data is collected in the continuation of the columns, and in the extension of the rows, customer opinions of the products are included. The roof matrix provides space for the interrelationships of the designed product characteristics.

During the practical implementation of the QFD, the planning can be further divided into 4 stages: product and component design, process, and production planning. Accordingly, we can apply a four-phase QFD procedure with the construction of 4 different quality housings, which creates an analysis based on customer specifics, taking into account the specifics of the fields [16].

Throughout the four-phase QFD process, product parameters can be determined from the conceptual characteristics of the product, and process characteristics can be determined in turn from them. Due to the transparency of the design, they provide a firm basis for the preparation of complete product documentation, which can be presented during an audit or when the product is approved, if required.

4 Use of Multiphases QFD Method in Logistics

By applying the QFD method in the field of logistics, the target values of the design and production process can be determined through the derivation of primary customer needs.

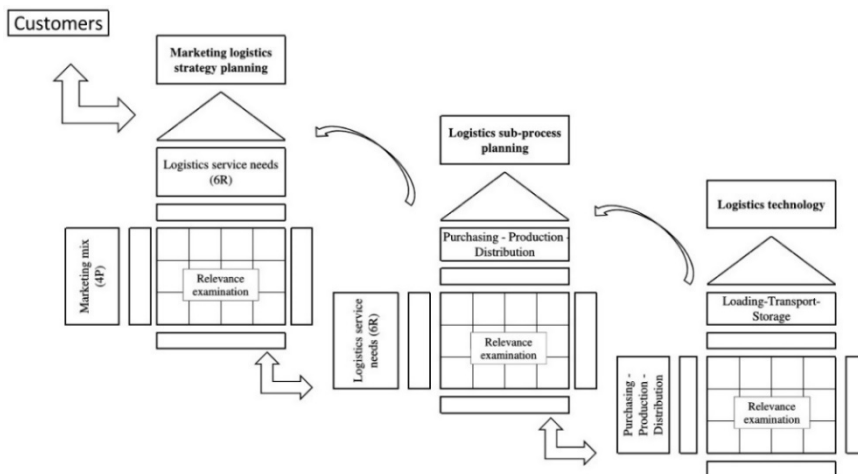


Figure 2

Three phases in the process used to design logistics strategies strategies

The use of the matrix and the study of the interaction of processes and parameters support the selection of the appropriate logistics strategy for a given product.

One of the main differences from the original four-step process is that when designing logistics systems, we can discuss a three-phase process (Figure 2), as the

component design is already integrated into the process as an element of logistics systems.

Another difference between the quality houses of the QFD technique and the houses used in logistics is that we perform a relevance test of the indicated sub-processes, thus clarifying the effect of the processes on each other by examining a given product.

The elements of the relationship matrix for examining the relevance of the three-phase logistics procedure have been elaborated in detail. The starting point of the study is in each case the marketing mix elements are defined by customer expectations. These are compared through three steps to the needs related to the logistics service, then to the logistics sub-processes and technological parameters.

In the following, the elements of the relationship matrix are detailed for each phase.

		Logistics service						
		R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	
		λ ₁	λ ₂	λ ₃	λ ₄	λ ₅	λ ₆	
		The right product	In the right quantity	Of the right quality	At the right time	In the right place	For the right cost	
Marketing mix (4P)	P ₁	Y ₁ Quality						
		Y ₂ Physical appearance (design)						
		Y ₃ Product availability						
		Y ₄ Manufacturing technology						
		Y ₅ Customer service						
		Y ₆ Logistics resilience *						
	P ₂	Y ₆ Price						
		Y ₇ Discounted price						
	P ₃	Y ₈ Product life cycle						
		Y ₉ Number of distribution channels						
		Y ₁₀ Product location						
		Y ₁₁ Assortment						
		Y ₁₂ Stock						
	P ₄	Y ₁₃ Advertisement						
Y ₁₄ Sales promotion								

Figure 3

The first phase: the marketing-logistics strategic planning

The first house (Figure 3) is designed to support marketing-logistics strategy planning. Therefore, among the elements of the marketing mix, we introduced a new concept called “logistics resilience”. This is necessary as a product may be suitable in all respects, but if it is unsuitable for storage and transport, it will not reach the buyer.

The product alone does not represent value in use, only if we add both the space value and the time value produced by logistics. Thus, the characteristics of the product related to its traditional value in use should not contradict its logistics parameters (e.g., an excellent product but not deliverable). For this reason, even at the design stage, it is necessary to examine how suitable the product is for logistics, i.e., whether it can be transported and stored properly.

Logistical resilience, therefore, means that we have to produce a product with a value in use on which we can still place the logistical burdens needed to produce place value and time value.

			Procurement					Manufacturing					Distribution				
			I ₁	I ₂	I ₃	I ₄	I ₅	M ₁	M ₂	M ₃	M ₄	M ₅	O ₁	O ₂	O ₃	O ₄	O ₅
Logistics service needs (6R)	R ₁	λ ₁	The right product														
	R ₂	λ ₂	In the right quantity														
	R ₃	λ ₃	Of the right quality														
	R ₄	λ ₄	At the right time														
	R ₅	λ ₅	In the right place														
	R ₆	λ ₆	For the right cost														
			Procurement planning	Price and cost analysis	Selection of suppliers	Track orders	Determining material requirements plans	Production schedule	Lead time	Constant production volume	Inventory level	Fast, low-cost transition	Delivery item sizes	Number of goods handling	Fast and reliable delivery method	Inventory costs	Strong IT relationship between manufacturer and customer

Figure 4

The second phase: the logistics sub-process planning

The second phase (Figure 4) is the logistics sub-process design, where we examine logistics service needs (6R) in relation to procurement, production, and distribution processes.

In this relationship matrix, the levels of relevance between 6R and logistics sub-processes can be mapped to help clarify the relevance of logistics sub-processes to the expectations for logistics services when examining a given product.

The third phase (Figure 5) is to examine the components of logistics technology, where the elements of procurement, production, and distribution are compared with the parameters of loading – transport - storage.

In this relationship matrix, the relevance of the previously formulated logistics sub-processes with technological elements is examined. It can be used to clarify how logistics sub-processes and technological parameters interact for a given product, thus clarifying the applicable logistics strategy.

		Loading			Transport				Storage				IT	
		L ₁	L ₂	L ₃	T ₁	T ₂	T ₃	T ₄	T ₅	W ₁	W ₂	W ₃	W ₄	β ₁
Procurement	I ₁	Procurement planning												
	I ₂	Price and cost analysis												
	I ₃	Selection of suppliers												
	I ₄	Track orders												
	I ₅	Determining material requirements plans												
Manufacturing	M ₁	Production schedule												
	M ₂	Lead time												
	M ₃	Constant production volume												
	M ₄	Inventory level												
	M ₅	Fast, low-cost transition												
Distribution	O ₁	Delivery item sizes												
	O ₂	Number of goods handling												
	O ₃	Fast and reliable delivery method												
	O ₄	Inventory costs												
	O ₅	Strong IT relationship between manufacturer and customer												
		Number of distribution channels												
		Documentation of incoming goods												
		Quality control												
		Shipping distance												
		Delivery frequency												
		Delivery unit												
		Distance of depots												
		Order item size												
		Storage costs												
		The size of the stock to be stored												
		Warehouse administration												
		Number of types of goods												
		IT support												

Figure 5
The third phase: the logistics technology

4.1 Practical Application of the Method to Select the Appropriate Logistics Strategy

In order to be able to properly illustrate the application of the QFD technique method in the field of logistics, we created a relationship matrix that forms the “trunk” of the first, second, and third houses for 4 products. For each product tested,

a relevance study was performed using the three-phase logistics procedure, which can be found in Appendices 1, 2, 3, and 4.

The relationship matrices included the parameters compiled on the principle from the previous chapter. Our decision was based on the fact that we can see the relevance of the processes in context through the application of the three-phase logistics procedure. In this way, it can be clarified what differences can be discovered during strategic decisions.

The starting point for defining the products is the Push and Pull systems used in logistics, which already assume a logistics strategy of sorts. For this, we examine products with different lead times. In the present study, we selected the examples that form the basis of the relevance study based on these parameters. The exemplary products in Appendices 1, 2, 3, and 4 represent our present decision and can be modified as needed during further studies.

During the completion of the relationship matrices, we analyze the effect of the listed, previously known expectations and the different sub-processes on each other.

In the relationship matrices in Appendices 1, 2, 3, and 4, the relevance assessment is examined in the range 0, 1, and 2, to which a meaning is also associated in the legend (Table 1). These values can be normalized later with Fuzzy [17]. Since the use of Fuzzy sets in logistics is often closer to the practical approach, it is necessary to convert the crisp numbers in the results into Fuzzy sets, that is, to properly evaluate the results, it is necessary to use the Fuzzy QFD technique [21].

Table 1
Legend for Appendices 1, 2, 3, 4

Legend	
2	Significantly relevant
1	Relevant
0	Neutral /not relevant

On the relationship matrices of the product tests in the appendix, it clearly seems that the Push and Pull strategy already fundamentally defines the logistics sub-processes of the product.

Products manufactured in the Push principle system are made based on forecasted order data, in larger quantities. In contrast, the products of the Pull system are made to a specific order, an express customer demand [18].

Lead time is a parameter that affects all logistics processes, as one of the important “milestones” of customer satisfaction is when the customer receives the product [19].

As can be clearly seen from the prepared contact matrices, lead time and customer expectations related to the product already cause significant differences in terms of logistics sub-processes as well as the marketing concept to be developed.

The examination of the four different products (Appendices 1, 2, 3, 4) clearly confirms that, given the specificities of a given product, the same system is implemented through different logistical processes.

Applying the method of QFD technique in the field of logistics clarifies what kind of logistics strategy a given product or service requires.

4.2 Impact of Push and Pull Logistics Strategy and Lead Time on Logistics Sub-Processes and Logistics Technology

The product-specific relevance studies in the appendix were based on the Push and Pull strategies and the different lead times. Using visual control techniques, we illustrate:

- differences in the relevance of products with the same logistics strategy but different lead times,
- and the difference in relevance between products with a defined logistics strategy and the same lead time.

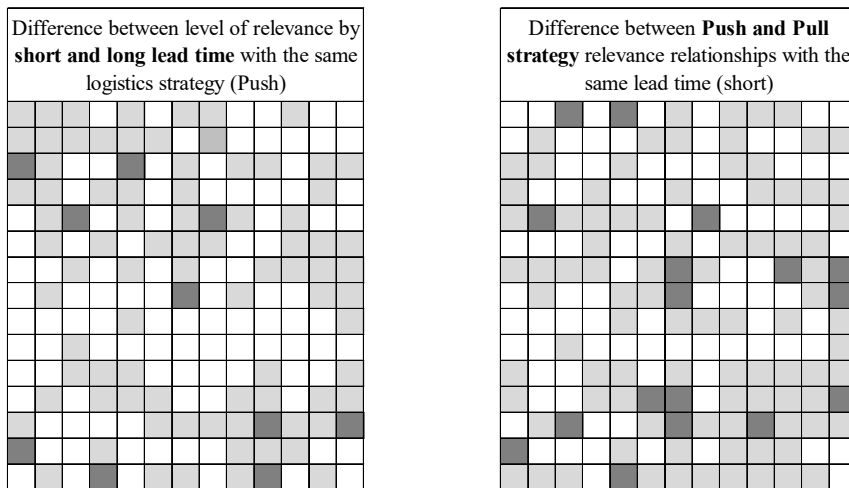


Figure 6

Using visual control technique to display the difference in relevance level between appendices 1 and 2 (left) and between appendices 1 and 3 (right)

Table 2

Legend for comparison with visual control technique (Figures 6 and 7)

Legend	
	Same level of relevance
	Relevance slightly different
	Relevance significantly different

Table 3
Data for the analysis of Figure 6

Legend	Level of difference	Difference between level of relevance by short and long lead time with the same logistics strategy (Pull)		Difference between Push and Pull strategy relevance relationships with the same lead time (long)	
		Number of deviations	%	Number of deviations	%
■	2	17	0,09	6	0,03
■	1	83	0,42	77	0,40
□	0	95	0,49	112	0,57

The visual control technique helps make the differences clear (Figures 6 and 7; Table 2). If everything in the figures were white, it would mean that a uniform logistics system could be applied to all products. As different customer expectations emerge, so do increasing differences in product-related logistics processes. The visualizations are produced to highlight the tendency for dissimilar products to require different logistics strategies.

Legend	Level of difference	Difference between level of relevance by short and long lead time with the same logistics strategy (Push)		Difference between Push and Pull strategy relevance relationships with the same lead time (short)	
		Number of deviations	%	Number of deviations	%
■	2	10	0,05	17	0,09
■	1	82	0,42	91	0,47
□	0	103	0,53	87	0,44

Figure 7

Using visual control technique to display the difference in relevance level between appendices 3 and 4 (left) and between appendices 2 and 4 (right)

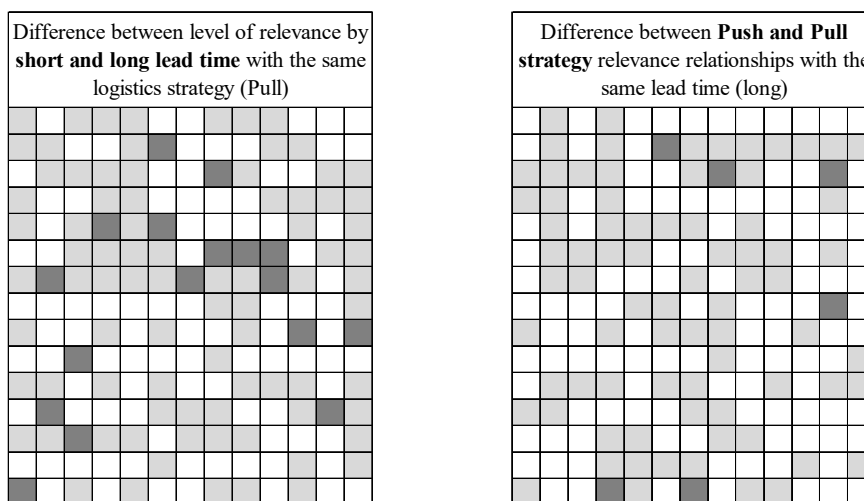


Table 4
Data for the analysis of Figure 7

In our present study, the visual control technique shows how changing the lead time and logistics strategy changes the interaction of product logistics processes and technological parameters. As can be seen in the legend, the darker the area, the more significant the difference. For the test, we used the third matrix of the relevance tests filled out with example products in appendices 1, 2, 3, and 4 called "Logistics technology", where the relevance level of logistics sub-processes was examined from the perspective of logistics technology elements. During the examination carried out with the help of visual control, we examined the differences between the relevance levels determined for the example products shown in Appendices 1, 2, 3, and 4. Dark gray refers to a significant difference between the examined products, which in this case takes the value "2". Light gray represents minor deviations, which in this case takes the value "1". White indicates that the relevance of the examined products shows no difference. The numbers show - based on the example products in Appendices 1, 2, 3, and 4 - how products with different logistics strategies and different lead times differ in terms of the logistics environment. As can be seen in Tables 3 and 4, the degree of deviations show similar percentages when examining the same logistics strategy with different lead times and when examining different logistics strategies with the same lead time.

The relevance studies of the example products with the Push - short lead time, Push - long lead time, Pull - short lead time, and Pull - long lead time logistics strategies in Appendices 1, 2, 3, and 4 also clearly show the significant differences between the distinct strategic decisions.

Examined in its context, it can clearly be seen through the prepared visualizations that a change in the applied logistics strategy or lead time already generates significant differences in the logistics sub-processes and the applied logistics technology.

Most of the available literature suggests strategic decisions to decide whether to use a Push or Pull logistics strategy for a particular product [20]. However, based on our analysis, it appears that at least as strategically important a decision or aptitude is whether a product has a long or short lead time. In our present study, the difference between short and long lead times based on the patch effect is larger than between the Push and Pull strategies.

The differences show that the choice between short and long lead times is of the same or sometimes greater strategic importance from a logistical point of view than when using a Push or Pull logistics strategy.

The conversion of the QFD method to logistics use and the preparation of a relevance study for a specific product provides a new scientific approach to the selection of a logistics strategy related to a specific product and to the further development of the already existing logistics environment.

Conclusions

There has been a significant recent increase in the importance of the marketing and logistics specialties, yet companies still fail to address the two areas together when developing strategies. As the market has a growing number of products with increasingly more choices, the marketing-logistics strategy has progressively gained more space.

In order to properly understand the importance of collaboration between the two areas, a technique must be used that helps meet the needs expressed by customers and the processes of engineering design. The QFD technique is used by a few, although it is one of the best methods for translating customer needs into engineering design parameters.

We found that the application of the QFD technique in the field of logistics can be used to examine the impact of each sub-area on each other, and the analysis also highlights the shortcomings of the logistics environment, thus supporting the design of logistics strategies.

In this article, we have used 4 different examples to show how this method can be used to analyze the logistics environment. We performed a relevance study, which can be used to determine the marketing-logistics strategy of a given product, through the logistics sub-processes and the applied logistics technology. In addition, we introduced the concept of logistics resilience as an element of the marketing mix, which is a new approach to product examination. It can be used to analyze whether a given product is suitable for logistics.

The analysis of the prepared connection matrices and the diagrams illustrated with the visual control technique clearly demonstrate that the logistics sub-processes and logistics technology parameters of a given product are significantly influenced by the applied logistics strategy and lead time. Therefore, it is not possible to apply a uniform logistics strategy to all companies; the specifics of the product must also be taken into account. The visual control technique provides an opportunity to subject the presented data to further, more detailed numerical analysis.

Acknowledgment

Project no. TKP2021-NKTA-48 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NKTA funding scheme.

References

- [1] V. E. Bottani, A. Rizzi: Strategic management of logistics service: A fuzzy QFD approach. *International Journal of Production Economics*, 2006, pp. 585-599, <https://doi.org/10.1016/j.ijpe.2005.11.006>

-
- [2] G. Z. Jia, M. Bai: An approach for manufacturing strategy development based on fuzzy-QFD. *Computers & Industrial Engineering*, 2011, pp. 445-454, <https://doi.org/10.1016/j.cie.2010.07.003>
- [3] S. Y. Sohn, I. S. Choi: Fuzzy QFD for supply chain management with reliability consideration. *Reliability Engineering & System Safety*, 2001, pp. 327-334, [https://doi.org/10.1016/S0951-8320\(01\)00022-9](https://doi.org/10.1016/S0951-8320(01)00022-9)
- [4] I. V. Kozlenkova, G. T. M.Hult, D. J. Lund, J. A. Mena, P. Kecec: The Role of Marketing Channels. *Supply Chain Management. In Journal of Retailing*, 2015, pp. 586-609, <https://doi.org/10.1016/j.jretai.2015.03.003>
- [5] M. S. Akdogan, A. Durak: Logistic and marketing performances of logistics companies: A comparison between Germany and Turkey. *Procedia - Social and Behavioral Sciences* 235, 2016, pp. 576-586, <https://doi.org/10.1016/j.sbspro.2016.11.084>
- [6] W. Ho, T. He, C. K. M. Lee, A. Emrouznejad: Strategic logistics outsourcing: An integrated QFD and fuzzy AHP approach. *Expert Systems with Applications*, 2012, pp. 10841-10850, <https://doi.org/10.1016/j.eswa.2012.03.009>
- [7] Y.Akao: QFD: Past, Present, and Future. 3rd International Symposium on Quality Function Deployment (ISQFD'97), Linköping, 1997. pp. 1-12
- [8] H. Wang, S. Liu: An integrated fuzzy QFD and grey decision-making approach for supply chain collaborative quality design of large complex products. *Computers & Industrial Engineering*, 106212, 2020, <https://doi.org/10.1016/j.cie.2019.106212>
- [9] A. Jahan, K. L. Edwards, M. Bahraminasab: 2 - Materials selection in the context of design problem-solving. *Multi-criteria Decision Analysis for Supporting the Selection of Engineering Materials in Product Design (Second Edition)*, 2016, pp. 25-40
- [10] D. R. Kiran: Quality Function Deployment. In *Total Quality Management, Key Concepts and Case Studies*, 2017, pp. 425-437, <https://doi.org/10.1016/B978-0-12-811035-5.00030-1>
- [11] A. M. Oddershede, L. E. Quezada, J. E. Valenzuela, P. I. Palominos, H. Lopez-Ospina: Formulation of a Manufacturing Strategy Using the House of Quality. In *Procedia Manufacturing*, 2019, pp. 843-850, <https://doi.org/10.1016/j.promfg.2020.01.417>
- [12] E. Bottani: A fuzzy QFD approach to achieve agility. *International Journal of Production Economics*, 2009, pp. 380-391, <https://doi.org/10.1016/j.ijpe.2009.02.013>
- [13] B. El-Haik: Quality Function Deployment. *Wiley StatsRef: Statistics Reference Online*, 2014, <https://doi.org/10.1002/9781118445112.stat04036>

- [14] Md. M. H. Chowdhury, M. A. Quaddus: A multi-phased QFD based optimization approach to sustainable service design. *International Journal of Production Economics*, 2016, pp. 165-178, <https://doi.org/10.1016/j.ijpe.2015.09.023>
- [15] F. Zhang, M. Yang, W. Liu: Using integrated quality function deployment and theory of innovation problem solving approach for ergonomic product design. *Computers & Industrial Engineering*, 2014, pp. 60-74
- [16] L. Buglione, 2AlainAbran, Maya Daneva, Andrea Herrmann: Chapter 7 - "Filling in the blanks": A way to improve requirements management for better estimates. *Software Quality Assurance*, 2016, pp. 151-176, <https://doi.org/10.1016/B978-0-12-802301-3.00007-7>
- [17] J. Menyhárt, R. Szabolcsi: Support Vector Machine and Fuzzy Logic. *Acta Polytechnica Hungarica* Vol. 13, No. 5, 2016, pp. 205-220, http://epa.niif.hu/02400/02461/00067/pdf/EPA02461_acta_polytechnica_hungarica_2016_05_205-220.pdf
- [18] D. F. Pyke, Morris A. Cohen: Push and pull in manufacturing and distribution systems. *Journal of Operations Management*, 1990, pp. 24-43, [https://doi.org/10.1016/0272-6963\(90\)90144-3](https://doi.org/10.1016/0272-6963(90)90144-3)
- [19] L. Liu, H. Xu, S. X. Zhu: Push verse pull: Inventory-leadtime tradeoff for managing system variability. *European Journal of Operational Research*, 2020, pp. 119-132, <https://doi.org/10.1016/j.ejor.2020.04.033>
- [20] E. Glistau, N. I. C. Machado, M. Schenk: Logistics Strategies and Tools. *MultiScience - XXIX. microCAD International Multidisciplinary Scientific Conference*, 2015, DOI:10.26649/musci.2015.027
- [21] E. Sós, P. Földesi: Fuzzy QFD assessment of logistics coherence. *IEEE 6th International Conference on Logistics Operations Management (GOL) 2022*, pp. 1-6, DOI: 10.1109/GOL53975.2022.9820009

Appendix

Appendix 1

PUSH System with short lead time. Example product: soft drink, self-service vending machine.

Marketing logistics strategic planning			Logistics service					
Marketing mix (4P)	P1	Y1	R1	R2	R3	R4	R5	R6
			λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
		Y2	2	1	0	0	0	1
		Y3	2	2	0	2	2	1
		Y4	0	0	2	0	0	1
		Y5	1	0	1	1	1	1
		Y6	1	0	1	0	0	1
	P2	Y7	1	1	2	0	2	2
		Y8	1	2	1	0	2	1
		Y9	1	0	2	0	2	1
	P3	Y10	2	2	0	2	2	2
		Y11	1	0	0	0	1	2
		Y12	2	2	0	2	2	2
	P4	Y13	2	2	1	2	2	2
		Y14	2	2	1	2	2	2

Logistics sub-process planning			Logistics service									
Logistics sub-process	R1	λ_1	R2	λ_2	R3	λ_3	R4	λ_4	R5	λ_5	R6	λ_6
Procurement planning	1	1	2	0	1	2	2	2	2	2	2	2
Price and cost analysis	2	2	2	1	0	2	2	2	2	2	2	2
Selection of suppliers	2	2	2	1	0	2	2	2	2	2	2	2
Track orders	2	2	2	1	0	2	2	2	2	2	2	2
Determining material requirements plans	0	2	2	2	1	0	0	1	0	0	1	2
Production schedule	1	1	1	1	1	1	1	1	1	1	1	1
Lead time	1	1	1	1	1	1	1	1	1	1	1	1
Constant production volume	0	1	0	0	0	0	0	0	0	0	0	0
Inventory level	1	1	1	2	2	0	2	2	2	2	2	2
Fast, low-cost transition	0	0	1	0	0	0	0	0	0	0	0	0
Delivery item sizes	2	2	2	2	2	2	2	2	2	2	2	2
Number of goods handling	2	2	2	2	2	2	2	2	2	2	2	2
Fast and reliable delivery method	2	2	2	2	2	2	2	2	2	2	2	2
Inventory costs	2	2	2	2	2	2	2	2	2	2	2	2
Strong IT relationship between manufacturer and customer	1	1	1	1	1	1	1	1	1	1	1	1

Legend	
2	Significantly relevant
1	Relevant
0	Neutral /not relevant

Logistics technology												
Logistics sub-process	L1	L2	L3	T1	T2	T3	T4	T5	W1	W2	W3	IT
11	0	2	2	1	0	1	1	1	1	1	1	2
12	0	0	0	1	0	1	1	2	2	1	0	1
13	2	1	1	2	0	1	0	2	0	2	1	2
14	1	1	1	0	0	0	1	0	0	0	1	0
15	0	2	2	1	0	1	0	0	1	2	1	2
M1	1	0	1	2	2	1	1	2	1	1	1	1
M2	1	1	1	2	2	0	2	1	1	2	2	2
M3	0	1	0	0	2	1	2	1	0	1	0	1
M4	1	0	0	1	2	1	1	1	2	2	1	2
M5	0	0	1	0	0	0	0	1	0	0	0	2
O1	0	0	0	1	1	2	1	2	1	2	1	1
O2	2	0	1	2	2	2	2	2	1	1	1	2
O3	2	0	0	2	2	2	2	2	0	2	1	2
O4	2	0	0	0	1	1	1	1	1	1	1	0
O5	1	0	0	0	2	1	1	1	0	2	2	1

Appendix 3

PULL System with short lead time. Example product: pizza.

Marketing logistics strategic planning

		Logistics service					
		R ₁	R ₂	R ₃	R ₄	R ₅	R ₆
		λ ₁	λ ₂	λ ₃	λ ₄	λ ₅	λ ₆
P ₁	Y ₁ Quality	1	0	2	2	2	2
	Y ₂ Physical appearance (design)	2	0	2	0	0	1
	Y ₃ Product availability	1	2	0	2	2	2
	Y ₄ Manufacturing technology	2	2	2	2	2	2
P ₂	Y ₅ Customer service	2	0	2	2	2	2
	Y ₆ Logistics resilience *	2	0	2	1	1	1
	Y ₇ Price	0	2	0	1	1	2
	Y ₈ Discounted price	0	2	0	0	0	2
P ₃	Y ₉ Product life cycle	0	0	0	0	0	0
	Y ₁₀ Number of distribution channels	2	0	1	1	2	1
	Y ₁₁ Product location	1	0	0	2	2	0
	Y ₁₂ Assortment	2	2	2	0	0	1
P ₄	Y ₁₃ Stock	1	0	0	2	2	2
	Y ₁₄ Advertisement	1	0	2	2	2	2
	Y ₁₅ Sales promotion	1	0	2	2	2	2

Legend	
2	Significantly relevant
1	Relevant
0	Neutral /not relevant

Logistics sub-process planning

		Procurement				Manufacturing				Distribution						
		I ₁	I ₂	I ₃	I ₄	M ₁	M ₂	M ₃	M ₄	O ₁	O ₂	O ₃	O ₄	O ₅		
Logistics service needs (SR)	R ₁ λ ₁ The right product	1	1	1	1	2	2	2	0	0	2	2	0	2		
	R ₂ λ ₂ In the right quantity	2	1	1	2	2	2	0	0	0	2	2	0	1		
	R ₃ λ ₃ Of the right quality	1	1	2	2	0	0	2	0	0	1	1	1	0		
	R ₄ λ ₄ At the right time	2	0	1	1	0	2	2	0	1	2	2	0	1		
	R ₅ λ ₅ In the right place	2	0	1	1	0	1	0	0	1	1	1	0	2		
	R ₆ λ ₆ For the right cost	1	2	1	0	2	0	1	1	1	2	1	0	1		
		Procurement planning	Price and cost analysis	Selection of suppliers	Track orders	Determining material requirements plans	Production schedule	Lead time	Constant production volume	Inventory level	Fast, low-cost transition	Delivery item sizes	Number of goods handling	Fast and reliable delivery method	Inventory costs	Strong IT relationship between manufacturer and customer

Logistics technology

		Loading			Transport			Storage			IT			
		L ₁	L ₂	L ₃	T ₁	T ₂	T ₃	T ₄	T ₅	W ₁	W ₂	W ₃	W ₄	P ₁
Procurement	I ₁ Procurement planning	0	2	0	1	2	1	0	1	2	2	0	2	2
	I ₂ Price and cost analysis	0	1	0	1	0	2	0	2	1	1	0	2	2
	I ₃ Selection of suppliers	1	2	1	2	1	1	0	2	1	1	1	1	2
	I ₄ Track orders	0	1	1	1	0	0	0	0	0	1	0	1	1
Manufacturing	M ₁ Production schedule	1	0	1	1	2	1	0	2	2	2	0	2	1
	M ₂ Lead time	0	0	0	1	2	1	0	2	1	2	0	1	0
	M ₃ Constant production volume	0	0	0	1	0	0	1	0	1	0	1	0	0
	M ₄ Inventory level	1	0	0	1	1	1	0	0	1	2	0	2	0
Distribution	O ₁ Fast, low-cost transition	0	0	2	0	0	0	0	1	0	0	0	2	1
	O ₂ Delivery item sizes	1	0	0	2	2	2	0	2	0	1	0	0	0
	O ₃ Number of goods handling	1	0	1	1	1	0	0	2	0	0	0	0	0
	O ₄ Fast and reliable delivery method	2	1	2	2	1	0	1	1	0	0	0	0	1
Distribution	O ₅ Inventory costs	0	0	0	0	0	1	0	1	2	2	0	1	0
	O ₆ Strong IT relationship between manufacturer and customer	0	1	1	0	0	0	0	0	1	1	1	2	2

Appendix 4

PULL System with long lead time. Example product: custom made car.

Marketing logistics strategic planning		Logistics service						
		R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	
		λ ₁	λ ₂	λ ₃	λ ₄	λ ₅	λ ₆	
Marketing mix (4P)	P ₁	Y ₁ Quality	2	0	2	0	0	1
		Y ₂ Physical appearance (design)	1	0	2	0	0	1
		Y ₃ Product availability	0	0	0	0	0	0
		Y ₄ Manufacturing technology	1	1	2	1	0	2
		Y ₅ Customer service	0	0	1	1	1	2
		Y ₆ Logistics resilience *	2	0	2	1	0	2
	P ₂	Y ₇ Price	1	0	1	1	0	2
		Y ₇ Discounted price	0	0	0	0	0	1
		Y ₈ Product life cycle	1	0	1	0	0	2
		Y ₉ Number of distribution channels	0	0	0	2	0	1
	P ₃	Y ₁₀ Product location	0	0	0	1	0	0
		Y ₁₁ Assortment	0	0	0	0	0	0
		Y ₁₂ Stock	0	0	0	0	0	0
	P ₄	Y ₁₃ Advertisement	1	0	0	0	0	1
	Y ₁₄ Sales promotion	1	0	1	0	0	1	

Legend	
2	Significantly relevant
1	Relevant
0	Neutral /not relevant

Logistics sub-process planning		Logistics service												
		Procurement				Manufacturing				Distribution				
		I ₁	I ₂	I ₃	I ₄	M ₁	M ₂	M ₃	M ₄	O ₁	O ₂	O ₃	O ₄	
Logistics service links (6R)	R ₁ λ ₁	The right product	0	0	1	0	0	1	0	0	0	1	0	2
	R ₂ λ ₂	In the right quantity	1	0	1	1	2	1	1	1	1	1	0	1
	R ₃ λ ₃	Of the right quality	0	1	2	0	1	0	0	0	0	0	0	1
	R ₄ λ ₄	At the right time	2	0	1	1	2	2	2	1	1	1	0	2
	R ₅ λ ₅	In the right place	1	0	0	0	1	0	0	1	0	0	1	2
	R ₆ λ ₆	For the right cost	2	2	1	1	1	2	0	2	2	1	1	2
	Procurement planning		1	0	0	0	0	0	0	0	0	0	0	0
	Price and cost analysis		1	0	0	0	1	1	0	0	2	1	0	1
	Selection of suppliers		1	1	0	1	2	1	0	0	0	1	1	0
	Track orders		1	1	1	2	1	0	0	0	0	0	1	0
Determining material requirements plans		0	0	0	0	0	0	0	0	2	1	2	2	
Production schedule		1	0	0	0	1	0	0	0	0	0	0	1	
Lead time		1	2	1	0	1	0	2	1	0	0	1	1	
Constant production volume		0	0	0	0	1	0	0	0	1	1	0	0	
Inventory level		0	0	0	0	1	0	0	0	2	2	2	2	
Fast, low-cost transition		0	0	0	0	0	0	0	0	0	0	0	2	
Delivery item sizes		0	1	0	1	2	1	0	2	1	2	1	0	
Number of goods handling		1	2	1	1	1	1	1	0	0	1	2	1	
Fast and reliable delivery method		1	0	0	1	1	1	1	0	0	0	0	0	
Inventory costs		0	0	0	0	0	0	0	1	2	2	1	1	
Strong IT relationship between manufacturer and customer		2	1	0	0	1	0	1	0	1	2	2	2	

Logistics technology		Logistics service												
		Loading			Transport			Storage			IT			
		L ₁	L ₂	L ₃	T ₁	T ₂	T ₃	T ₄	W ₁	W ₂	W ₃	W ₄	β ₁	
Procurement	I ₁	Number of distribution channels	1	2	1	2	1	1	0	2	1	1	0	2
	I ₂	Documentation of incoming goods	1	0	0	1	1	0	0	2	1	0	1	2
	I ₃	Quality control	1	1	0	1	2	1	0	0	0	1	1	0
	I ₄	Shipping distance	1	1	1	2	1	0	0	0	0	0	1	0
Manufacturing	M ₁	Delivery frequency	0	0	0	0	0	0	0	2	1	2	2	
	M ₂	Delivery unit	1	0	0	0	0	0	0	0	0	0	1	
	M ₃	Distance of depots	0	0	0	0	0	0	0	0	0	0	2	
	M ₄	Order item size	1	0	0	0	1	0	0	0	0	0	1	
Distribution	O ₁	Storage costs	0	0	0	0	0	0	0	0	0	0	1	
	O ₂	The size of the stock to be stored	1	2	1	1	1	0	2	1	0	1	1	
	O ₃	Warehouse administration	0	1	0	1	1	1	0	2	2	1	0	
	O ₄	Number of types of goods	1	2	1	1	1	1	1	0	0	1	2	
IT support		2	1	0	0	1	0	1	0	1	2	2	2	

Cueing of Parkinson's Disease Patients by Standard Smart Devices and Deep Learning Approach

Pavol Šatala¹, Peter Butka¹, Arsen Samaiev¹, Petra Levická²

¹ Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 04001 Košice, Slovakia; E-mails: pavol.satala@tuke.sk, peter.butka@tuke.sk, arsen.samaiev@student.tuke.sk

² Department of Neurology, Pavol Jozef Šafárik University in Košice, Faculty of Medicine, and Louis Pasteur University Hospital in Košice, Trieda SNP 1, 04011 Košice, Slovakia; E-mail: petra.levicka@student.upjs.sk

Abstract: Parkinson's disease is one of the most common neurological diseases. The patients suffer from different symptoms, e.g., tremor, bradykinesia, or walk gait disorders. One of the gait disorders which affects Parkinson's disease patients is freezing of gait, which shows as a sudden gait interruption without the ability to take the next step. It is hard to manage this symptom by medication. However, there are ways to address this symptom by applying different visual or vibration aids. This work presents a system for automatic detection and cue of freezing of gait events provided by ordinary smart devices. We used a deep learning approach to detect such events automatically. The test results and the doctors' opinions on the practical experience with the patients suggest the benefits of the provided solution.

Keywords: Parkinson's disease; deep learning; smart devices; freezing of gait

1 Introduction

Freezing of gait is a common symptom and severe complication, affecting every third patient with Parkinson's disease [1]. It is characterized clinically by sudden, relatively brief episodes of inability to produce effective forward stepping that typically occurs during gait initiation or turning while walking [2]. Parkinson's disease is currently the second most common neurodegenerative disease [3], and many patients suffer from the freezing of gait symptoms. The estimated epidemiology of Parkinson's disease shows about 1% of people older than 60 years and up to 3% of people older than 80 years [3, 4]. The freezing of gait events often leads to falls, which cause serious injuries, significantly decreasing patients' quality of life [5]. As many authors agreed, treatment of freezing of gait is a challenging task.

Currently, medicine offers various ways to treat the freezing of gait, including drugs, surgery, and physiotherapy [6]. The most common way is the usage of Levodopa, which reduces the freezing of gait events. Unfortunately, this drug may cause a severe side-effect called dyskinesia, characterized by unwanted movements of patients' body parts [7-9]. Other treatment strategies are dopamine agonist or STN stimulation. Instead of many pharmacological strategies to manage the freezing of gait phenomena, the gait problems remain persistent for patients with Parkinson's disease [10].

Medication can effectively manage some symptoms of Parkinson's disease and improve the quality of life. However, symptoms like postural instability or gait impairments still cause unfortunate limitations [11, 12]. In addition to pharmacological methods, physiotherapy has become essential in treating Parkinson's disease patients in recent years. Many complications of Parkinson's disease are due to patients' muscle weakness, postural problems, and a general decline of physical activity [13]. This problem becomes more significant as the disease progress and motion abilities are getting worse, which causes the patient to have less and less physical activity [11].

Various approaches have been developed for gait rehabilitation in recent years, including individual, group, or home-based rehabilitation [14]. However, several authors agreed that home-based therapy could bring more injuries and falls and is ineffective in improving gait or balance [15-17]. An effective way to improve patients' walking ability appears in auditory and visual cueing [18]. The cueing is based on rhythmic visual or acoustic stimuli applied continuously or on-demand [19].

It is important to apply low-cost, widely used devices to make them available to home users. Mobile smart and wearable devices can be the right choice. Such devices are equipped with three-axis accelerometers, which can be used to detect the presence of freezing of gait. Anti-freeze aids can be divided into three main groups: visual, acoustic, and vibrational (tactile). All of them can be provided by a standard smartphone in conjunction with any other wearable device, such as smart glasses, smart watches, or headphones.

The presented paper discusses the possibilities of creating a home system for on-demand freezing of gait cueing, which can be a cheap and simple-to-use alternative to more complex systems. The paper begins with an overview of the related work in Section 2. We describe the current knowledge about walking freezes and their cueing using different approaches. The following section describes an experiment performed with patients to test devices in real conditions. In Section 4, we describe model training and the system architecture. Section 5 provides an evaluation of the experiments with patients and freezing of gait detection models, followed by conclusions of the paper.

2 Related Works

2.1 Cueing of Parkinson's Disease Patients

The reason why cueing works for Parkinson's disease patients is still not fully understood. We can help patients using various types of different cues. The main idea is to provide sensory cueing based on rhythmic stimuli – acoustic, visual, or tactile. We can use simple metronome or rhythmic music for auditory stimulation. Usually, in studies, it is metronome beep in various ranges of frequencies from low frequencies with about 1 Hz [19] to high frequencies between 60 to 120 Hz [20]. Visual cues can be considered a wide range of regular patterns generated by digital devices or drawn on the floor [19, 21]. The tactile cues are rhythmic vibration impulses applied in low frequencies by wearable devices to any part of the patient's body (usually legs) [19]. In 2010, Bachlin and colleagues identified limitations with metronome cueing devices, such as those previously reported by Enzensberger et al., and Cubo et al. [38, 39]. When activated, these devices continually delivered auditory cueing regardless of whether freezing of gait was present or not. Cubo et al. identified this design limitation, reporting that patients may become habituated to the auditory stimuli, thus reducing the effect of cueing [39].

2.2 Cueing of Parkinson's Disease Patients by Wearable Devices

As Sweeney et al. show, many wearable devices for cueing freezing of gait symptoms have been built [22]. They found 4480 publications from January 2009 to December 2018, which contain Parkinson and Cue/cueing in its title. After a review process, they selected 18 publications that describe take-home systems for cueing Parkinson patients. Most of the reviewed works use continuous cueing, applied all the time. In the other cases, the authors provide the system that first detects freezing of gait events and applies on-demand cueing. The system usually uses more accelerometers placed on various parts of the user's body. One of the first works dealing with freezing of gait detection was Moore et al. [27]. The authors recorded 46 freezing of gait episodes during the experiment with eleven patients with idiopathic Parkinson's disease. Four of the patients does not show any freezing event during the experiment. The patients walk up to 100 m (based on their ability to walk). The patient's motion has been recorded by a 3-axis accelerometer placed on the patient's ankle with a frequency of 100 Hz. The work finds differences in dominant frequencies between freezing and regular gait events. The Freezing index, defined as a division of the area under the power spectra of 'freeze' band (3-8 Hz) by the square of the area under the spectra in the 'locomotor' band (0.5-3 Hz), can be then used to detect freezing of gait events. Based on the thresholds of the Freezing index, they were able to detect 78% of freezing of gaits events successfully, while the system also incorrectly labeled 20% of stand events as

freezing of gait by global threshold. Lima *et al.* in [28] bring a detailed review of existing works focusing on freezing gait detection using wearable devices. The 27 works have been selected from PubMed and Web of Science databases. Four of them focus on falls and 23 on freezing of gait. The mentioned works use various sensors differently positioned on the patient's body. Only three of the mentioned works provide a system suitable for its use at home. The following Table 1 provides their overview, with the type of sensors and evaluation of detection results. As we can see, all these approaches only detect the events and none of them provide cueing functionality.

Table 1
Review of articles focusing on freezing of gait detection in the home environment [28]

Article	Sensors	Location	Results	Cueing
Rodríguez-Martín [29]	Accelerometer	Waist	Sensitivity: 91.7% Specificity: 87.4%	NO
Ahlich [30]	Accelerometer	Waist	Sensitivity: 92.3% Specificity: 100.0%	NO
Tzallas [31]	Accelerometer Gyroscope	Wrist Shin Waist	Accuracy 79%	NO

Thanks to the development of wearable visual headsets such as HoloLens or smart glasses, it has made it possible to create visual cues using wearable devices. One example of the use of HoloLens to guide freezing of gait was presented by Geerse *et al.* [32]. It is a solution called Holocue that displays 2D or 3D holographic elements for walking, as shown in Figure 1:

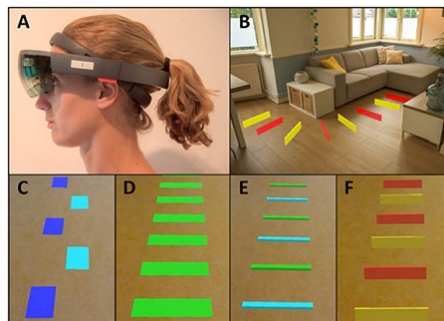


Figure 1
Holographic cues created by Holocue (source: [32])

According to [33], visual signals may be more effective than vibration or auditory. However, in [32], authors show that when wearable devices present visual stimuli, patients' immediate response generated by wearable devices worsens symptoms.

After several sessions, when patients became accustomed to wearing Holocue, the negative effect of Holocue disappeared but showed no significant improvement compared to control studies [32]. The Holocue solution offers an on-demand launch. However, it does not offer freezing of gait recognition, and a voice command triggers cueing. In addition to the mentioned solutions using wearable smart devices, even simpler solutions are already available. These simpler solutions use visual stimuli continuously, consisting of a laser lamp placed on boots or a wheelchair, as seen in Figure 2.



Figure 2

Simple solutions for visual FoG cueing (source: [35, 36])

Unfortunately, continuous application of the cues may reduce their performance during a time or cause cue addiction [19].

2.3 Machine-Learning Algorithms to Freezing of Gait Detection

Lima et al. in [28] state that most freezing of gait detection works with data from an accelerometer or inertial measurement unit (IMU) located on the ankle, knee, and belt. In addition, some work such as [37] use plantar pressure data recorded from users' shoes to detect freezing of gait. Many machine learning algorithms have been used to detect the events of freezing of gait. Random forests and Support Vector Machines (SVM) show good performance [38, 39]. Recurrent neural networks (RNNs) can perform better due to handling time-series data. Long Short Term Memory (LSTM) units have been invented to process data streams in neural networks. When training with acceleration data from three accelerometer sensors located on the knee, hip, and ankle, the network showed 83% accuracy in freezing of gait detection [40]. Most research focused on participant-dependent models [35, 39, 40]. When FoG detection is performed on its wearable device, it is severely limited by the microcontroller's performance and memory size. Then, filtering, data pre-processing, and another high-performance task affect the detection time [37]. The scope of this paper is to overcome the performance limitations of wearable devices by using server-side detection. The second point is to improve patient comfort with a solution that uses only an ordinary smartphone placed in the patient's pocket without additional sensors on the feet or inside the shoe.

3 Setup of Aids Prototypes Testing with Patients

To verify the theoretical information mentioned in the introduction and its usability with available devices, we first tested cueing using wearable and smart devices with a small group of patients with Parkinson's disease.

We created a simple mobile application to test acoustic and tactile cues. The acoustic cues were applied using wireless headphones. The application makes metronome beeps in 60, 90, and 120 Hz frequencies. The vibration of the wearable watch has applied tactile cues placed on the patient's legs. It consisted of a short interval of vibrations in frequency based on the patient's standard gait period. We tried the period of patient native gait frequency, then half and quarter of this period.

The experiment was performed on the neurological clinic of University Hospital L. Pasteur in Košice. Patients were asked to participate in our experiment during the regular control visit. Those who agreed made several walks through our test trial path. The test trial consists of two walks through narrowed space and making two U-turns, where freezing of gait usually appears. The patient stood up from the chair, walked 4 meters in the first room, then walked through the door and continued 4 meters in the second room. After that, patients make a U-turn and continue back the same way. In the end, he should make a U-turn. Figure 3 depicts the visualization of the testing trial path. The total walk time and the number of freezing of gait occurrences were observed during the test walk. The medical specialist has counted the freezing of gait events.

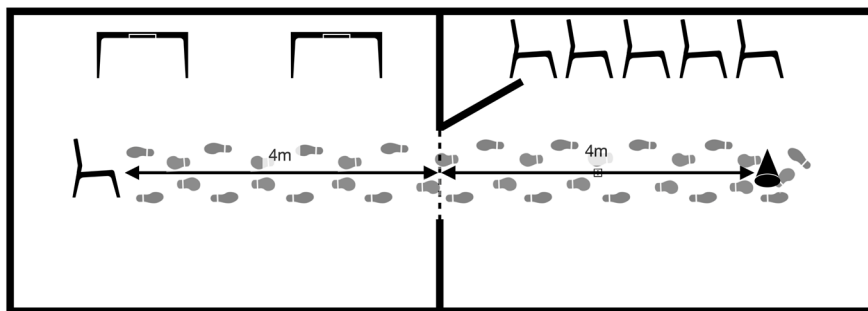


Figure 3

Visualization of the testing trial path walked by the patients

Tactile (vibration) and acoustic aids have been tested. During the test walk (first 4 meters usually without freezing of gait event), medical specialists measure the time and count the number of steps to get the patient's step period.

Then we added the smartwatch providing vibration aid onto the patient's wrist or ankle (the place with better sensitivity). We tested vibration aid with three periods - the base period of the patient's step and the half and quarter period of the steps period. The number of freezing episodes and time was measured using these aids.

We also tested acoustic aids on the same trial. As mentioned above, the acoustic aid was applied using wireless headphones, using simple beeps for selected frequencies of 60, 90, or 120 BPM. All these trials provided us with a dataset that showed practical usage of relatively cheap and usual devices (smartwatch with vibration ability, headphones) as tactile or acoustic cues for actual patients. We show the evaluation of the trial's results in Section 5.

4 Proposed System – Freezing of Gait Detection and Final System Architecture

After testing cues, we had a solid knowledge of how our tactile and acoustic aids work for the patients. We were able to design the final system architecture at this moment. Of course, we needed to work on a detection algorithm for freezing of gait (FoG) events for the final system setup. Therefore, this section describes the details on designed deep learning architectures for FoG detection, one of the proposed system modules. Then, we also added some details of the final system architecture itself.

4.1 Dataset for Detection of Freezing of Gait Events

Due to data-gathering issues through the COVID-19 pandemic with our pool of patients (testing of cues was shortly before pandemic), we decided to use the available dataset to design deep learning architecture to detect FoG events. The deep learning models trained on Parkinson's disease patients are expected to be transferable to new patients. To design and test the FoG detection algorithm, we used Daphnet Freezing of Gait Data Set [41], which contains more than 8 hours of walk data from ten patients. Eight of these patients experienced freezing of gait events during measurement. The professional physiotherapists identified 237 FOG events by video analysis. Three 3-axis accelerometers, placed on the patient's ankle, leg, and trunk, recorded data during the walking test trial in the laboratory. Data are sampled with a frequency of 64 Hz. Of course, we would like to use standard intelligent devices in our proposed system - smartphones, which are usually held in pockets. It supports our view to provide a solution as cheap as possible. It means that we can then minimize the need for placing any custom sensors. We use only data from the trunk sensor that are the same as data from smartphones in the pocket. Therefore, we will use only this part of the available dataset for learning, evaluation, and application for the deep learning architecture.

4.2 Deep Learning Architectures for Detection of Events

Contrary to most available works, we used data only from a sensor placed on the leg. This limitation may cause less accuracy of the FoG detection, but brings the

pros that users do not need to wear any additional sensors than smartphones in their pocket. For FoG events recognition, we designed and tested different deep learning architectures.

For the primary reference on the foundations of neural networks and deep learning architectures, follow Goodfellow *et al.* in [42]. The usual architecture consists of some feature extraction and classification parts, where the second one is often realized as one or more fully connected feed-forward layers. For the extraction part in time series classification tasks, the main architecture types or elements within the neural network area include:

- Convolutional layers – a type of architecture often used for features extraction in image recognition extraction, but in a 1-D setup are also used to identify features from time series. Convolution 1-D layers are suitable for identifying local changes in time series.
- Long Short-Term Memory (LSTM) – recurrent network architecture for time series sequence learning, significantly better in learning longer relations between the elements within the series. Several extensions of the approach, like bidirectional LSTM, enhance features extraction from time series windows in both directions (forward and backward in time).
- Specific layers for preprocessing features – in time series, there are often some specific operations that are useful for extracting features, e.g., Fourier transformation, which helps for the analysis of frequencies. Such processing elements can also be applicable as a layer within the architecture [43].
- Hybrid architectures – a combination of previous architecture types (elements) within the one network.

We decided to combine such elements with another architectural idea used to improve classification and scalability in our work. It means that we combine parallel blocks within the architecture (with different setups in every block), producing one concatenated feature vector, followed by a fully connected part for final classification. This combined approach helps the classifier to enhance its granularity and robustness. The main idea comes from inception models from Google for more complex networks, which were introduced in [44], but we applied the principle for smaller networks. The main advantage of such architectures is that information is processed on different scales simultaneously within parallel blocks and their features are aggregated for one feature vector, which leads to better classification results and robustness of the models. We had a good experience with such architectures in other works in the domain of astrophysics, where we used it to classify eclipsing binary stars [45] or radio galaxies [46].

In the modeling phase, we used ReLU activation function in hidden layers and softmax for the output layer. We used categorical cross-entropy as a loss function and SGD as an optimizer. We compared four architectures to find the best for our purposes:

- **CNN model** – contains three parallel CNN (Convolutional Neural Network) blocks of layers with an input batch of 64 samples, each with three values. The CNN layers differ in the number of filters (20, 40, or 50). The output has two neurons with a soft-max activation function.
- **CNN + LSTM model** – has two parallel flows. One with CNN layer and the second with bidirectional Long Short-Term Memory layer.
- **2xCNN + LSTM model** – has three parallel flows. One block contains a bidirectional Long Short-Term Memory layer, and the second and third blocks contain the CNN layers. Same as in the first model, CNN-based blocks differ in the number of filters in layers.
- **Fourier + CNN** – adds to CNN model Fourier transformation as a pre-processing layer. Moore et al. in [27] found the importance of Fourier transformation in the detection based on frequency analysis.

4.3 Final System Architecture

It is essential to combine all the aspects mentioned above for practical application. Therefore, we proposed a cloud-based recognition system that connects acoustic or tactile aids. The cloud-based system enables the online model upgrade. The system consists of a mobile device as a sensor and an actuator for aids. As the first step, it is needed to label test data for a user, and then the system is able to work by itself.

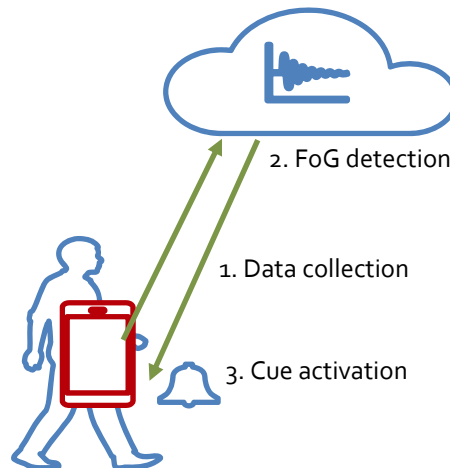


Figure 4

High-level architecture of the proposed system

Thanks to cloud-based messaging, the system may change the recognition system without changing any client application running on the user device. Figure 4 depicts the high-level architecture of the system.

The solution can be divided into two parts: server and client application. The client application runs on a smart device, which can be easily held in the patient's pocket. Because the recognition algorithm was developed for 64 Hz data, the mobile device must be compatible with the need for an acceleration sensor with a frequency of at least 64 Hz. According to the Android 12 Compatibility Definition document [47], it is strongly recommended that Android phones include an accelerometer sensor, which (if included) must provide data with a frequency of at least 100 Hz. The second requirement for the device is a functional internet connection. When no internet connection is available, the application does not offer a recognition feature.

The task of the client application is to record data from the accelerometer at the required frequency (for the presented models, it is 64 Hz). The recorded data are aggregated and sent as one batch (size N) to the server. FoG cueing is activated depending on the server response. After each $N / 2$ measurement, the batches are sent to improve performance and reduce response time. It means that for a batch of 64 samples (1 second at 64 Hz), we send data to the server every 32 new samples. One batch consists of $N / 2$ rows of old data (data used in the previous recognition), followed by $N / 2$ rows of new data. This halves the recognition time delay.

The server application uses a cloud platform as a service (PaaS) solution from a local provider running the Ubuntu operating system. The application is written in NodeJS and offers a Rest API. The Rest API offers a POST method at `[host]/fog`. The method accepts a json file containing the person's id and a recorded data, as the following code example shows:

```
{
  "person":12,
  "data": [
    [202,1803,386...],
    [292,1598,-49...],
    .....
  ]
}
```

The server provides a simple answer, containing a flag if the batch is FoG or not:

```
{
  "fog": true
}
```

After the API method is called, the server receives the data and evaluates the model. The machine learning operations run using the TensorFlow JavaScript machine learning solution. According to the person ID sent in the actual request, the correct TensorFlowJS model is loaded and evaluated using predict function applied to received values.

The solution with presented settings (64 Hz frequency and batch of 64 samples) creates requests up to 1.2 KB. The application makes two calls in one second. It creates a maximum of 2.4 KB (19.2 Kb) network traffic per second. This requirement can be met with huge reserve by the current mobile networks. If the number of clients connected to the server raises the required performance of the server application, it could be easily scaled thanks to using a SaaS cloud service.

5 Results and Evaluation

The success of the presented solutions depends on two of its abilities. The first is to provide cueing that effectively reduces the number of freezing of gait (FoG) episodes, shortens FoG time, or helps overcome it. The second is the use of a machine learning algorithm that is efficient enough to recognize an FoG episode and provide on-demand guidance.

5.1 Evaluation of Wearable Aids Testing

Eleven patients – four women and seven men with an average age of 66.44 years took part in the experiment. The average duration of patients' disease has been eight years. The acoustic and tactile cueing has been tested. Due to the motion limitation and fatigue, patients usually do not finish all tests. Two patients show no FoG events during the test walk. We excluded these patients from the evaluation.

We observed improvement in total walk time and the number of freezing of gait events during the test. The average total time of walk without any cueing was $42,43 \pm 19,43$ seconds, and the average number of FoG occurrences was $1,95 \pm 1,27$. As we mentioned before, the acoustic cueing was tested with three frequencies - 60, 90, and 120 BPM. There was a decrease in the number of FoG events and total walk time in all tested frequencies, leading to smoother and safer walks of the patients. The average reduction of occurrences of FoG events was about 60% in all three tested frequencies. However, the frequency of 120 BPM shows the best results in walking speeds (23,03%). Table 2 shows an overview of the results of acoustic aids.

Similarly to acoustic aids, we tested vibration aids on three different frequencies. The basic level is the patient's base step frequency (p). Then, the tested frequencies are p , $p/2$, and $p/4$. The number of tested subjects was lower than in acoustic aids testing. The main reason is that some patients had problems with sensitivity. We also issued problems with their fatigue after the previous testing. Table 3 depicts the results of the testing. We achieved the best results with the frequency of $p/4$, but only two people took part in the test with this setup. These people were patients in good conditions who were able to take part in all tests. The good health conditions of patients may affect the results. However, generally, we can see that vibration aids can significantly decrease the number of FoG events by at least 80%.

We can see different numbers of tested subjects for various frequencies. During experiments, not all patients could finish all experiments due to their physical abilities – fatigue or other complications. In these cases, we had to end measurements earlier. That caused the different number of patients in columns in Table 2 and Table 3. Average walk time is calculated from times for users who took part in the experiment with the required frequency. Average walk time with aids can be calculated as "Average walk time T" plus "Average decrease of T".

Table 2
Results of acoustic aids testing

	60 BPM	90 BPM	120 BPM
Number of tested subjects	5	8	6
Average walk time – T (seconds)	28,03	36,78	25,26
Average decrease of T (seconds)	2,51	5,63	7,35
Average decrease of T (%)	9,30	11,25	23,03
Average decrease of FoG (occurrences)	1,60	1,56	2,25
Average decrease of FoG (%)	64,28	62,50	58,33

Table 3
Results of vibration cueing

	p	p/2	p/4
Number of tested subjects	5	5	2
Average walk time – T (seconds)	34,12	35,21	26,23
Average decrease of T (seconds)	2,99	1,86	2,76
Average decrease of T (%)	5,88	2,45	11,09
Average decrease of FoG (occurrences)	1,4	1,4	1,5
Average decrease of FoG (%)	80	80	100

Due to the results of experiments, and according to theoretical background known from the literature, the designed acoustic and vibration aids show their usability for the proposed system.

5.2 Evaluation of FoG Recognition Algorithm

To evaluate FoG recognition models based on deep learning architectures, we used the following metrics:

- **Accuracy** = $(TP+TN)/(TP+FP+FN+TN)$. It's the ratio of the correctly labeled subjects to the whole pool of subjects.
- **Recall** = $TP / (TP+FN)$. Recall (sensitivity) is the ratio of the correctly positively labeled subjects by our models to all truly positive subjects.
- **Precision** = $TP / (TP + FP)$. Precision is the ratio of the correctly positive labelled subjects by our models to all positive labelled subjects.
- **F1 score** = $2 * (Precision * Recall) / (Precision + Recall)$. It is the harmonic mean of the precision and recall.

where:

- **TP – True Positive** - event marked as FoG is really FoG,
- **TN – True Negative** - event correctly marked as normal gait,
- **FP – False Positive** – event marked as FoG, but is normal gait in reality,
- **FN – False Negative** – event is predicted to be normal gait but is FoG in reality.

The accuracy metric may be misleading for the unbalanced dataset. The freezing of gait (FoG) appears as a relatively short episode, interrupting the long duration of normal gait. Due to it, FoG datasets are strongly unbalanced. For practical application, the most crucial metric is recall. We considered as important to mark data batch as FoG also when it sometimes is marked as a false positive. In addition, we also considered precision and F1 score to take into account. The results showed significant differences between patients. Table 4 depicts personalized results for five selected patients from the dataset (who offer enough records).

Table 4

Personalized metrics of FoG class detection for every type of model. Bold marked numbers are the best results for five patients (S01-05). The results are considered for every batch of data separately.

3xCNN	recall	precision	F1
S01	38%	62%	47%
S02	62%	88%	73%
S03	50%	72%	59%
S05	39%	48%	43%
S07	7%	27%	11%

CNN + LSTM	recall	precision	F1
S01	36%	63%	46%
S02	63%	88%	73%
S03	51%	73%	60%
S05	21%	46%	29%
S07	5%	40%	8%
2xCNN+LSTM	recall	precision	F1
S01	51%	52%	52%
S02	62%	90%	73%
S03	61%	71%	66%
S05	30%	46%	37%
S07	12%	62%	20%
Fourier + CNN	recall	precision	F1
S01	72%	40%	52%
S02	44%	96%	60%
S03	56%	68%	61%
S05	57%	50%	53%
S07	21%	32%	25%

If we choose the same model architecture for all users, the average recall will be about 50%. However, if we use a personalized architecture and model with the best results for each user, we may achieve an average sensitivity of 55%. One of the problems with learning models and evaluation of results per batch is the small proportion of FoG events for patients in their data stream from the sensor. It varies from 5%-24%, where worse results are for the more imbalanced dataset because 5% means that only a small amount of batches had detectable events. It is especially a problem for S07 patient's data. For other patients, the average recall is better. While our solution shows lower sensitivity than other works, the deeper analysis showed some promising details:

- Again, it is essential to mention that we use only one sensor compared to the three used in most other works. Therefore, we provide a cheaper alternative for detection combined with an eventual type of aid (acoustic or vibration-based).

- For more balanced streams of data in the learning process, from patients with more than 15% and/or higher number of events in general, average recall of their personalized models for FoG batch detection is at least 65%.
- In a practical setup, FoG events in batches are not uniformly distributed but combined within one longer event (episode – more batches during some time). It means that there are more FoG batches in real life near each other, and with an average of 65% recall, it is possible to detect at least one FoG batch during the actual FoG event of a patient.

Especially the last observation has an important impact on practical implementation, when the achieved sensitivity for particular batches may be sufficient for the final usability of the system. It is important to note that the algorithm evaluates many batches during a single real-life FoG event (e.g., 20 batches for a 10 s FoG event). Figure 5 shows an example of five freezing events with a number of batches during every FoG event. One rectangle illustrates one batch as a one-second interval. The intervals overlap by 500 ms, increasing the number of batches and decreasing detection delay.

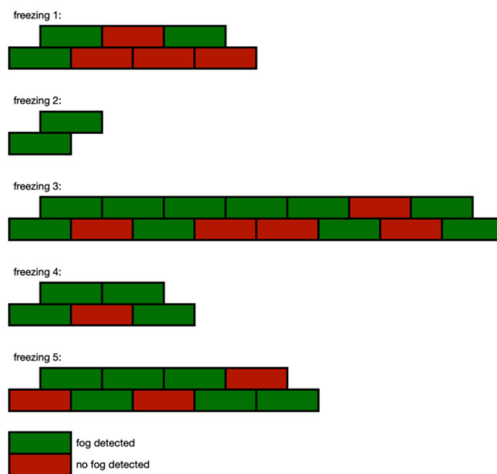


Figure 5

Test API calls for cueing with sample data from one of the patients for his FoG episodes. All rectangles are data batches recognized as FoG events by annotators. For every freezing event (even for their different length), our algorithm has several correct FoG detections, and API will start cueing the patient successfully.

In practice, we need to detect at least one of the first few batches during the FoG event to help overcome it for a workable solution. To try the performance in real conditions, we tested the system by calling created API function with sample data from the dataset. We found out, that in all of the FoG events, more than a half of batches in the event have been marked as a FoG. The visualization for one of the patient and his five freezing episodes is available in Figure 5. Thanks to detection

of at least some of the batches from the start of the episode as FoG, our system will start cueing the patient after the first detection. Moreover, for most freezing episodes, the first detection is in one of the first batches, usually less than 1-2 seconds. Therefore, the presented solution will help the patient during the FoG episode with the current technological setup and detection algorithm while providing a cheaper solution for the system's sensor and cueing parts.

Conclusions

We proposed the system combining smart and wearable devices for detection of freezing of gait events of Parkinson's disease patients and their cueing using simple acoustic or vibration aids. Our experiments have shown that we can significantly improve walkability for patients suffering from the freezing of gait symptoms and improve their quality of life. The proposed system is one of the solutions to make a practical system based on cheap devices, which took part of life for almost everyone. Our detection algorithm has some limits, but it can still detect freezing of gait episodes soon after starting with only one sensor in mobile devices. It is the more convenient solution for patients because they only need their mobile devices. The medical assistant, who helped us with the experiments with patients during the experiment, also provided us with his feedback on the potential of the proposed solution: "...in comparison of acoustic and vibration aids, acoustic aids were more convenient for application and more convenient for patients. In the future, I would welcome the opportunity to install the application on a mobile phone for every patient with a stronger freezing of gait to improve their quality of life."

Acknowledgement

This work was supported by Slovak VEGA research grant No. 1/0685/21 and Slovak APVV research grant under contract No. APVV-17-0550. We would also like to thank medical specialists and patients from University Hospital L. Pasteur in Košice.

References

- [1] Macht, M., Kaussner, Y., Möller, J. C., et al.: Predictors of freezing in Parkinson's disease: a survey of 6,620 patients, *Movement disorders* 22(7), 2007, pp. 953-956
- [2] Nutt, J. G., Bloem, B. R., Giladi, N., et al.: Freezing of gait: moving forward on a mysterious clinical phenomenon, *The Lancet Neurology* 10(8), 2011, pp. 734-44
- [3] Lee, A., Gilbert, R. M.: Epidemiology of Parkinson disease, *Neurologic clinics* 34(4), 2016, pp. 955-965
- [4] De Lau, L. M., Breteler, M. M.: Epidemiology of Parkinson's disease, *The Lancet Neurology* 5(6), 2006, pp. 525-535

- [5] Gao, C., Liu, J., Tan, Y., Chen, S.: Freezing of gait in Parkinson's disease: pathophysiology, risk factors and treatments, *Translational neurodegeneration* 9(12), 2020, pp. 1-22
- [6] Nonnekes, J., Snijders, A. H., Nutt, J. G., et al.: Freezing of gait: a practical approach to management, *The Lancet Neurology* 14(7), 2015, pp. 768-778
- [7] Schaafsma, J., Balash, Y., Gurevich, T., et al.: Characterization of Freezing of gait subtypes and the response of each to Levodopa in Parkinson's disease, *European journal of neurology* 10(4), 2003, pp. 391-398
- [8] Fahn, S.: Does Levodopa slow or hasten the rate of progression of Parkinson's disease?, *Journal of neurology* 252(Suppl 4), 2005, pp. iv37-iv42
- [9] Thanvi, B., Lo, N., Robinson, T.: Levodopa-induced dyskinesia in Parkinson's disease: clinical features, pathogenesis, prevention and treatment, *Postgraduate medical journal* 83(980), 2007, pp. 384-388
- [10] Spaulding, S. J., Barber, B., Colby, M., et al.: Cueing and gait improvement among people with Parkinson's disease: a meta-analysis, *Archives of physical medicine and rehabilitation* 94(3), 2013, pp. 562-570
- [11] Chester, E. L., Turnbull, G. I., Kozey, J.: The effect of auditory cues on gait at different stages of parkinson's disease and during "on /" off" fluctuations: a preliminary study, *Topics in Geriatric Rehabilitation* 22(2), 2006, pp. 187-195
- [12] McNeely, M. E., Duncan, R. P., Earhart, G. M.: Medication improves balance and complex gait performance in Parkinson disease, *Gait & posture* 36(1), 2012, pp. 144-148
- [13] Stelmach, G. E., Teasdale, N., Phillips, J., Worringham, C. J.: Force production characteristics in Parkinson's disease, *Experimental Brain Research* 76(1), 1989, pp. 165-172
- [14] King, L. A., Wilhelm, J., Chen, Y., et al.: Effects of group, individual, and home exercise in persons with Parkinson disease: a randomized clinical trial, *Journal of Neurologic Physical Therapy* 39(4), 2015, pp. 204-212
- [15] Schenkman, M., Hall, D. A., Barón, A. E., et al.: Exercise for people in early- or mid-stage Parkinson disease: a 16-month randomized controlled trial, *Physical therapy* 92(11), 2012, pp. 1395-1410
- [16] Canning, C. G., Allen, N. E., Dean, C. M., Goh, L., Fung, V. S.: Home-based tread-mill training for individuals with Parkinson's disease: a randomized controlled pilot trial, *Clinical rehabilitation* 26(9), 2016, pp. 817-826
- [17] Quinn, L., Busse, M., Khalil, H., et al.: Client and therapist views on exercise programmes for early-mid stage Parkinson's disease and Huntington's disease, *Disability and rehabilitation* 32(11), 2010, pp. 917-928

- [18] Picelli, A., Camin, M., Tinazzi, M., et al.: Three-dimensional motion analysis of the effects of auditory cueing on gait pattern in patients with Parkinson's disease: a preliminary investigation, *Neurological Sciences* 31(4), 2010, pp. 423-430
- [19] Velik, R.: Effect of on-demand cueing on freezing of gait in Parkinson's patients, *International Journal of Biomedical Engineering* 6, 2013, pp. 280-285
- [20] De Icco, R., Tassorelli, C., Berra, E., et al.: Acute and chronic effect of acoustic and visual cues on gait training in Parkinson's disease: a randomized, controlled study, *Parkinson's Disease* 2015, article id 978590, 2015, pp. 1-9
- [21] Janssen, S., Bolte, B., Nonnekes, J., et al.: Usability of three-dimensional augmented visual cues delivered by smart glasses on (freezing of) gait in Parkinson's disease, *Frontiers in neurology* 8:279, 2017, pp. 1-10
- [22] Sweeney, D., Quinlan, L. R., Browne, P., et al.: A technological review of wearable cueing devices addressing freezing of gait in Parkinson's disease, *Sensors* 19(6):1277, 2019, pp. 1-35
- [23] Bächlin, M., Plotnik, M., Roggen, D., et al.: A wearable system to assist walking of Parkinson's disease patients, *Methods of information in medicine* 49(1), 2010, pp. 88-95
- [24] Samà, A., Pérez-López, C., Rodríguez-Martín, D., et al.: A double closed loop to enhance the quality of life of Parkinson's Disease patients: REMPARK system, *Studies in health technology and informatics* 207, 2015, pp. 115-124
- [25] Mazilu, S., Blanke, U., Dorfman, M., et al.: A wearable assistant for gait training for Parkinson's disease with freezing of gait in out-of-the-lab environments. *ACM Transactions on Interactive Intelligent Systems*, 5(1), article no. 5, 2015, pp. 1-31
- [26] Ahn, D., Chung, H., Lee, H. W., et al.: Smart gait-aid glasses for Parkinson's disease patients, *IEEE Transactions on Biomedical Engineering* 64(10), 2017, pp. 2394-2402
- [27] Moore, S., MacDougall, H., Ondo, W.: Ambulatory monitoring of freezing of gait in Parkinson's disease, *Journal of Neuroscience Methods* 167, 2008, pp. 340-348
- [28] De Lima, A. L. S., Evers, L. J., Hahn, T., et al.: Freezing of gait and fall detection in Parkinson's disease using wearable sensors: a systematic review, *Journal of neurology* 264(8), 2017, pp. 1642-1654
- [29] Rodríguez-Martín, D., Samà, A., Pérez-López, C., et al.: Comparison of Features, Window Sizes and Classifiers in Detecting Freezing of Gait in

- Patients with Parkinson's Disease through a Waist-Worn Accelerometer, *Frontiers in Artificial Intelligence and Applications* 288, 2016, pp. 127-136
- [30] Ahlrichs, C., Samà, A., Lawo, M., et al.: Detecting freezing of gait with a tri-axial accelerometer in Parkinson's disease patients, *Medical & biological engineering & computing* 54(1), 2016, pp. 223-233
- [31] Tzallas, A. T., Tsipouras, M. G., Rigas, G., et al.: PERFORM: a system for monitoring, assessment and management of patients with Parkinson's disease, *Sensors* 14(11), 2014, pp. 21329-21357
- [32] Geerse, D. J., Coolen, B., Hilten, J. J., Roerdink, M.: Hologue a Wearable Holographic Cueing Application for Alleviating Freezing of Gait in Parkinson's Disease, *Front. Neurol.*, 2022
- [33] Cole, B. T., Roy, S. H., Nawab, S. H.: Detecting freezing-of-gait during unscripted and unconstrained activity, *Proceedings of 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2011, pp. 5649-5652
- [34] Rezvanian, S., Lockhart, T. E.: Towards real-time detection of freezing of gait using wavelet transform on wireless accelerometer data, *Sensors* 16(40):475, 2016, pp. 1-11
- [35] Donovan, S., Lim, C., Diaz, N., et al.: Laserlight cues for gait freezing in Parkinson's disease: An open-label study. *Parkinsonism & Related Disorders* 17, 2011, pp. 240-245
- [36] Roberts, J.: Shoe-mounted laser to 'unfreeze' people with Parkinson's scoops €1 million prize, *Horizon The EU Research & Innovation Magazine*, 2019 (online: <https://ec.europa.eu/research-and-innovation/en/horizon-magazine/shoe-mounted-laser-unfreeze-people-parkinsons-scoops-eu1-million-prize>, available: 12.9.2022)
- [37] Shalin, G., Pardoel, S., Lemaire, E. D. et al.: Prediction and detection of freezing of gait in Parkinson's disease from plantar pressure data using long short-term memory neural-networks. *J NeuroEngineering Rehabil* 18, 2021
- [38] Samà A., Rodríguez-Martín D., Pérez-López C., et al.: Determining the optimal features in freezing of gait detection through a single waist accelerometer in home environments. *Pattern Recognit Lett.* 2018; pp. 135-143
- [39] Mazilu S., Hardegger M, Zhu Z., et al.: Online detection of freezing of gait with smartphones and machine learning techniques: 6th International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health) and Workshops 2012, pp. 123-130
- [40] Ashour A. S., El-Attar A., Dey N., Abd El-Kader H., Abd El-Naby M. M.: Long short-term memory based patient-dependent model for FOG detection in Parkinson's disease. *Pattern Recognit Lett.* 2020; pp. 23-29

- [41] Bachlin, M., Plotnik, M., Roggen, D., et al.: Wearable assistant for Parkinson's disease patients with the freezing of gait symptom, *IEEE Transactions on Information Technology in Biomedicine* 14(2), 2009, pp. 436-446
- [42] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, 2016
- [43] Pratt, H., Williams, B., Coenen, F., Zheng, Y.: FCNN: Fourier Convolutional Neural Networks, *Lecture Notes in Computer Science* 10534, Machine Learning and Knowledge Discovery in Databases, 2017, pp. 786-798
- [44] Szegedy, C., et al.: Going deeper with convolution, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9
- [45] Cokina, M., Maslej-Kresnakova, V., Butka, P., Parimucha, S.: Automatic classification of eclipsing binary stars using deep learning methods, *Astronomy and Computing* 36, art. no. 100488, 2021, pp. 1-12
- [46] Maslej-Kresnakova, V., El Bouchefry, K., Butka, P.: Morphological classification of compact and extended radio galaxies using convolutional neural networks and data augmentation techniques, *Monthly Notices of the Royal Astronomical Society* 505(1), 2021, pp. 1464-1475
- [47] Google LLC: Android 12 Compactibility Definition, 2021, pp. 92-93

Tuning Parameter-free Model Predictive Control with Nonlinear Internal Model Control Structure for Vehicle Lateral Control

Adorjan Kovacs* and Istvan Vajk

Department of Automation and Applied Informatics, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary; vajk@aut.bme.hu

*Corresponding author: adorjan.kovacs@aut.bme.hu

Abstract: The paper presents a new methodology for the lateral control of autonomous vehicles. The proposed cascade structure realizes two main components: a model predictive control (MPC)-based outer loop and an internal model control (IMC) based inner loop. In the outer loop, a unique model predictive control is introduced that eliminates the tuning parameters of the system by introducing a hierarchical optimization system. Each cost function of the hierarchical optimization focuses on minimizing a physical phenomenon. The inner loop handles system dynamics and nonlinearities, providing a robust system against external disturbances and parameter changes. After presenting the proposed structure, proper comparisons were performed: firstly, to see the advantages of the tuning parameter-free method and secondly, to highlight the benefits of the IMC-based method. Finally, the whole system is compared to a reference controller, available in MatLab.

Keywords: model predictive control; internal model control; minimal tuning parameters; vehicle lateral control

1 Introduction

The lateral control problem of vehicles is one of the main emphasized questions of autonomous vehicles. One of the biggest goals of autonomous vehicles is to increase safety on roads. That means the vehicle control methods should be prepared to handle unexpected or unmeasured effects such as external disturbances, parameter changes, and suddenly changing environments. Therefore, the nonlinearities and the dynamics of the vehicle should be considered, giving a solution that is universal under all circumstances. It is supposed that the path is given as a reference, e.g., calculated by a receding horizon control algorithm [1] or a dynamic optimal control problem [2], so the path planning part of the problem is not included in this paper.

Several solutions handle the nonlinearities, parameter uncertainties, and dynamics of the vehicle, even using machine learning or fuzzy techniques [3, 4]. Adaptive control methods manage model imperfections using the multivariable fixed point iteration method [5]. An iterative feedback tuning controller can handle the strong nonlinearities of systems [6]. Integral backstepping control realizes a feedback control rule for the problem [7], ensuring stability based on the Lyapunov theory. Two parameters are weighting the lateral position and the orientation errors for the feedback loop that should be tuned for the controller. The feedback linearization method leads to a chained system that can be handled by a linear matrix inequalities problem using the peak-to-peak performance approach [8]. However, this method includes a trial-and-error-based parameter tuning method, which provides knowledge of the behavior of the system only in the tested cases. Another feedback solution, the potential-field-based method, was introduced in [9], but only a proportional-derivative controller is tuned for the feedback control. This method also lacks the usage of existing knowledge of the model. The flatness-based method deals with the dynamics of the vehicle, but only linear tire models are included [10, 11]. The proper knowledge of the nominal model parameters is crucial in this method, and a tuning process should also be performed on the gains. The main disadvantage of the different feedback-based methods [8] is that system noises can result in unnecessary control actions compared to methods that consider future references.

The feedforward-feedback method was introduced in [13], which is a virtual potential field-based solution. However, the steering control signal is determined from three independent signals (yaw damping, lane-keeping, and feedforward branch), which are hard to handle if the system reaches its rate limit or final value limit. A path planning and tracking algorithm realizes both feedforward and feedback parts of the control, but separately. The physical limitations are handled by the curve-based feedback loop [14].

The model predictive controller (MPC) is a method that integrates the feedforward and feedback loops into one system. This method can determine the control signal based on optimization, using the existing knowledge about the system: its model with accurate complexity (including nonlinearities and dynamics) and its parameters [15]. The methods presented in [2] and [16] use linearization around the prescribed nominal trajectory to gain a real-time solvable problem. However, this method operates with more than a dozen of parameters, and the control structure excludes direct feedback, so the reaction of the controller to sudden disturbances or changes could be improved. The parameter changes can be handled with an adaptive MPC [17], but the adaptation rule presupposes that specific parameters are measured. The MPC method can be formulated based on the input-output variables and internal states such as yaw-rate [18]. The advantage of the second approach is that the behavior of the vehicle can be controlled directly, concerning the states, despite the indirect methods.

The main contributions of this paper include a methodology that can eliminate the intuitively tuned parameters from MPC controllers, providing non-specific control rules. The optimization is performed based on physical phenomena instead of summing up different expressions with weighting coefficients in the cost function. The proposed algorithm includes a novel hierarchization method that ensures feasibility for the parameter-free approach. This method is placed in a cascade control structure [19, 20], formulating the outer loop. The inner loop handles the dynamics and the nonlinearities of the system, providing robustness against external disturbances and parameter changes. The proposed control approach is compared with the classical MPC methods to see its advantages. The parameter-free method is examined in the simulation, comparing the outer loop. Then, the inner loop is compared to see the performance of the IMC structure. The whole structure was compared to the lane-keeping assist (LKA) reference controller available in MATLAB.

In the following, in Section 2, the system modeling approaches are detailed. After, Section 3 introduces the solution of the proposed structure for dynamics handling together with the custom solution. The classical and the proposed model predictive approaches are detailed in Section 4. The controllers developed for the comparison-based evaluation are introduced in Section 5. The proposed and the reference controllers are compared and evaluated in simulation, and the results are written in Section 6. Finally, the conclusions are gathered in Section 7.

2 Modeling Considerations for Lateral Vehicle Control

In this section, the modeling considerations are presented. The proposed controller uses the kinematic model in the outer and the dynamic model in the inner loop. The simulation framework uses the nonlinear dynamic bicycle model.

2.1 Kinematic Model and the Frenet Frame

The kinematic unicycle model can determine the planar behavior of the vehicle. This model is described in the Frenet frame, as can be seen in Fig. 1. This frame defines the states of the vehicle in a path-based coordinate system with three parameters: the distance from the reference path (d), the orientation compared to the orientation of the reference path (ψ_p), and the distance taken along the reference path (s). The kinematic behavior of the vehicle can be described by a nonlinear state equation system [21]:

$$\dot{s} = \frac{\cos\psi_p}{1 - dC(s)} u_x, \quad \dot{d} = u_x \sin\psi_p, \quad \dot{\psi}_p = r - C(s) \frac{\cos\psi_p}{1 - dC(s)} u_x, \quad (1)$$

where u_x is the longitudinal speed, r is the yaw-rate ($r = \dot{\psi}$, where ψ is the orientation of the vehicle), and $C(s)$ is the curvature of the path.

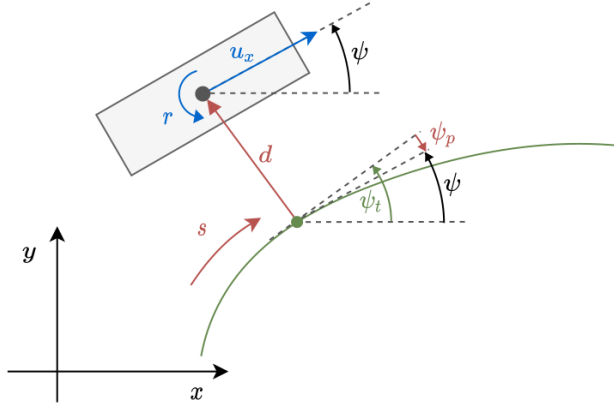


Figure 1

The unicycle in the Frenet frame

It is assumed that the vehicle goes with constant velocity to make the latter controller comparison methods clearer by omitting the longitudinal dynamics. Also, the small-angle assumptions and the first-order Taylor-series approximation can be used on this model [20]. The derivative of the yaw-rate $\dot{r} = \ddot{\psi} = \rho$ is chosen to be the control signal since it describes the control effort performed during the maneuver. By using ρ , the linear state equation can be derived for the Frenet frame model:

$$\begin{bmatrix} \dot{r} \\ \dot{d} \\ \dot{\psi}_p \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & u_x \\ 1 & -C(s)^2 u_x & 0 \end{bmatrix} \begin{bmatrix} r \\ d \\ \psi_p \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \rho + \begin{bmatrix} 0 \\ 0 \\ -C(s) u_x \end{bmatrix}. \quad (2)$$

This linear equation system can be used for state prediction. The future state values of the model can be determined by using a pre-known input vector and the initial state values [20].

2.2 Dynamic Bicycle Model

The dynamic nonlinear bicycle model is needed to have proper knowledge of the behavior of the vehicle. This model can be seen in Fig. 2. The state equations of this model can be derived by writing up the forces and moments balance on the center of the gravity (COG):

$$\dot{u}_y = \frac{F_{yF} \cos(\delta) + F_{yR}}{m} - r u_x, \quad \dot{r} = \frac{a F_{yF} \cos(\delta) - b F_{yR}}{I_z}, \quad (3)$$

where u_y is the lateral speed, F_{yi} , $i=\{F, R\}$ is the lateral tire force for the front (F), and the rear (R) wheels, m is the mass of the vehicle, I_z is the inertia around axis z , a is the distance between the front axle and the COG, b is the distance between the rear axle and the COG, and δ is the road wheel angle.

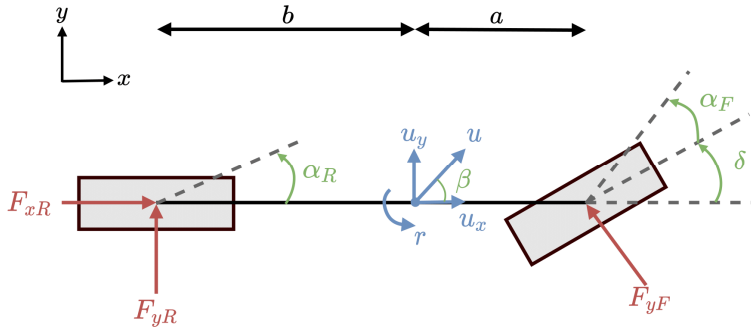


Figure 2

The bicycle model of the vehicle

The nonlinear tire model is created by a two-parameter approximation [22] of the Pacejka tire model [23] to determine the lateral tire forces:

$$F_{yt} = F_{zt} \mu \cdot \sin(c_{lat} \arctan(-b_{lat} \alpha_i)), \quad i \in [F, R], \quad (4)$$

where c_{lat} is the shape factor, and b_{lat} is the stiffness factor, α_i is the tire sideslip, and F_{zi} is the vertical force, calculated from the geometry of the model:

$$F_{zF} = mg \frac{b}{a+b}, \quad F_{zR} = mg \frac{a}{a+b}. \quad (5)$$

The tire sideslips can be determined based on geometrical considerations:

$$\alpha_F = \arctan \frac{u_y + ar}{u_x}, \quad \alpha_R = \arctan \frac{u_y - br}{u_x}. \quad (6)$$

It is common to handle the vehicle with the linearized model, which can be gained by the small angle assumptions and substituting the tire model to a linear one. This linear tire model can be described by only one parameter, the cornering stiffness $F_{yi} = C_i F_{zi}$, $i=\{F, R\}$ [17]. This way, the linearized dynamic model can be derived from Eq. 3:

$$\begin{aligned} \dot{u}_y &= -\frac{C_F + C_R}{mu_x} u_y - \left(u_x + \frac{aC_F - bC_R}{mu_x} \right) r + \frac{C_F}{m} \delta \\ \dot{r} &= -\frac{aC_F - bC_R}{I_z u_x} u_y - \frac{a^2 C_F + b^2 C_R}{I_z u_x} r + \frac{aC_F}{I_z} \delta. \end{aligned} \quad (7)$$

The nominal values of the vehicle model used in the simulation can be found in Table 1, together with the parameter names and units.

3 Handling Model Dynamics

Two ways of handling the dynamics of the system are presented in this paper. The widespread method among model predictive controls is that the model with its dynamics is included in the prediction. It means that the linearized dynamic model should have the problem manageable with the convex optimization methods, or global nonlinear solvers should be included to handle the model without linearization. The other way is that the dynamics are not included in the prediction, only the kinematics, creating a cascade structure. In this structure, there is an outer loop for handling the predictive control with a kinematic approach (considering limitations derived from the system dynamics) and an inner loop driving the dynamics of the system.

Table 1
Parameters of the vehicle

Symbol	Name	Value
m	Vehicle mass	1523 kg
I_z	Inertia around z-axis	2330 kgm ²
a	Distance between COG and front axle	1.5 m
b	Distance between COG and rear axle	1.2 m
c_{lat}	Shape factor	1.472
b_{lat}	Stiffness factor	10.87
$ \delta _{max}$	Maximum of road wheel angle	1.05 rad
$ d\delta/dt _{max}$	Maximum steepness of road wheel angle	1.35 rad/sec

3.1 Linearized Dynamic Model

The first presented method handles the dynamics by linearization. The linearized dynamic model can be written up in the Frenet frame, so the steering wheel angle can be determined directly from this model using an MPC formulation. The state equations are gained from the linearized Frenet frame (Eq. 2) and linearized dynamic model of Eq. 7 using the state vector $x^d = [d, \psi_p, u_y, r]^T$:

$$\begin{bmatrix} \dot{d} \\ \dot{\psi}_p \\ \dot{u}_y \\ \dot{r} \end{bmatrix} = \begin{bmatrix} 0 & u_x & 0 & 0 \\ c_k^2 u_x & 0 & 0 & 1 \\ 0 & 0 & -\frac{C_F + C_R}{m u_x} & -u_x \frac{\alpha C_F - b C_R}{m u_x} \\ 0 & 0 & -\frac{\alpha C_F - b C_R}{L_\pi u_x} & -\frac{\alpha^2 C_F + b^2 C_R}{L_\pi u_x} \end{bmatrix} \begin{bmatrix} d \\ \psi_p \\ u_y \\ r \end{bmatrix} + \begin{bmatrix} 0 \\ C_F \\ m \\ \alpha C_F \\ L_\pi \end{bmatrix} \delta + \begin{bmatrix} 0 \\ -C_R u_x \\ 0 \\ 0 \end{bmatrix}. \quad (8)$$

The control structure of the MPC using the linearized dynamic model can be seen in Fig. 3. The localization block is responsible for determining the Frenet frame state variables and the curvature of the reference path for the prediction and control horizon. The dynamic MPC uses the model described in Eq. 8 for determining the requested control signal (δ_c). This MPC can be both the

parameter-free and the classical method in this approach (as these two types are detailed later in Section 4. It should be noted that in this structure, the errors are not fed back directly. Only the prediction-based part compensates for them, which naturally introduces a delay in the reaction.

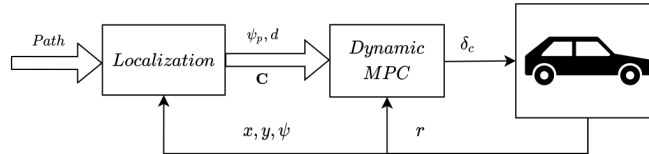


Figure 3

Control structure using the linearized dynamic model

3.2 Nonlinear Dynamic Model-based Feedback Structure

In this section, a different dynamics model handling method is proposed. This method handles the problem in a cascade structure, as can be seen in Fig. 4. The outer loop is a model predictive method using the linearized kinematic model (Eq. 2), and the inner loop realizes an internal model control (IMC) structure. The kinematic-based MPC does not determine the required road wheel angle but determines the required ρ , which is the most important state variable of the system concerning the lateral behavior. This MPC can also use the parameter-free approach or the classical method detailed later. It should be noted that proper limitations should be used in the outer loop to ensure feasibility in this cascade structure.

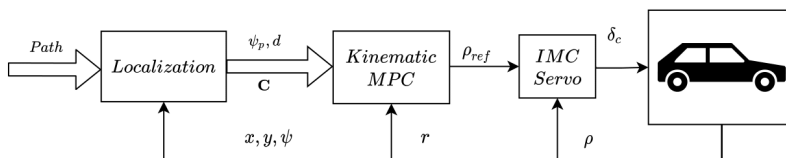


Figure 4

Control structure using kinematic and dynamic models in a cascade structure

The inner IMC structure can be seen in Fig. 5. A realizable inverse is placed on the feedforward branch that can determine the required road wheel angle (RWA). Here, the physical limitations (rate limit and final value limit) of the given system can be enforced. The inverse of the model is calculated based on the nonlinear model of the vehicle (Eq. 3) by solving the nonlinear equation for ρ_{ref} , using numerical approximation. The calculated RWA is then actuated in the vehicle and inputted to a model connected parallel with the plant. Then, the difference between the model and the plant is fed back through an autoregressive-like filter. This filter is responsible for noise suppression, using the following equation:

$$\rho_{fb}(k) = a_f(\rho(k) - \rho_m(k)) + (1 - a_f)\rho_{fb}(k - 1), \quad (9)$$

where a_f is the filter parameter, k is the discrete-time step-index, ρ is the value measured on the vehicle, and ρ_m is the value calculated by the model. The whole feedback loop is responsible for compensating the external noises and the effect of parameter mismatch. The presented structure exceeds the classical MPC method with a feedback loop. This system can react faster to the disturbances since the most critical internal state parameter is controlled in the inner loop. Additionally, this method can be extended to handle multi-actuator systems [19].

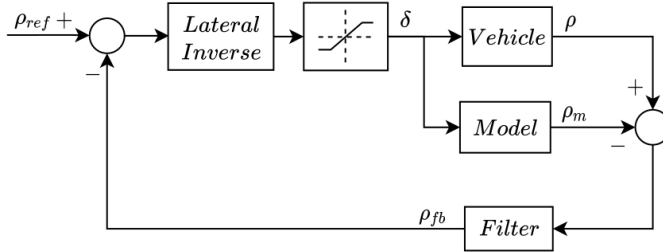


Figure 5

The inner loop realizing IMC

4 Model Predictive Control

The model predictive control method is an advanced technique widely used in various fields of optimal control problems [24]. The Model Predictive Control approach mentioned above will be detailed in this section by presenting the classical MPC method. The paper focuses on the main disadvantage of the classical MPC methods: they have cost functions with mixing values by tuning weights. The proposed parameter-free, hierarchical method that answers this problem will be detailed in this section.

4.1 Classical Model Predictive Control

Model predictive control is the most common approach among predictive controllers. The most obvious use case of predictive controllers is the discrete-time version with finite prediction and control horizon. For the sake of simplicity, the prediction and the control horizons are defined to be equal.

In general, the cost functions of predictive controllers include two parts. The control signals and the reference tracking error are included in these cost functions [24]. Defining N as the length of the horizon, N_x as the number of the

states, N_u as the number of inputs, N_y as the number of outputs, $\mathbf{x}=[x_1, x_2, \dots, x_{N_x}]$ as the state vector, $\mathbf{u}=[u_1, u_2, \dots, u_{N_u}]$ as the input vector, and $\mathbf{y}=[y_1, y_2, \dots, y_{N_y}]$ as the output vector, the general MPC problem formulation can be given. Minimize:

$$J(\mathbf{x}(0), \mathbf{u}(1 \dots N)) = \sum_{j=1}^{N_y} \sum_{k=1}^N W_{e_j}(k) e_j^2(k) + \sum_{i=1}^{N_u} \sum_{k=1}^N W_{u_i}(k) \Delta u_i^2(k), \quad (10)$$

where $e_j(k)$ is the difference between the j^{th} reference and the predicted state value at the k^{th} time step ($e_j(k)=y_j(k)-y_j^{\text{ref}}(k)$), and W_{e_j} is the corresponding weight. For the control signal, $\Delta u_j(k)^2$ corresponds to the control effort performed at the j^{th} input at the k^{th} time step, and $W_{u_j}(k)$ is the corresponding weight.

The minimization should be performed subject to constraints coming from the state equations of the controlled system:

$$\begin{aligned} \mathbf{x}(k+1) &= f(\mathbf{x}(k), \mathbf{u}(k)), \quad k = 1, \dots, N \\ \mathbf{y}(k) &= g(\mathbf{x}(k), \mathbf{u}(k)), \quad k = 1, \dots, N, \end{aligned} \quad (11)$$

and to the constraints derived from the limitations of the states, the inputs, and the outputs.

It can be seen that this cost function formulation has $N \cdot (N_y + N_u)$ weighting parameters. These weights provide the possibility for the designers of the controller to determine different weighting strategies in the cost function. The simplest solution is when the weights are constant for the whole horizon for each input or output. In some cases, the cost concerning the final state is highlighted compared to the running cost.

The existence of this amount of tuning parameters is twofold. On the one hand, the system performance can be maximized in predefined specific scenarios by finding the proper parameter tuning. On the other hand, the cost function including these parameters is a mixture of different values (considering physical meaning) on a different scale. Even if these values are normalized in some approaches, it is hard to interpret the real meaning of the cost function in the control environment, and it is not defined by physical law. However, there is no proper method given to find this parameter tuning. Additionally, there is no insurance that using the found parameter set, the performance of the system will remain if the test scenario or the system parameters change.

Linear state equations are created in Eq. 11. In our case, the linearized kinematic or the dynamic model (Eqs. 2 and 8) can be used for the state prediction. The future states and outputs can be determined in a closed form, using the linear equations if the future inputs are known. The linearized problem results in convex quadratic programming (QP) optimization problem. The problem complexity is crucial concerning the real-time applicability of the control method [25]. Due to the improvement of the available computing capacities, this problem can be solved in real-time, so the MPC method has become a widespread solution [24].

4.2 Parameter-free Model Predictive Control

The proposed method eliminates the weights and creates a cost function that has a physical interpretation. This is reached by decomposing the mixed cost function presented in Eq. 10. For the actual lateral control problem, the cost function should be used with the following variables:

$$\mathbf{y} = [\psi_p, \dot{a}], \quad \mathbf{u} = r, \quad \Delta \mathbf{u} = \rho, \quad (12)$$

According to the approach of a chauffeur, the path following is performed by minimizing the lateral and orientation error at a certain looking ahead distance. In the ideal case, these errors can be driven to zero during the control horizon so that these parts can be transformed to equality constraints. Finally, it results in that the remaining part includes only the control effort that was requested at the control input.

If the case is not ideal, this transformation of the cost function may cause infeasibility. A hierarchic solver method is introduced in order to solve this feasibility problem. This method drives the system step-by-step towards feasibility. Each equality constraint is first introduced as a cost function to minimize the distance from reaching equality. If equality is reached, it is introduced as a constraint while the following constraint is transformed into a cost function. After introducing all constraints that continuously maintain feasibility, the original cost function minimizing the control effort can be used in the optimization.

The outputs of the lateral control problem are formulating an integrator chain since the lateral error is connected to orientation via integration, as can be seen in Eq. 2. This chain determines the order of the introduction of the constraints: firstly, the orientation constraint is satisfied, then the position, to prevent overshoot.

In the following, the optimization problems of the sequential algorithm are given for the kinematic model defined in Eq. 2. Using the notations $\boldsymbol{\rho} = [\rho_1, \rho_2, \dots, \rho_N]$, and the state vector $\mathbf{x} = [r, d, \psi_p]$, the first optimization can be formulated:

$$\begin{aligned} & \min_{\boldsymbol{\rho}} |\psi_{pN}(\boldsymbol{\rho})| \\ \text{s. t. : } & |\rho_i| \leq \rho_{max}, \quad i = 1, \dots, N \\ & |r_i(\boldsymbol{\rho})| \leq r_{max}, \quad i = 1, \dots, N, \end{aligned} \quad (13)$$

where the limitations are considered for the yaw rate and its derivative, coming from the physics of the vehicle, and the state values are calculated in a closed form using the measured states and the linear models presented in section 2. After ensuring the orientational constraint, the lateral error is minimized by replacing the cost function with $|d_N(\boldsymbol{\rho})|$ and adding an equality constraint $\psi_{pN}(\boldsymbol{\rho})=0$ to equation (13). Finally, in the third optimization, the control effort is minimized, transforming the last goal into a constraint:

$$\begin{aligned}
& \min_{\rho} \sum_{i=1}^N \rho_i^2 \\
\text{s. t.: } & |\rho_i| \leq \rho_{\max}, \quad i = 1, \dots, N \\
& |r_i(\rho)| \leq r_{\max}, \quad i = 1, \dots, N \\
& d_N(\rho) = 0 \\
& \psi_{pN}(\rho) = 0.
\end{aligned} \tag{14}$$

In this hierarchical method, the algorithm performs the following optimization only if the minimization reaches zero, showing that the constraint is feasible. After the optimization, the first element of the optimal control vector is actuated, realizing the receding horizon approach.

This way, the weights are eliminated, but on the other hand, instead of one, three optimizations should be performed to ensure feasibility. Since the model is stable and the problem is feasible, the controller is stable [24]. However, using the linearized model, the optimization problems are created to be Convex problems, so they have unique solutions and can be solved in real-time [26].

5 MPC Controllers for Comparison

Firstly, the outer and the inner loop of the proposed method will be examined separately, and then the whole structure will be compared to a nominal solution. In this section, the three MPC models used as a base for the proper comparison are presented to support the presented cost function and dynamics handling approach.

5.1 Kinematic MPC with Mixed Cost

In this case, the classical MPC with mixed cost function uses the kinematic bicycle model presented in Eq. 2, within the structure presented in Fig. 4. This controller is the reference for the outer loop comparison, denoted as MIXIMC. The optimization problem is formulated as follows:

$$\begin{aligned}
& \min_{\rho} \sum_{i=1}^N \left(K_d d_i^2(\rho) + K_{\psi_p} \psi_{p_i}^2(\rho) + K_{\rho} \rho_i^2 \right) \\
\text{s. t.: } & |\rho_i| \leq \rho_{\max}, \quad i = 1, \dots, N \\
& |r_i(\rho)| \leq r_{\max}, \quad i = 1, \dots, N,
\end{aligned} \tag{15}$$

where K_d , K_{ψ_p} , and K_{ρ} are the weights for the lateral error, orientation error, and yaw acceleration, respectively.

In this case, it should be noted that these weights are considered constant over the control horizon. However, in some cases, better performance can be reached by having weights that are changing over the horizon. It results that the cost function of the optimization problem in Eq. 14 may contain $3N$ parameters.

5.2 Dynamic MPC with Parameter-Free Cost Function

The second comparison focuses on the presented inner loop approach, the IMC structure. The reference controller for this comparison is created based on the linear dynamics model described in Section 3.1, using the structure presented in Fig. 3. This controller is denoted as PFD. Since the model includes the road wheel angle, it serves as an input for the system. The optimization problem is formulated similarly to the sequential, hierarchical approach presented in Section 4.2.

Using the dynamic model, the MPC handles the system dynamics and determines the control signal, using the parameter-free approach, but without having feedback for the internal states of the system. The cost function is calculated similarly to the kinematic MPC, based on ρ for a proper comparison. The only difference is that the linear dynamic model is used instead of the kinematic model. However, this approach does not consider the nonlinearities in the system and does not have direct feedback for the dynamics behavior.

5.3 LKA Subsystem

The publicly available most complex and advanced controller is chosen for the complete comparison of the system. The lane-keeping assist (LKA) subsystem [27] includes the linearized dynamic model (Eq. 8) expanded with state estimation for handling the input-output disturbances.

This is an adaptive model predictive control structure, implemented using the Frenet-frame. The MPC formulation of this system is quite similar to the one presented in the previous Section. The disturbance rejection is realized by estimating the plant model and the controller states based on a disturbance model and the measurement noise model, using a linear-time-varying Kalman filter (LTVKF). The state estimation introduces further tuning parameters since two gain matrices are needed for its algorithm. Additionally, this system uses scale factors that the controller designer should also specify.

In this paper, the default values of this subsystem were used (estimator gains, scale factors, etc.). Only the cost function weights were tuned during the comparison method.

6 Simulation Results

In this section, the simulation results will be presented. Firstly, the simulation environment and the evaluation methods are detailed, followed by the three comparisons for the inner loop, the outer loop, and the whole proposed structure.

6.1 Simulation Environment

The simulation environment was implemented in MatLab & Simulink. The plant was modeled using the nonlinear dynamic bicycle model using the nonlinear tire model, presented in Section 2.2, using the nominal parameters (Table 1). The simulation run with fixed step size ($dt = 0.002$ s), using the ode4 solver. The optimization problems were implemented as MatLab function blocks, using the Optimization Toolbox of MatLab. The controller runs on a lower frequency with $f_c = 50$ [Hz]. The horizon was set to be $N = 15$ (similar to the LKA reference controller), and the discrete step of the linear system prediction was set to be $dt_p = 0.05$ [s]. The filter parameter (α_f) was 0.3, which is sufficient against the numerical errors of the simulation.

Due to the structure of the simulation software, it is easy to modify the parameters of the plant, emulating the mismatch between the controller model and the plant. Also, external disturbances can be added to the model for testing the disturbance rejection performance of the controller.

6.2 Evaluation Methods

In this paper, two types of evaluation methods were used to compare. Scalar-based evaluation gives a scalar number as a result of a successful measurement. The scenario-based comparison is performed based on a higher level overlook on the system, where a single scalar value is not enough to characterize the system. The following integral and maximal values were considered for the scalar evaluation of the presented controllers:

- Max error: $e_{\max} = \max|e(k)|$, $e = \{d, \Psi_p\}$, calculating the maximum absolute value of the error. The errors are the lateral and orientation deviation at the k^{th} time step.
- Error integral: $e_{\text{int}} = dt \sum e^2(k)$, $e = \{d, \Psi_p\}$, the discrete-time integral of the squared value of the error.
- Max control value: $u_{\max} = \max|u(k)|$, the maximal value of the actual control input of the system (that is usually included in the mixed cost function or is at the end of the hierarchic solution).
- Control integral: $u_{\text{int}} = dt \sum u^2(k)$, the discrete-time integral of the squared value of the control input.

The scenario-based comparison is based on the examination of the behavior of the vehicle during different predefined test scenarios [28]. Concerning the lateral control, the overshooting and the setting ability of the reference tracking were investigated. Three test cases were defined for examining the controllers: straight-line following with initial lateral error (dy), straight-line following with initial orientation error ($d\psi$), and the lane change maneuver (LC).

These tests can give a picture of the controller upon its state error rejection and path-following performance. Also, these tests were expanded with further examinations with disturbance rejection and handling the plant parameter changes.

6.3 Comparison of the Cost Functions

Firstly, the outer loops were compared, implementing the proposed IMC-based inner loop for the dynamics handling as it is described in Section 3.1. The tuning parameter-free (PF) method was compared with the MPC with the mixed cost function-based (MIX) method to examine the outer loop. The three parameters of the MIX controller were tuned so that during the lane-change maneuver, it reaches the same control effort as the PF has, resulting in the parameter tuning: $K_d = 6$, $K_p = 0.5$, $K_\psi = 10$. The performed maneuver can be seen in Fig. 6.

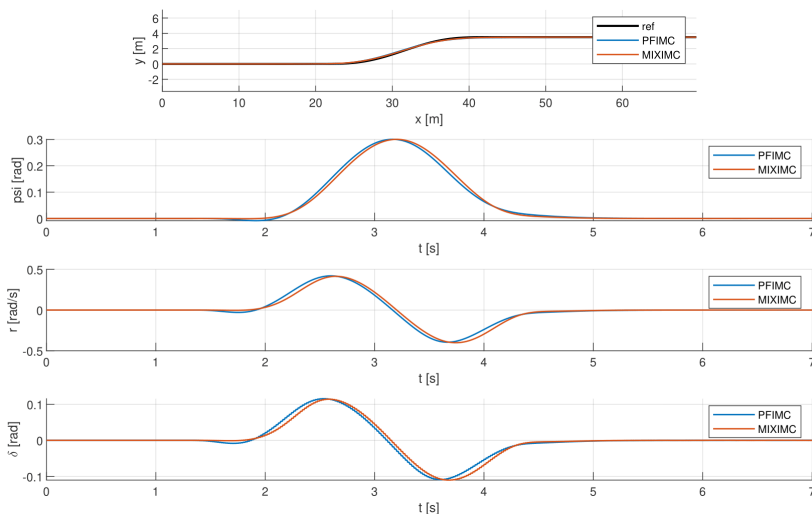


Figure 6

The trajectories of the Parameter-free and the Mixed MPC methods performing the lane change maneuver while using the same lateral control effort

Another scenario was performed, using the same parameter set tuned for the lane change to see the sensitivity of the parameter tuning. The results of the lateral error test (dy) can be seen in Fig. 7.

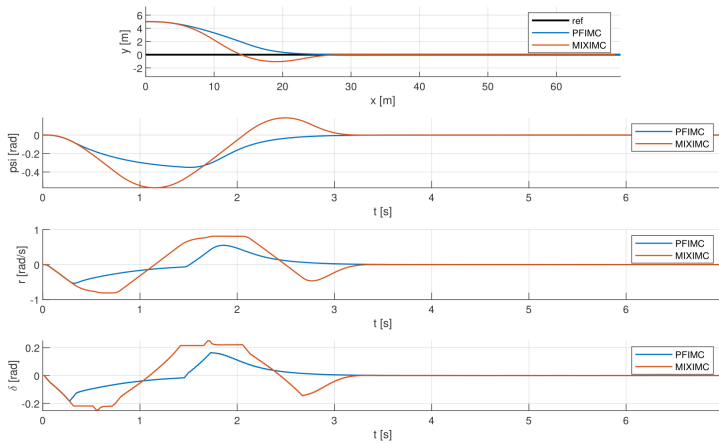


Figure 7

Comparison of lateral error elimination with the parameter set tuned for the lane change maneuver

Table 2

Maneuver evaluation of controllers

Man.	Cont.	u_{int}	u_{max}	d_{int}	d_{max}	Ψ_{pint}	Ψ_{pmax}
LC	PF	1.263	1.271	$2.10 \cdot 10^{-2}$	$1.43 \cdot 10^{-1}$	$7.76 \cdot 10^{-4}$	$2.84 \cdot 10^{-2}$
	MIX	1.251	1.262	$1.26 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$8.55 \cdot 10^{-4}$	$2.94 \cdot 10^{-2}$
dy	PF	2.740	2.108	24.15	5	0.132	0.350
	MIX	7.437	2.220	20.51	5	0.288	0.571

Both maneuvers were evaluated, and the results can be seen in Table 2. It can be seen that in the LC maneuver, as it was the goal, both controllers perform with almost identical control effort (u_{int}). Also, in the first test case, the MIX controller performs better in all the evaluations corresponding to the control input and the lateral error. The PF controller beats only the orientation error. There is a big difference between the controllers in the second test case. The MIX controller has an aperiodic setting in the position and the orientation; therefore, the results of the evaluations are significantly worse. The parameter-free method is now shown to be independent of the test scenario, providing an aperiodic setting in all the test cases. This stability and reliability are a great advantage among predictive controllers, even when the MIX-based approach performs better concerning the errors or the control signal.

6.4 Comparison of Dynamics Handling

In this section, the dynamics handling solutions, presented in Section 3: the proposed tuning parameter-free (PF) method with IMC in the inner loop is compared with the method that uses the linearized dynamic model with the PF

method in the optimization problem formulation. Therefore, only the dynamics handling method differs between the two approaches.

The first test case was the orientation error rejection ($d\psi$) under low- μ conditions. The parameters of the test case were: $\mu=0.5$, $\psi_0 = \pi/6$ [rad], and $u_x=10$ [m/s]. The results of this test can be seen in Fig. 8.

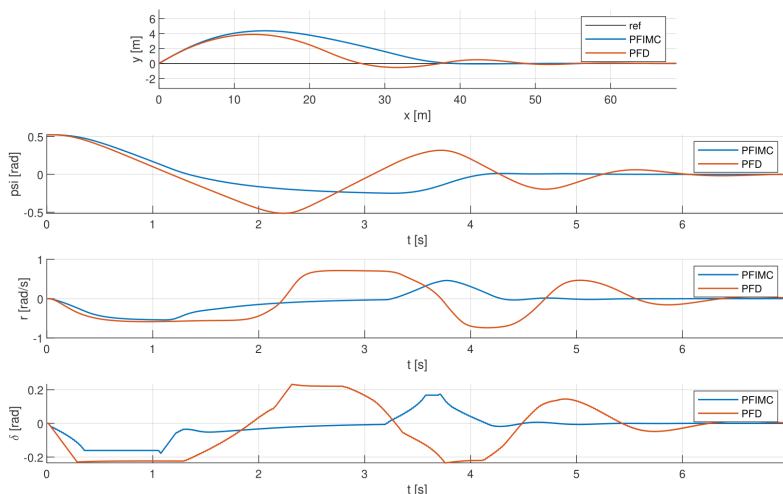


Figure 8

Comparison of the Dynamic model-based MPC and the kinematic-IMC-based controllers in low- μ situation

The IMC-based controller can handle the nonlinearities of the bicycle model successfully, even if the vehicle to road coefficient changes. This induces that at steering movements with high amplitude, the vehicle gets closer to its limits. The linearized dynamic model-based method controller results in a periodic setting in the errors. However, the IMC-based method can handle the nonlinearities with an aperiodic setting.

The second test was performed during straight-line following, examining the external disturbance handling ability of the controller. In this test case, the vehicle ran straight, with constant speed ($u_x=10$ [m/s]), and then at time $t = 0.5$ [s], a constant torque disturbance ($M_d = 9000$ [Nm]) around axis Z was added, inducing yaw moment into the system. The results of this test can be seen in Figure 9.

Both controllers compensate for the disturbance by turning the steering wheel in the proper direction. Due to the IMC loop, the proposed algorithm can react much faster to the disturbance, resulting in total disturbance rejection, eliminating the position error of the vehicle. However, the dynamic model-based controller has a significant constant lateral error, resulting from its structure since only the output states are fed back within the MPC method. To sum up, the proposed IMC method

can handle the nonlinearities and parameter changes, together with the appearing unmeasured external disturbances of the vehicle.

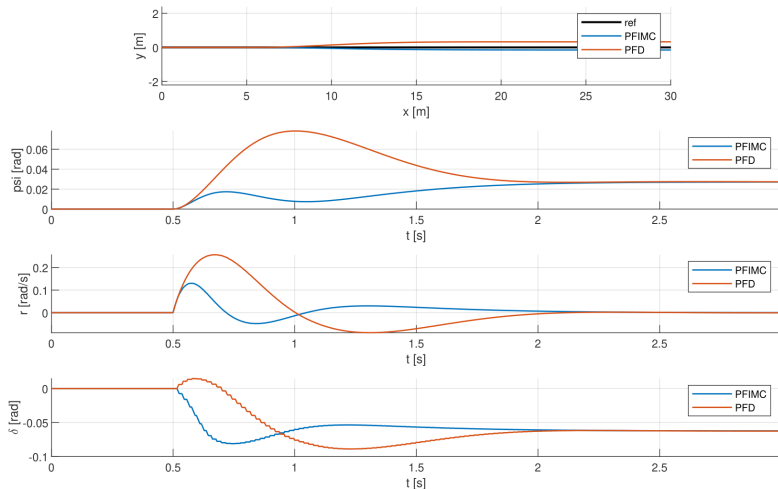


Figure 9

Comparison of the Dynamic model-based MPC and the kinematic-IMC-based controllers under Torque disturbance

6.5 Comparison with the LKA Subsystem

Finally, the proposed controller structure is compared with the LKA subsystem. Since the LKA subsystem should have the same control frequency and prediction frequency, both were set to be $f_c = 50$ [Hz], $dt_p = 0.05$ [s]. The parameters of the LKA subsystem were tuned similarly to the method presented in Section 6.3. The weight of the manipulated variables rate was set to be 4, and the output variables weight (concerning the default scaling factor given by the subsystem) was equally 1. The test was the dynamic lane change test performed on low μ ($\mu = 0.6$). The results of this test can be seen in Figure 10. and the scalar evaluations in Table 3.

Table 3
Numerical results comparison of LKA and PF-IMC under low μ

Controller	u_{int}	u_{max}	d_{int}	d_{max}	Ψ_{pint}	Ψ_{pmax}
PF-IMC	2.112	1.538	$1.524 \cdot 10^{-3}$	$3.591 \cdot 10^{-2}$	$2.634 \cdot 10^{-2}$	$1.516 \cdot 10^{-2}$
LKA	2.488	1.930	$4.510 \cdot 10^{-3}$	$6.536 \cdot 10^{-2}$	$4.168 \cdot 10^{-4}$	$2.049 \cdot 10^{-2}$

The results show that both systems can sufficiently perform the maneuver. However, the LKA system has an overshoot at the end of the maneuver, resulting in a small oscillation in the control signal. Also, it is significant that the proposed controller performs better in all the points of comparison.

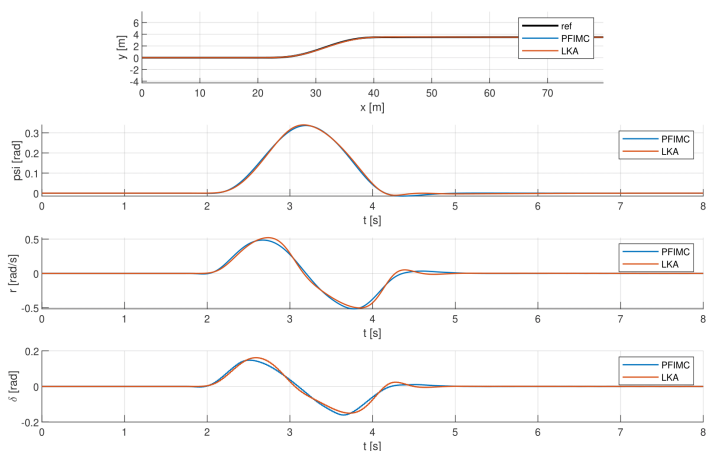


Figure 10

Path comparison of LKA and PF-IMC under low μ

Conclusions

In this paper, a novel approach is presented for lateral vehicle control. This parameter-free model predictive control transforms the classical model predictive control problem into a series of optimization problems, where the cost functions are hierarchized. Therefore, each cost function consists of a physical phenomenon. Due to this formulation, there is no need for tuning parameters in the system. This method is combined with the IMC structure for handling the dynamics and the nonlinearities of the system and implementing robustness against parameter changes and external disturbances.

The proposed method is compared with the well-known and widespread MPC methods. Three different comparisons were performed to see the advantages of the proposed method in detail, each focusing on a specified field of the proposed system. In Table 4, the main differences between the presented structure and the classical structure are gathered.

Table 4
Comparison table

LKA system	PF-IMC
Weights tuned intuitively	Hierarchical solver without weights
Different tuning and evaluation goals	Goal based on a physical phenomenon
Scenario-specific parameters	Consistent response overall scenarios
Single optimization task	Three optimization tasks
Linearized dynamics	Linear kinematics with nonlinear dynamics

From the simulation results and the aspects described in Table 4, it can be seen that the proposed algorithm has an outstanding contribution considering the

predictive controllers. This method can be generalized so that the reduction of tuning parameters can be reached in other control problems where MPC is used.

References

- [1] Nilsson J, Falcone P, Ali M, et al. Receding horizon maneuver generation for automated highway driving. *Control Engineering Practice*. 2015 08;41
- [2] Max G, Lantos B. Time optimal control of four-in-wheel-motors driven electric cars. *Periodica Polytechnica Electrical Engineering and Computer Science*, 2014;58(4):149-159
- [3] Chen T, Babanin A, Muhammad A, et al. Modified evolved bat algorithm of fuzzy optimal control for complex nonlinear systems. *Rom J Inf Sci Technol*. 2020;23:T28-T40
- [4] Precup RE, Preitl S, Petriu EM, et al. Generic two-degree-of-freedom linear and fuzzy controllers for integral processes. *Journal of the Franklin Institute*. 2009; 346(10):980-1003
- [5] Redjimi H, Tar JK. Multiple components fixed point iteration in the adaptive control of single variable 2nd order systems. *Acta Polytechnica Hungarica*. 2021;18(9):69-84
- [6] Roman RC, Precup RE, Hedrea EL, et al. Iterative feedback tuning algorithm for tower crane systems. *Procedia Computer Science*. 2022;199:157-165
- [7] Tan Y, Chang J, Tan H, et al. Integral backstepping control and experimental implementation for motion system. *Proceedings of the 2000. IEEE International Conference on Control Applications;02;2000*, pp. 367-372
- [8] Arogeti SA, Berman N. Path following of autonomous vehicles in the presence of sliding effects. *IEEE Transactions on Vehicular Technology*. 2012;61(4):1481-1492
- [9] Galceran E, Eustice RM, Olson E. Toward integrated motion planning and control using potential fields and torque-based steering actuation for autonomous driving. In: *2015 IEEE Intelligent Vehicles Symposium (IV)*; 2015, pp. 304-309
- [10] Menhour L, D'Andr'ea-Novel B, Fliess M, et al. Coupled nonlinear vehicle control: Flatness-based setting with algebraic estimation techniques. *Control Engineering Practice*. 2014;22:135-146
- [11] Bodo Z, Lantos B. High level kinematic and low level nonlinear dynamic control of unmanned ground vehicles. *Acta Polytechnica Hungarica*. 2019 may;16(1)
- [12] De Luca A, Oriolo G, Samson C. Feedback control of a nonholonomic car-like robot. Berlin, Heidelberg: Springer Berlin Heidelberg; 1998, Chapter 4; pp. 171-253
- [13] Talvala K, Kritayakirana K, Gerdes J. Pushing the limits: From lanekeeping to autonomous racing. *Annual Reviews in Control*. 2011 04;35:137-148

-
- [14] Li X, Sun Z, Liu D, et al. Combining local trajectory planning and tracking control for autonomous ground vehicles navigating along a reference path. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC); 2014, pp. 725-731
- [15] Babqi AJ, Alamri B. A comprehensive comparison between finite control set model predictive control and classical proportional-integral control for grid-tied power electronics devices. *Acta Polytechnica Hungarica*. 2021; 18(7):67-87
- [16] Jalali M, Khajepour A, Chen SK, et al. Integrated stability and traction control for electric vehicles using model predictive control. *Control Engineering Practice*. 2016 09; 54:256-266
- [17] Lin F, Chen Y, Zhao Y, et al. Path tracking of autonomous vehicle based on adaptive model predictive control. *International Journal of Advanced Robotic Systems*. 2019 09; 16:1729881419880089
- [18] Huang C, Li B, Kishida M. Model predictive approach to integrated path planning and tracking for autonomous vehicles. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC); 2019, pp. 1448-1453
- [19] Kovacs A, Vajk I. Integrated path planning and lateral-longitudinal control for autonomous electric vehicles. In: 2021 AEIT International Conference on Electrical and Electronic Technologies for Automotive; 2021, pp. 1-6
- [20] Kovacs A, Vajk I. Integrated lateral and longitudinal control with optimization-based allocation strategy for autonomous electric vehicles. *Journal of Advanced Transportation*. 2021 11;2021:1-18
- [21] Kanatnikov A, Liu W, Tkachev S. Path coordinates in a 3d path following problem. *Mathematical Models and Computer Simulations*. 2018 05;10:265-275
- [22] Kabzan J, Hewing L, Liniger A, et al. Learning-based model predictive control for autonomous racing. *IEEE Robotics and Automation Letters*. 2019; 4(4):3363-3370
- [23] Pacejka H. *Tire and vehicle dynamics*. Elsevier; 2005
- [24] Rakovic SV, Levine W. *Handbook of model predictive control*. Springer; 2018
- [25] Preitl Z, Precup RE, Tar JK, et al. Use of multi-parametric quadratic programming in fuzzy control systems. *Acta Polytechnica Hungarica*. 2006; 3(3):29-43
- [26] Alizadeh F, Goldfarb D. Second-order cone programming. *Mathematical Programming*. 2003 Jan;95:3-51
- [27] MATLAB. Lane keeping assist system [<https://www.mathworks.com/help/mpc/ref/lanekeepingassistsystem.html>]; 2022, Accessed: 2022-02-10
- [28] Ni J, Hu J, Xiang C. Envelope control for four-wheel independently actuated autonomous ground vehicle through afs/dyc integrated control. *IEEE Transactions on Vehicular Technology*. 2017;66(11):9712-9726
-

A Hybrid Multi-criteria and Creative, Problem-solving Approach, for Measuring Local Values of Information Technology Products

Vesna Čančer

University of Maribor, Faculty of Economics and Business
Razlagova 14, 2000 Maribor, Slovenia, vesna.cancer@um.si

Abstract: The frame of the procedure for multi-criteria decision making, that support complex problem solving, has been well-verified in business practice, but lacks a fully defined approach, for the determination of local alternative values. The purpose of this paper is to develop a hybrid multi-attribute value model and creative problem-solving approach for measuring local alternatives' values. It also aims to verify the applicability of this approach in an Information Technology company. Within measuring local alternatives' values, the paper describes how to create increasing and decreasing piecewise, linear functions by using a bisection method. It introduces a systematic approach for the determination of the local alternatives' values, by using the "six questions" technique. In addition to the theoretical statement of the hybrid multi-criteria and creative problem-solving approach in determining the local alternatives' values, the approach is applied to the "real-life" problem of choosing the most appropriate switch, for small and medium-sized companies. The resultant increasing and decreasing piecewise linear functions, can serve as a good approximation of exponential value functions, that would otherwise, require a large series of data and a demanding statistical knowledge. The presented approach can be applied to a wide range of organizational and management problems for the selection, assessment, and evaluation of alternatives.

Keywords: creative problem solving; information technology; multi-criteria decision making; piecewise linear value function; prescriptive approach

1 Introduction

Consideration of a prescriptive approach to decision making [22], which advises against the exclusive treatment of people, as perfectly rational individuals, resulted in systematic decision-making procedures to support smart decisions. They follow the decision-making phases and consist of well-described steps [26]. Among them, the frame procedure for multi-criteria decision making (MCDM) by using the group of methods based on assigning weights [7] that follows the phases of the Belton and Stewart's decision-making process [2] has been well-verified

in practice, mainly to support the preparation of business decisions for complex problem solving in small and medium-sized enterprises. The particularities of the above-mentioned frame procedure for MCDM, which includes the following steps: problem definition, elimination of unacceptable alternatives, problem structuring, measuring local alternatives' values, criteria weighting, synthesis, ranking and sensitivity analysis [7], have been introduced in the selection of Information Technology (IT) services and products [7]. The growing role of IT in meeting the needs in enterprises' growing businesses and supporting their integration into global economic processes [13], which also stood out during the Corona crisis period [29], underlines the need for the methodological development of individual steps.

In MCDM based on assigning criteria weights, measuring alternatives' values encompasses measuring local alternatives' values with respect to each criterion on the lowest hierarchy level, and synthesis, i.e., measuring alternatives' values with respect to all criteria structured in a problem hierarchy. The purpose of this paper is to develop a hybrid multi-criteria and creative problem-solving approach to measuring alternatives' values with respect to criteria on the lowest hierarchy level, the so-called local alternatives' values.

The local values of alternatives can be measured indirectly, e.g., by value functions or pairwise comparisons, or directly. According to Kadziński *et al.* [18], a direct specification of a set of parameter values can be difficult for decision makers since it requires considerable cognitive effort. For this reason, indirect specification of preference information is considered more user-friendly. The recognized advantage of the indirect over the direct approach is that it allows decision makers to investigate their evaluation of parts of the problem, i.e., alternatives according to criteria, and to elicit their preferences to alternatives with respect to each criterion on the lowest level. Rezaei [23] noted that the existing MCDM methods often use simple monotonic linear value functions for measuring alternatives' values and pointed out that the assumption of an increasing or decreasing linear function between a criterion level (over its entire range) and its value might lead to improper results. Ghaderi and Kadziński [14] pointed out that the shape of value function is of great importance in different areas of research in decision analysis, including multi-criteria decision making as it decides upon the contribution of various performances into the comprehensive value of an alternative. They found that accounting for the structural patterns at the population level considerably improves the predictive performance of the constructed value functions at the individual level [14]. Greco *et al.* [16] introduced the concept of a representative value function in robust ordinal regression applied to multiple criteria sorting problems and proposed a way of selecting a representative value function among the set of compatible ones. In [16] the authors introduced several examples of level-increase value function on multiple sections in real world decision problems. Rezaei [23] proposed a set of the following piecewise linear functions: increasing, decreasing, V-shape, inverted V-shape, increase-level,

level-decrease, level-increase, decrease-level, increasing stepwise, and decreasing stepwise. This set of piecewise linear functions, however, does not explicitly expose piecewise linear increasing nor piecewise linear decreasing value functions with multiple (at least two) sections on which the absolute value of the slope coefficient is between 0 and 1. To fill this gap, this paper deals with the piecewise linear increasing and piecewise linear decreasing value functions, with multiple (at least two) sections on which the absolute value of the slope coefficient is between 0 and 1. The sections can be defined by using the bisection method [1] [27]. The first goal of this paper is therefore to delineate how to create increasing and decreasing piecewise linear functions by using a bisection method.

Since decision makers and/or the experts who measure the values of alternatives often do not have either specific mathematical knowledge or do not have enough time to study mathematical expressions and procedures, we propose that the elicitation of their preferences to determine value functions can be supported by using methods based on questions, e.g., W technique, six questions technique, Why and 5 Whys [3] [5]. The second goal of this paper is to introduce a systematic approach to determine the local alternatives' values by using a six questions technique.

The organization of the rest of this paper is as follows. The next section delineates how to create the increasing and the decreasing piecewise linear value functions with four sections, based on the bisection method, proposes a process on how to support the determination of the local alternatives' values by using the six questions technique, and defines the real-life problem, together with the data presentation. Then the approach proposed in this paper is illustrated in detail on a real-life case. The paper also discusses the obtained results, together with the approach's limitations and further research possibilities. The concluding part highlights the theoretical and practical implications of the proposed hybrid multi-criteria and creative problem-solving approach to measuring local alternatives' values.

2 Methods

2.1 A Systematic Approach to Determine Value Functions

It is well known that the choice of an appropriate technique for assessment of value function depends on the decision problem, its context, and the decision maker's characteristics [19]. According to Segura and Maroto [25], decision making not only considers opinions and judgments, but also integrates historical data and expert knowledge. Based on the research, knowledge and experience in measuring local alternatives' values of the author of this paper, it has to be pointed

out that the set of influential factors to the assessment of value function depends on the type of a criterion, the data, and decision maker's preferences.

The systematic approach introduced in this paper includes the creation of piecewise linear functions by using the bisection method. In this method, two objects are presented to a decision maker; he is asked to define the attribute level that is halfway between the objects in respect of the relative strengths of the preferences. This paper delineates how to create the increasing and then also the decreasing piecewise linear functions with four sections by using a bisection method.

Let us delineate how to create the increasing piecewise linear function with four sections by using a bisection method. First, the two extreme points, the least preferred evaluation object x_{min} and the most preferred evaluation object x_{max} are identified and associated with values $v(x_{min}) = 0$, $v(x_{max}) = 1$. Then, a decision maker is asked to define a midpoint x_1 , for which:

$$(x_{min}, x_1) \sim (x_1, x_{max}) \quad (1)$$

where \sim indicates the decision maker's indifference between the changes in value levels. While x_1 is in the middle of the value scale, we must have:

$$v(x_1) = 0.5 v(x_{min}) + 0.5 v(x_{max}) = 0.5 \quad (2)$$

Thus, we determined the increasing piecewise linear function with two sections. To create four sections, each of the existing two sections obtained by (1) and (2) must be halved according to the alternative's value. For the midpoint x_2 between x_{min} and x_1 , for which:

$$(x_{min}, x_2) \sim (x_2, x_1) \quad (3)$$

we obtain:

$$v(x_2) = 0.5 v(x_{min}) + 0.5 v(x_1) = 0.25 \quad (4)$$

and for the midpoint x_3 between x_1 and x_{max} , for which:

$$(x_1, x_3) \sim (x_3, x_{max}) \quad (5)$$

we obtain:

$$v(x_3) = 0.5 v(x_1) + 0.5 v(x_{max}) = 0.75 \quad (6)$$

Let us also delineate how to create the decreasing piecewise linear function with four sections by using a bisection method. First, the two extreme points, the most preferred evaluation object x_{min} and the least preferred evaluation object x_{max} are identified and associated with values $v(x_{min}) = 1$, $v(x_{max}) = 0$. Then, a decision maker is asked to define a midpoint x_1 to which it applies (1). Again, while x_1 is in the middle of the value scale, we must have (2). Similarly, for the midpoint x_2 (between x_{min} and x_1) to which it applies (3), we obtain:

$$v(x_2) = 0.5 v(x_{min}) + 0.5 v(x_1) = 0.75 \quad (7)$$

and, for the midpoint x_3 (between x_1 and x_{max}), to which it applies (5), we obtain:

$$v(x_3) = 0.5 v(x_1) + 0.5 v(x_{max}) = 0.25 \quad (8)$$

2.2 Use of the Six Questions Technique in Measuring Local Alternatives' Values

When measuring local alternatives' values with respect to each criterion on the lowest hierarchy level, it is important to ask the decision maker good questions (the term 'decision maker' includes both an individual and a group). For this purpose, we can use the six questions technique – the creative problem-solving method for problem definition, based on questions. The six questions technique is namely a structured method that examines a problem from multiple viewpoints. According to Cook [5], it is best used with rational problems due to its complexity. Moreover, it can be used individually or in groups. A general summary of the six questions technique includes stating the problem using the question 'In what ways might...?', writing down who, what, when, where, why and how questions that are relevant to the problem, answering the above written questions, and examining responses and using them for problem redefinitions [5]. In MCDM, the technique can be used to define problems in the first step of the frame procedure of MCDM [8]. The technique has already proven useful in indirect criteria weighting [6] [9].

In addition, we propose the following process of determining the local alternatives' values:

- 1) In what ways might the local alternatives' values be determined?
- 2) The who, what, where, when, why and how questions regarding the local alternatives' values are put and written down.
- 3) The questions are answered, and the local alternatives' values are determined and re-determined.

2.3 Data

The systematic approach to determine value functions is illustrated in detail on a real-life case of choosing the most appropriate switch, from the viewpoint of an IT company that offers switches to small and medium-sized companies. Alternatives are the switches that can be offered: Alternative 1 is Dell EMC Switch N1524P [10], Alternative 2 is C1000-24P-4G-L [4] and Alternative 3 is 6300M 24x 1G PoE / 4x SFP56 (JL662A) [17]. In Table 1, the data of alternatives with respect to criteria on the lowest hierarchy level (see Figure 1) are compiled from [4] [10] [17].

Table 1
Alternatives' data

Criterion	Data Type	Alternatives		
		Alternative 1	Alternative 2	Alternative 3
Ports total	Quantitative: number of choices	28	24	28
Switching bandwidth	Quantitative: Gbps	176	128	880
Forwarding rate	Quantitative: Mpps	164	95.23	660
Power over Ethernet	Quantitative: W	600	195	600
Maximum power consumption	Quantitative: W	871	250	674
Acoustic noise	Quantitative: dB	45	0	34.2
Power supply	Quantitative: number of choices	1	1	2
Warranty	Mixed: years or verbal description	3	For the period of ownership or use	5
Training	Quantitative: €	400	500	600
Price	Quantitative: €	2500	2125	3500

3 Results

The criteria hierarchy is presented in Figure 1. The criteria importance was together with the IT company's experts determined hierarchically. The criteria importance with respect to the global goal, which is choosing the most appropriate switch, was determined indirectly, by using the SWING method [28]: the change from the worst to the best level of technical criteria was considered the most important and was assigned 100 points; 70 points less, i.e., 30 points were assigned to the change from the worst to the best level of environmental criteria to reflect the importance of this change relative to the most important criterion change, and 30 points less than to the change from the worst to the best level of technical criteria, i.e., 70 points were assigned to the change from the worst to the best level of economic criteria; the first level criteria weights were obtained by normalization. The importance of economic sub-criteria was determined indirectly, too, by using the SMART method [11]: the change from the worst to the best training was considered the least important and was assigned 10 points; 20 points more, i.e., 30 points were assigned to the change from the worst to the best warranty to reflect the importance of this change relative to the least

important criterion change, and 30 points more than to the change from the worst to the best training, i.e., 40 points were assigned to the change from the highest to the lowest price. The SMART method was also used to indirectly determine the technical sub-criteria weights. Again, the above-mentioned sub-criteria weights were calculated by using normalization. The environmental sub-criteria weights were determined directly.

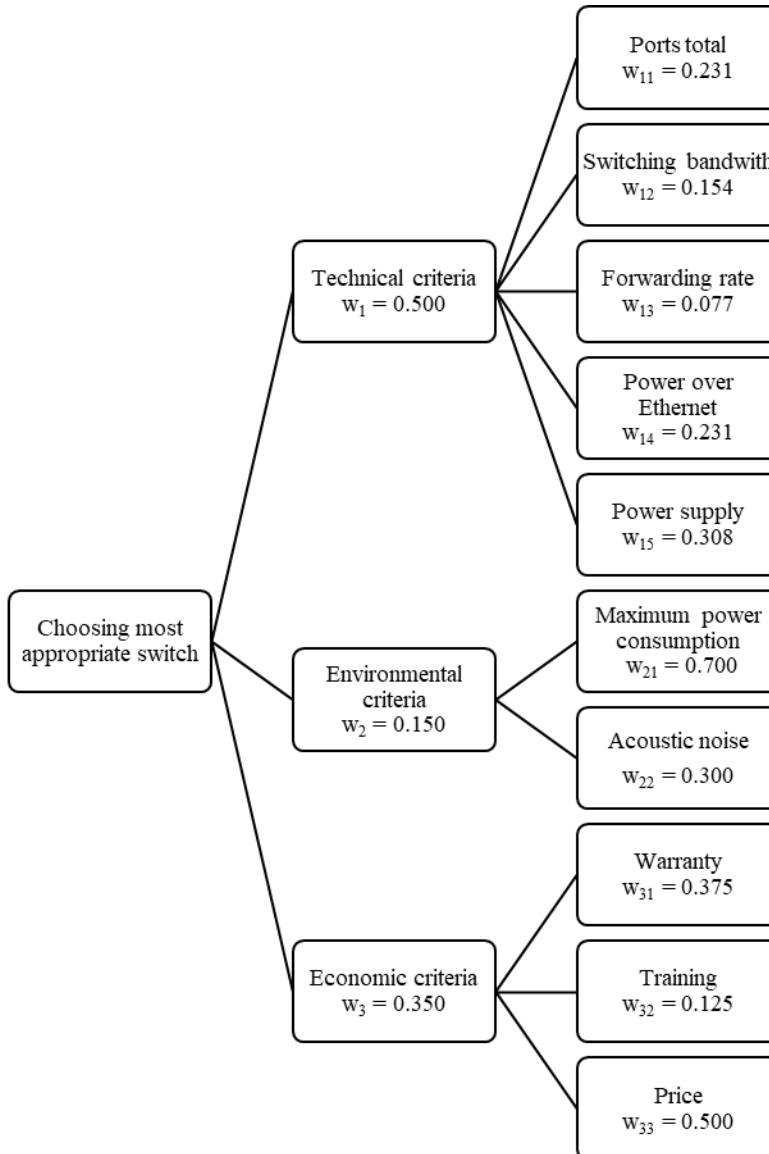


Figure 1
Criteria hierarchy and the weights

For measuring local values of alternatives, the coordinator, with appropriate knowledge for creative thinking techniques and for MCDM, asked and answered the typical question of the first step of the six questions technique process:

Q: In what ways might the local alternatives' values be determined?

A: Individually, in groups. Directly, indirectly, i.e., by using pairwise comparisons (verbal, numerical and graphical), and value functions (simple-monotonic, exponential, piecewise linear).

As the local alternatives' values were determined in groups, the group participants were defined in the second and the third step of the proposed six questions technique process.

Q: Who is responsible for this model building, including local alternatives' values determination?

A: Project manager, responsible for the defined problem solving, and IT experts.

Q: Who is competent to express preferences about the local alternatives' values?

A: Problem experts and/or experts in the field described by the considered criterion.

After the participants of the group for solving the problem were defined, they answered the questions regarding the model, successively put by the coordinator:

Q: Where will the model be used?

A: In small and medium-sized companies.

Q: When will the model be applied for problem solving?

A: In 2021 and beyond, for the next five years.

As this paper is focused on indirect specification of preference information about alternatives with respect to each criterion on the lowest hierarchy level with value functions, we present in more detail the questions (put by the coordinator) and the participants' answers expressing preferences to measure the local alternatives' values with value functions. Because the participants were not familiar with several ways of the local alternatives' ways determination, the coordinator briefly presented the ways of the determination of local alternatives' values. Then, the following question was asked for each criterion on the lowest hierarchy level:

Q: With respect to the criterion on the lowest hierarchy level, how will the local alternatives' values be determined?

When the response covered value functions, further questions referred to a more accurate determination of the value function. In this paper, we present questions for the bisection method to determine the piecewise linear function with multiple

– in this case four – sections, for measuring the local alternatives' values with respect to forwarding rate and with respect to price (Table 1).

To determine the increasing piecewise linear function for forwarding rate, the following questions were put and answered:

Q: Which is the least preferred evaluation object x_{min} so that $v(x_{min}) = 0$?

A: The least preferred evaluation object x_{min} is 50 Mpps.

Q: Which is the most preferred evaluation object x_{max} associated with $v(x_{max}) = 1$?

A: The most preferred evaluation object x_{max} is 850 Mpps.

Q: Why is x_{min} the least preferred evaluation object and x_{max} the most preferred evaluation object?

A: Because the greater the forwarding rate, the more favorable the alternative.

For the determination of sections, the following question based on (1) and (2) were put and answered:

Q: Which is a midpoint x_1 , for which $(x_{min}, x_1) \sim (x_1, x_{max})$, where \sim indicates the decision maker's indifference between the changes in value levels, so that $v(x_1) = 0.5 v(x_{min}) + 0.5 v(x_{max}) = 0.5$?

Because decision-makers were not familiar with mathematical expressions, a coordinator re-formulated the above written question:

Q: Which is a midpoint x_1 , which is considered equally good if the forwarding rate increases from x_{min} to x_1 , as if it increases from x_1 to x_{max} ?

A: The increase of the forwarding rate from 50 Mpps to 250 Mpps is equally favorable as its increase from 250 Mpps to 850 Mpps. The local value of x_1 is 0.5.

Thus, we determined the increasing linear function with two sections. To obtain the increasing linear function with four sections, the following questions based on (3) – (6) were asked:

Q: Which is a midpoint x_2 , which is considered equally good if the forwarding rate increases from x_{min} to x_2 , as if it increases from x_2 to x_1 ?

A: The increase of the forwarding rate from 50 Mpps to 100 Mpps is equally preferred as its increase from 100 Mpps to 250 Mpps. The local value of x_2 is 0.25.

Q: Which is a midpoint x_3 , which is considered equally good if the forwarding rate increases from x_1 to x_3 , as if it increases from x_3 to x_{max} ?

A: The increase of the forwarding rate from 250 Mpps to 500 Mpps is equally favorable as its increase from 500 Mpps to 850 Mpps. The local value of x_3 is therefore 0.75.

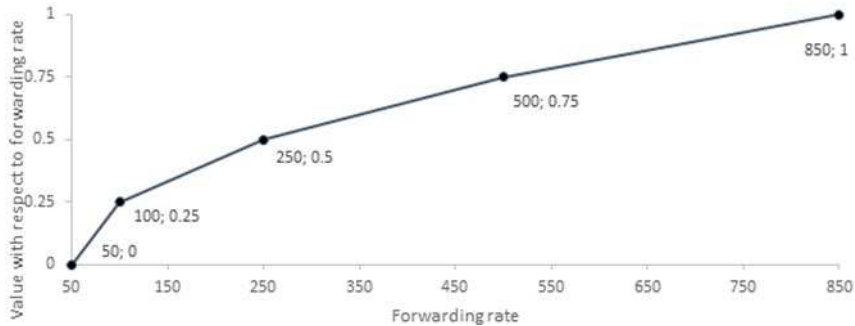


Figure 2

Piecewise linear value function for forwarding rate

The obtained increasing piecewise linear function with respect to forwarding rate is presented in Figure 2. The local alternatives' values with respect to forwarding rate are as follows: $v_{13}(\text{Alternative 3}) = 0.865$, $v_{13}(\text{Alternative 1}) = 0.359$, $v_{13}(\text{Alternative 2}) = 0.236$ and are higher than if they were obtained with monotonic linear increasing function.

To determine the decreasing linear piecewise linear function with four sections for price, the following questions were asked and answered:

Q: Which is the most preferred evaluation object x_{min} so that $v(x_{min}) = 1$?

A: The most preferred evaluation object x_{min} is 1500 €.

Q: Which is the least preferred evaluation object x_{max} associated with $v(x_{max}) = 0$?

A: The least preferred evaluation object x_{max} is 5000 €.

Q: Why is x_{min} the most preferred evaluation object and x_{max} the least preferred evaluation object?

A: Because the greater the price, the less favorable the alternative.

For the determination of sections, a question based on (1) and (2):

Q: Which is a midpoint x_1 , for which $(x_{min}, x_1) \sim (x_1, x_{max})$, where \sim indicates the decision maker's indifference between the changes in value levels, so that $v(x_1) = 0.5 v(x_{min}) + 0.5 v(x_{max}) = 0.5$?

was worded in a question that is more comprehensible to the decision-maker:

Q: Which is a midpoint x_1 , which is considered equally unfavorable if the price increases from x_{min} to x_1 , as if it increases from x_1 to x_{max} ?

A: The increase of the price from 1500 € to 2500 € is equally unfavorable as its increase from 2500 € to 5000 €.

The local value of x_1 is 0.5. So far, we determined the decreasing linear function with two sections. To obtain the decreasing linear function with four sections, the following questions based on (3), (5), (7) and (8) were put and answered:

Q: Which is a midpoint x_2 , which is considered equally unfavorable if the price increases from x_{min} to x_2 , as if it increases from x_2 to x_1 , so that $v(x_2) = 0.75$?

A: The increase of the price from 1500 € to 1800 € is equally unfavorable as its increase from 1800 € to 2500 €.

Q: Which is a midpoint x_3 , which is considered equally unfavorable if the price increases from x_1 to x_3 , as if it increases from x_3 to x_{max} , so that $v(x_3) = 0.25$?

A: The increase of the quantity from 2500 € to 4000 € is equally unfavorable as its increase from 4000 € to 5000 €.

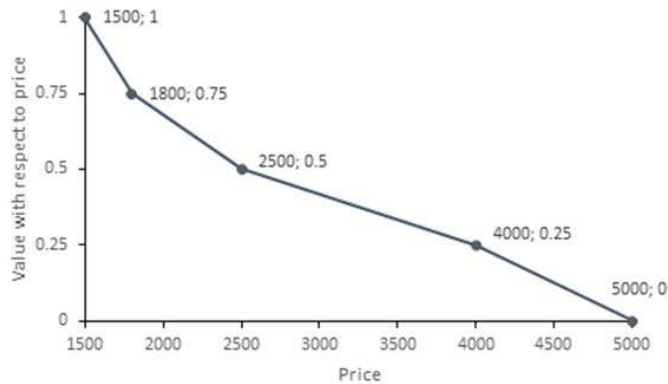


Figure 3

Piecewise linear value function for price

The obtained decreasing piecewise linear function, with respect to price, is presented in Figure 3. The local alternatives' values with respect to price are as follows: $v_{33}(\text{Alternative 2}) = 0.631$, $v_{33}(\text{Alternative 1}) = 0.5$, $v_{33}(\text{Alternative 3}) = 0.333$ and are lower than if obtained with a monotonic linear decreasing function.

The determination of value functions as a narrower professional task required the concentration and reflection of everyone in the group. The answers were written down by each participant. Then the coordinator reviewed all the answers and presented any differences to the participants. At the coordination meeting, the coordinator asked questions to provide justifications for the preferences expressed and to investigate the causes of differences, with an aim to bring the views of the participants closer. For example, in the case of increasing value function, the questions for the lower bound determinations were as follow: Why is x_{max} the most preferred evaluation object? How would the change of x_{max} affect the local alternatives' values? What do we want to achieve: greater or lesser differentiation of alternatives?

For measuring the local values of alternatives with respect to ports total, switching bandwidth, power over Ethernet, maximum power consumption and training, the monotonic-linear functions were used. In these cases, questions relating to the least and the most preferred evaluation object were put and answered. To measure the local values of alternatives with respect to power supply directly, the extreme values 0 and 1 were used. The expression of judgments on the local values of alternatives, with respect to acoustic noise and warranty was supported by pair-wise comparisons. The local values of alternatives are presented in Table 2.

Table 2
Local alternatives' values

Criterion	Alternative			Measuring Local Alternatives' Values
	Alternative 1	Alternative 2	Alternative 3	
Ports total	0.444	0.333	0.444	Value function: Lower bound: 12, Upper bound: 48
Switching bandwidth	0.064	0	1	Value function: Lower bound: 128, Upper bound: 880
Forwarding rate	0.359	0.236	0.865	Value function: Lower bound: 50, Upper bound: 850
Power over Ethernet	1	0.325	1	Value function: Lower bound: 0, Upper bound: 600
Power supply	0	0	1	Direct
Maximum power consumption	0	0.747	0.237	Value function: Lower bound: 40, Upper bound: 871
Acoustic noise	0.058	0.553	0.388	Pair-wise comparisons
Warranty	0.075	0.592	0.333	Pair-wise comparisons
Training	0.600	0.500	0.400	Value function: Lower bound: 0, Upper bound: 1000
Price	0.500	0.631	0.333	Value function: Lower bound: 1500, Upper bound: 5000

The aggregate alternatives' values obtained with an additive model [2] are presented in Table 3. The results in Table 3 show that Alternative 3 is best suited with respect to the technical criteria, and Alternative 2 is best suited with respect

to the environmental and to the economic criteria. With respect to all criteria that are structured in the hierarchy (Figure 1), the aggregate values of alternatives are as follows: $v(\text{Alternative 3}) = 0.592$, $v(\text{Alternative 2}) = 0.398$ and $v(\text{Alternative 1}) = 0.311$. It can be concluded that with respect to all criteria taken into consideration in the model presented in Figure 1, Alternative 3 is most appropriate (Table 3). The gradient sensitivity results showed that the order of alternatives to weight changes up to 0.1 is stable.

Table 3
Aggregate alternatives' values

Value with respect to:	Alternative		
	Alternative 1	Alternative 2	Alternative 3
Technical criteria	0.371	0.170	0.861
Environmental criteria	0.018	0.689	0.282
Economic criteria	0.352	0.600	0.341
All criteria	0.311	0.398	0.592

4 Discussion

The introduced systematic approach applied to a real-life problem of choosing the most appropriate IT product can be used in the IT companies that offer support to their customers.

The results in Tables 2 and 3 show that the most appropriate switch, Alternative 3, has the highest value with respect to technical criteria, too. Among the considered alternatives that are suitable for small and medium sized companies, Alternative 3 has therefore the best potential to enable communication among different networked devices in small and medium sized companies.

The presented approach to measuring local values of several IT products by value functions proved useful in the elicitation of expertly justified preferences. The determination of value functions included the coordinator with appropriate knowledge for creative thinking techniques and for MCDM, and problem experts and/or experts in the field described by the considered criterion. The engagement of the coordinator and the commitment of each expert provided the reviewed and the justified value functions. The hybrid multi-criteria and creative problem-solving approach has an application potential for other sectors, primarily for small and medium-sized enterprises, or local government decision-making.

Limitation of the measurement of local alternatives' values with value functions is the availability of numerical data, based on interval or ratio scale, of alternatives with respect to the considered criteria on the lowest hierarchy level. Further research possibility is therefore to complete the introduced systematic approach to

determine the local values of alternatives with other methods that enable dealing with data on nominal and ordinal scale, too [12] [15] [24]. In these cases, before measuring the local values of alternatives, it is necessary to define the problem requisitely holistically [8], to include comparable alternatives and to structure an appropriate set of criteria that allows for a comprehensive evaluation of alternatives.

In this paper we presented how to determine the increasing and the decreasing piecewise linear functions with four sections. The increasing or the decreasing piecewise linear functions with more than four sections can be determined according to the same principle by splitting existing sections.

In addition, several possibilities of group preference elicitation [20] [21] [26] in the step of measuring alternatives' values can be further explored in detail in the framework procedure for MCDM. Within this, the original procedure can be completed with other quantitative and qualitative methods, with an emphasis to several creative problem-solving methods for problem definition.

Conclusions

Piecewise linear functions are distinguished by simplicity and representativeness. To meet the first goal of our work herein, we defined how to determine piecewise linear increasing and decreasing linear functions, with four sections, by using the bisection method. The resultant increasing or decreasing piecewise linear functions, determined by using a hybrid approach that is proposed in this paper, can serve as a good approximation of exponential value functions, that would otherwise, require a large series of data and a demanding statistical knowledge base. Moreover, considering the expressed expert preferences, the approach also allows the creation of value functions, whose form deviates from simple monotonic-linear or exponential value functions. The simple monotonic-linear functions are easier to work with, but they might not be representative in non-linear cases. On the other hand, Rezaei [23] showed that exponential value functions might have a better representativeness, than simple linear functions, however, it is difficult for a practitioner to estimate a value for the shape parameter of the exponential value functions and cannot be easily interpreted.

Within the frame of the procedure for MCDM, we explored the possibilities of measuring alternatives' values and within this, recommended the original systematic approach, that includes both the quantitative and qualitative methods. The described approach is based on the "six questions" technique – a creative problem-solving qualitative method, which is usually used to define problems. The novelty of this paper is in the extension of the use of the six questions technique, to the measurement of the local alternatives' values, which has usually been seen as a quantitative step in the frame procedure for MCDM. By introducing the systematic approach to determine the local values of alternatives, by using the six questions technique, we met the second goal of this paper. A practical case has proven that the six questions technique can adopted in

group preference elicitation and thus, adequately supports the step of measuring local alternatives' values.

The practical case presented in this work is limited to choosing the most appropriate switch for small and medium sized companies, in the current era of digitalization, it is an important IT product. Further application possibilities of the presented approach can be extended to a wide range of organizational and management problems, for the selection, assessment and evaluation of various alternatives.

Acknowledgement

This work was supported by the Slovenian Research Agency under research program Entrepreneurship for Innovative Society [P5-0023].

References

- [1] E. Beinat, *Value Functions for Environmental Management*, Kluwer Academic Publishers, Dordrecht, 1997
- [2] V. Belton and T. J. Stewart, *Multiple Criteria Decision Analysis: An Integrated Approach*, Kluwer Academic Publishers, Dordrecht, 2002
- [3] G. P. Boulden, *Thinking creatively*. Dorling Kindersley Limited, London, 2002
- [4] Cisco, *Cisco Catalyst 1000 Series Switches Data Sheet*, 2021, Retrieved April 26, 2021, from <https://www.cisco.com/c/en/us/products/collateral/switches/catalyst-1000-series-switches/nb-06-cat1k-ser-switch-ds-cte-en.html>
- [5] P. Cook, *Best Practice Creativity*, Gower Publishing Limited, Hampshire, 1998
- [6] V. Čančer, Criteria weighting by using the 5Ws & H technique, *Business Systems Research*, Vol. 3, No. 2, pp. 41-48, 2012
- [7] V. Čančer, A frame procedure for multiple criteria selection of IT products and services, *Analele științifice ale Universității "Al.I. Cuza" din Iași, Științe economice*, Vol. 60, No. 1, pp. 94-106, 2013
- [8] V. Čančer and M. Mulej, The Dialectical Systems Theory's Capacity for Multi-Criteria Decision-Making, *Systems Research and Behavioral Science*, Vol. 27, No. 3, pp. 285-300, 2010
- [9] V. Čančer and M. Mulej, Multi-criteria decision making in creative problem solving, *Kybernetes*, Vol. 42, No. 1, pp. 67-81, 2013
- [10] Dell Inc., *Dell EMC PowerSwitch N1500 Series Switches*, 2021, Retrieved April 26, 2021, from https://www.dell.com/en-us/work/shop/povw/networking-n1500-series#features_section

-
- [11] W. Edwards, How to use multiattribute utility measurement for social decisionmaking, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 7, No. 5, pp. 326-340, 1977
- [12] A. Fernandez, P. Zaraté, J. C. Gardey, and G. Bosetti, Supporting multi-criteria decision-making across websites: the Logikós approach, *Central European Journal of Operations Research*, Vol. 29, No. 5, pp. 201-225, 2021
- [13] A. Fernández-Portillo, M. Almodóvar-González, and R. Hernández-Mogollón, Impact of ICT development on economic growth, A study of OECD European union countries, *Technology in Society*, Vol. 63, Article 101420, pp. 1-9, 2020
- [14] M. Ghaderi and M. Kadziński, Incorporating uncovered structural patterns in value functions construction, *Omega*, Vol. 99, Article 102203, 2021
- [15] R. Ginevicius, D. Gedvilaite, A. Stasiukynas, K. Suhajda, Complex Expert Assessment of the State of Business Enterprises, *Acta Polytechnica Hungarica*, Vol. 19, No. 2, pp. 135-150, 2022
- [16] S. Greco, M. Kadziński, and R. Słowiński, Selection of a representative value function in robust multiple criteria sorting, *Computers & Operations Research*, Vol. 38, No. 1, pp. 1620-1637, 2011
- [17] Hewlett Packard Enterprise, *Aruba 6300 Switch Series*, 2021, Retrieved April 26, 2021, from <https://h20195.www2.hp.com/v2/gethtml.aspx?docname=a00073540enw>
- [18] M. Kadziński, J. Badura, and J. R. Figueira, Using a segmenting description approach in multiple criteria decision aiding, *Expert Systems With Applications*, Vol. 147, Article 113186, 2020
- [19] R. L. Keeney and H. Raiffa, *Decisions with multiple objectives: preferences and value tradeoffs*, Cambridge University Press, 1999
- [20] S. Nikou, J. Mezei, and P. Sarlin, A Process View to Evaluate and Understand Preference Elicitation, *Journal of Multi-Criteria Decision Analysis*, Vol. 22, No. 5-6, pp. 305-329, 2015
- [21] J. Pictet and D. Bollinger, The silent negotiation or how to elicit collective information for group MCDA without excessive discussion, *Journal of Multi-Criteria Decision Analysis*, Vol. 13, No. 5-6, pp. 199-211, 2005
- [22] H. Raiffa, The prescriptive Orientation of Decision making: A Synthesis of Decision Analysis, Behavioral Decision Making, and Game Theory, In S. Rios (Ed.), *Decision Theory and Decision Analysis: Trends and Challenges*, pp. 3-13, Springer, Dordrecht, 1994
- [23] J. Rezaei, Piecewise linear value functions for multi-criteria decision-making, *Expert Systems With Applications*, Vol. 98, pp. 43-56, 2018

-
- [24] P. Rezaei, K. Rezaie, S. Nazari-Shirkouhi, M. R. J. Tajabadi, Application of Fuzzy Multi-Criteria Decision Making Analysis for Evaluating and Selecting the Best Location for Construction of Underground Dam, *Acta Polytechnica Hungarica*, Vol. 10, No. 7, pp. 187-205, 2013
- [25] M. Segura and C. Maroto, A multiple criteria supplier segmentation using outranking and value function methods, *Expert Systems With Applications*, Vol. 69, pp. 87-100, 2017
- [26] M. Škoda, M. Flegl, and C. Lozano, Fuzzy approach for group decision-making in crisis situations, *Business: Theory and Practice*, Vol. 22, No. 1, pp. 180-189, 2021
- [27] D. Vukovič and V. Čančer, The multi-criteria model for financial strength rating of insurance companies, In L. Zadnik Stirn and S. Drobne (Eds.), *Proceedings of the 9th International Symposium on Operational Research SOR '07*, pp. 109-114, Slovenian Society Informatika, Section for Operational Research, Ljubljana, 2007
- [28] D. von Winterfeldt and W. Edwards, *Decision analysis and behavioral research*, Cambridge University Press, 1986
- [29] S. Yang, P. Fichman, X. Zhu, M. Sanfilippo, S. Li, and K. R. Fleischmann, The use of ICT during COVID-19. *Proceedings of the Association for Information Science and Technology*, Vol. 57, No. 1, Article e297, 2020

Segmentation of Moiré Fringes of Scoliotic Spines Using Filtering and Morphological Operations

**Csaba Bogdán¹, Andor Dániel Magony¹, Wolfgang Birkfellner²,
Ákos Antal³, Miklós Tunyogi-Csapó⁴**

¹ Institute of Transdisciplinary Discoveries, Medical School, University of Pécs, Ifjúság útja 11, H-7624 Pécs, Hungary; csaba.bogdan@pte.hu; magony.andor@pte.hu

² Center for Medical Physics and Biomedical Engineering, Medical University of Vienna Spitalgasse 23, A-1090 Vienna, Austria; wolfgang.birkfellner@meduniwien.ac.at

³ Department of Mechatronics, Optics and Mechanical Engineering Informatics, Budapest University of Technology and Economics, Bertalan Lajos u. 4-6, H-1111, Budapest, Hungary; akos@mogi.bme.hu

⁴ National Center for Spinal Disorders, Buda Health Center, Királyhágó u. 1-3, H-1126 Budapest, Hungary; miklos.tunyogi@bhc.hu

Abstract: For reducing uncertainties in moiré pattern analysis, an accurate segmentation of moiré fringes (MF) is vital. In this study, an algorithm for segmenting MFs of scoliotic spines was provided in MATLAB® environment by an empirically established sequence of filtering and morphological operations defined by static function parameters: (1) contrast enhancement, (2) brightness increase, (3) contrast refinement, (4) 2-D Gaussian filter, (5) dilation, (6) thresholding, (7) skeletonization. The algorithm is simple, fast to process and, for the most part of the images, follows the MFs correctly. Further research on segmenting MFs is quite promising by improving the algorithm with adaptive and dynamic solutions, and exploring ways to replace time demanding and complex image processing techniques.

Keywords: moiré fringe segmentation; moiré method; shadow moiré; projection moiré; digital moiré; scoliosis

1 Introduction

Diagnosing spinal deformities has long been in the focus of medicine. Measurement of bending disorders of the vertebral column, such as scoliosis, is usually performed on radiographs by calculating the Cobb angle. Disadvantages of

X-ray imaging such as cost, time and repetition demands, tools and environmental conditions required and radiation exposure imparted to the patient, are not negligible, and justify methodological research of such moiré technique or moiré topography, that can lead to fast, cost-effective and non-ionizing diagnostic imaging of the spine.

The phenomenon of moiré [mwàkə] or moiré effect was elevated to scientific status, first, by Lord Rayleigh dealing with diffraction gratings in 1874 [1]. Lord Rayleigh concluded that moiré might be made useful for measurement purposes. About 100 years later, beyond industrial applications, further research proposed moiré topography for measurement of the human body [2], [3]. From the 1980s, MT is used on the whole human body, including oral cavity [4], bones [5], teeth and the skeletal system [6].

Moiré refers to an irregular wavy surface the pattern of which changes in accordance with its movement and can be observed, if two or more structures with similar geometry (nearly identical arrays of lines or dots) overlap, producing alternating bright and dark fringes (Fig. 1). Usually, dark fringes are called moiré fringes (MFs) or moiré stripes, but we can also consider bright ones as moiré surfaces when examining moiré images (MIs)—it is only a matter of agreement [7]. Based on the physical phenomenon of moiré, moiré techniques are defined as a group of methods usually used for surface mapping and shape or deformation measurement. In the scientific literature, shadow moiré and projection moiré techniques (SMT and PMT, respectively) seem to be the two primary methods of MT used for measurement of the human body (Fig. 2-3) [8].

An algorithm based on MT that proved to be suitable for calculating the curvature angle of the spine, may complement or substitute X-ray images—especially in follow-up examinations. The workload required for segmentation and evaluation of MIs is, however, not inconsiderable; some researchers see the best solution for that in an automatic system [8], [9], [10], [11], [12], [13]. And yet, processing of MIs requires several unique solutions that are especially influenced by optical arrangement, applied illumination (effect on intensity distribution), nature of noise and detection. Therefore, implementing a fully automated image analysis is a challenging, nevertheless desired objective in the field [7], [14]. For reducing uncertainties in moiré pattern analysis, an accurate segmentation of MFs is vital. The present study aims to contribute to the segmenting phase of MI analysis of scoliotic spines by providing an algorithm of filtering and morphological operations. This study presents the initial phase of the research, where static function parameters were applied to identify a possible segmenting sequence for MFs. Segmentation with adaptive and dynamic function parameters is not the subject of the present study but can be expected in a later phase of the research.

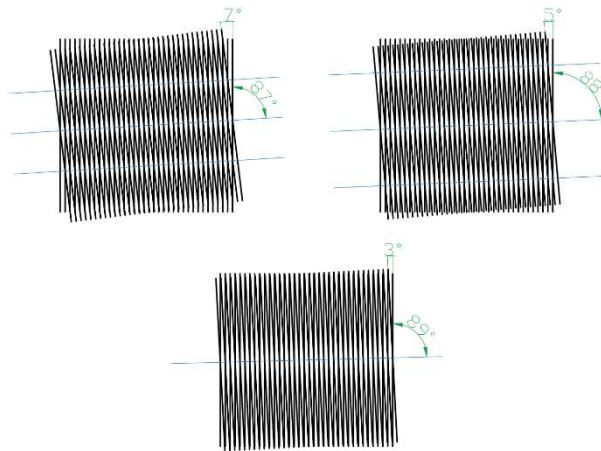


Figure 1

Moiré fringes of identical grids with different angle deviations

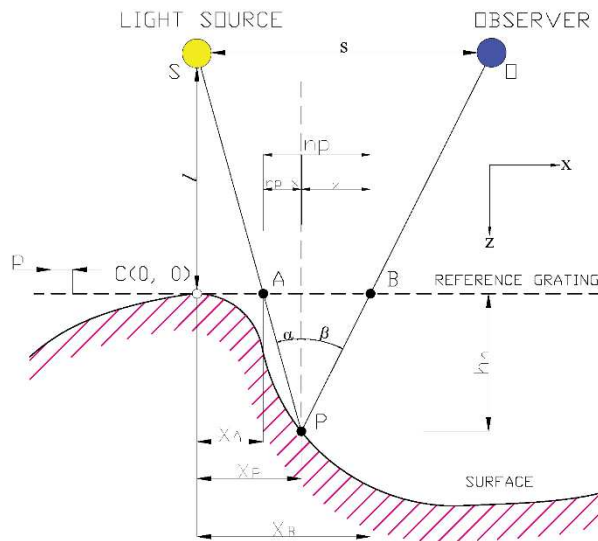


Figure 2

The base principle of shadow moiré technique

- (A) Point where source light S passes through the reference grating
- (B) Point where reflected source light S passes through the reference grating
- (C) Origin
- (O) Observer
- (p) Grating pitch (or period of the grating)
- (P) Measured point
- (S) Light source

- (α) Incidence angle (of incoming light)
- (β) Viewing angle
- (h_n) Depth of n^{th} -order moiré pattern measured from the reference grating
- (l) Distance of S and O from the reference grating surface
- (n) Order of the moiré pattern
- (s) Interseparation of S and O
- (X_A) X component of SA distance
- (X_B) X component of SB distance
- (X_P) X component of SP distance
- (x) X component of PB distance

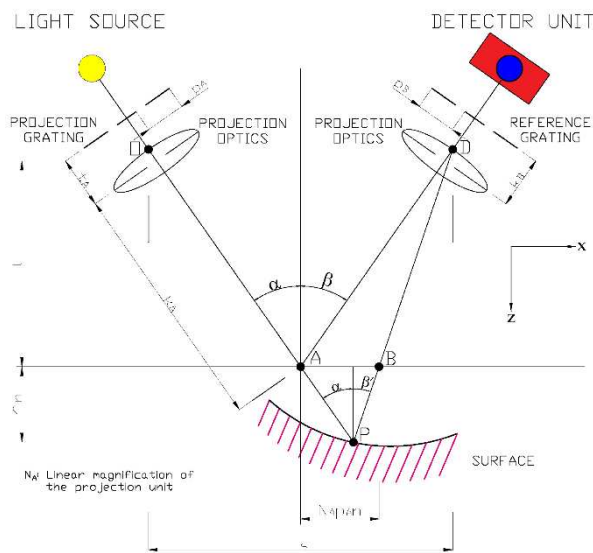


Figure 3

The base principle of projection moiré technique

- (A) Point where source light passes through the reference plane
- (B) Point where reflected source light passes through the reference plane
- (D) Projection optics on the detector's side
- (h_n) Depth of n^{th} -order moiré pattern measured from the reference plane
- (k_A) Distance between A and projection optics on the illumination's side
- (k_B) Distance between the reference grating and optics on the detector's side
- (l) Distance of projection and reference grating measured from the reference plane
- (n) Order of the moiré pattern
- (N_A) Linear magnification of the projection unit
- (O) Projection optics
- (P) Measured point
- (p_A) Pitch of projection grating

- (p_B) Pitch of reference grating
- (s) Distance between the projection optics O and D
- (t_A) Distance between the projection grating and optics on the illumination's side
- (α) Incidence angle (of incoming light)
- (β) Viewing angle at A
- (β') Viewing angle at P

2 Materials and Methods

The segmenting algorithm was developed in MATLAB[®] version 9.5.0.944444 (R2018b) [15] based on exploratory sequences and observations performed on 11 MIs made available by SALUS Ortopédtechnika Kft. Original MIs were software-generated applying digital PMT, and have a resolution of 2448 x 3264 px in 10 cases and 2736 x 3648 px in 1 single case with a bit depth of 24 and DPI of 96. MIs show patients in standing position facing the reference wall, and non-ROI areas that typically contain irrelevant image content of residual grating and non-moiré parts outside the back region. The procedure of image processing involves two main phases: (1) preprocessing MIs, (2) segmenting MFs using a sequence of filtering and morphological operations. Based on operation parameters applied in phase (2), two MI groups (G1, G2) are distinguished. For MIs in G1, identical parameters of processing phases result the similar outcome. For G2, different parameters are required for similar results, even in the group itself (Table 1). The reason for this is to be found in different image characteristics in terms of contrast, noise and moiré blur/confluence that are likely to result from dissimilar measurement setups. The phases of the code are illustrated in MI no. 2 of G1 (Fig. 4-14). For the other images, only the result of the segmentation is given in the overlap with grayscale ROIs (Fig. 15-20).

2.1 Preprocessing Moiré Images

In the first phase of image processing, original MIs are prepared for further processing by (a) manual selection of region of interests (ROIs) and (b) conversion into grayscale. As a result of these steps, input images for phase 2 are obtained. Selection of ROIs is performed by rectangular cropping using *imcrop* function. Original images are reduced in size based on selected ROI area with boundaries from the neck to the hip; and from the left upper arm to the right one (Fig. 4). Cropped images vary in size according to the size of specific ROI area and patient setup (distance from the camera). The resolutions of cropped images are between the range of 784x1044 px and 1136x1406 px.



Figure 4
Original (L) and ROI-cropped (R) moiré image

Function *rgb2gray* is applied in order to convert cropped RGB images into grayscale (Fig. 5). For morphological operations discussed in phase (2), grayscale images contain all the relevant information of moiré stripes. Colour intensity values do not carry additional information here, and their conversion to grayscale does not induce data loss.

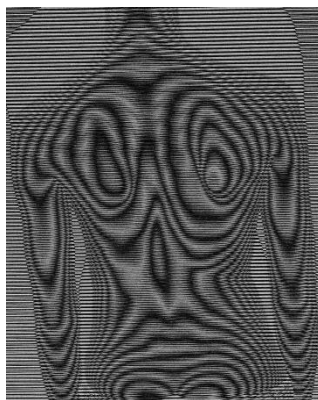


Figure 5
Grayscale moiré image with ROI

2.2 Steps of Moiré Fringe Segmentation

For segmenting MFs, a possible algorithmic approach is introduced. In the segmentation procedure, an empirically established sequence of filtering and morphological operations is used: (1) enhancing contrast, (2) increasing brightness, (3) refining contrast, applying (4) 2-D Gaussian filter and (5) dilation, (6) thresholding and (7) skeletonization.

Table 1
Summary of function parameters applied to moiré image (MI) groups G1 and G2

Phases	G1 (MI no. 1, 2, 3, 5, 6, 7, 9, 10)	G2 (MI no. 4, 8, 11)
Enhancing contrast [<i>imadjust</i> by def.]	3x	3x
Increasing brightness [increase in pixel value]	120	100 for MI no. 4 and 8 110 for MI no. 11
Refining contrast [range of intensity values]	[0.7 1]	[0.7 1]
2-D Gaussian filter [standard deviation]	6	6 for MI no. 4 and 11 8 for MI no. 8
Dilation [structuring element, pixel width, repetition]	square, 3, 3	square, 3, 3
Threshold value	0.41	0.35 for MI no. 4 and 11 0.38 for MI no. 8

2.2.1 Enhancing Contrast

For mapping the intensity values to new values and thereby enhancing contrast, *imadjust* function was applied in 2 steps. In the first step a slight contrast correction is performed via saturating the bottom 1% and the top 1% of all pixel values three times by the default operation of *imadjust* (Fig. 6).

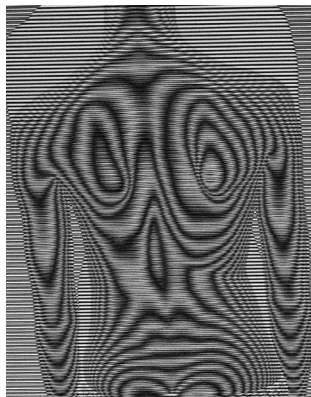


Figure 6

Result of contrast enhancement applied on grayscale moiré

2.2.2 Increasing Brightness

Before further contrast enhancement, brightness is increased by increasing pixel values of MIs of G1 by 120 and G2 by 100 and 110, respectively (Fig. 7).

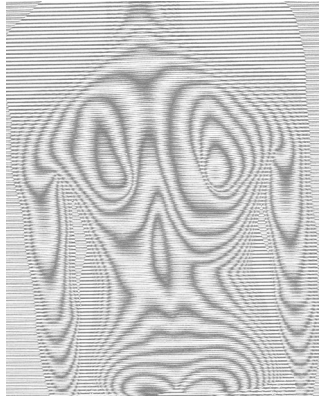


Figure 7
Result of brightness increase

2.2.3 Refining Contrast

As the second step in contrast adjustment, intensity values are mapped to new values between 0.7 and 1, by re-applying the function *imadjust*. The result is a stronger contrast of moiré stripes (Fig. 8).

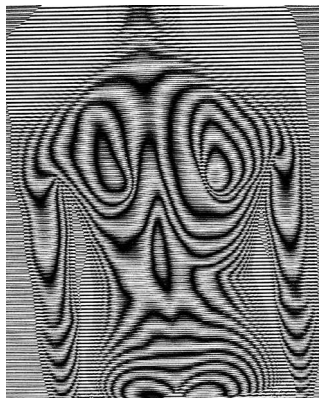


Figure 8
Result of contrast refinement

2.2.4 Applying 2-D Gaussian Filter

Image is filtered with a 2-D Gaussian smoothing kernel (*imgaussfilt*) with image specific standard deviation values 6 or 8 (Fig. 9).

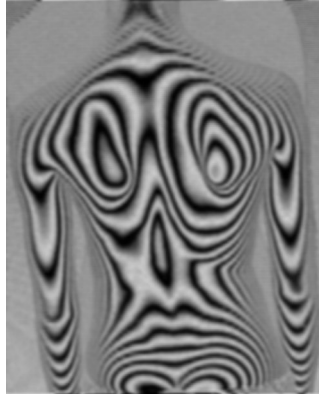


Figure 9
Result of 2-D Gaussian filtering

2.2.5 Applying Dilation

Blurred grayscale image is dilated by using *imdilate* and morphological structuring element *strel*. Dilation is applied three times in combination of square structuring element with a width of 3 pixels (Fig. 10). In Fig. 10, resulting differences from the previous step (2-D Gaussian filter) may not be obvious, however, dilation supports the results of thresholding and skeletonizing operations by gradually enlarging the boundaries of regions of foreground pixels.

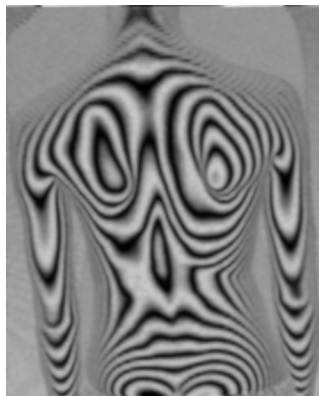


Figure 10
Result of dilation

2.2.6 Thresholding

Function *imbinarize* is used to convert the image to a binary image, based on a threshold value: all pixel values above a globally determined threshold are replaced with ones and all other values with zeros. The threshold value applied is covered in the range between 0.35 and 0.41 (Fig. 11).



Figurer 11
Result of thresholding

2.2.7 Skeletonization

Skeletonization was performed by using *bwmorph* function for morphological operations, specified as 'thin'. The operation 'thin' is applied for thinning binarized images to single lines, and repeated until the image no longer changes (i.e. parameter 'Inf' in the function input). The inputs of *bwmorph* are provided by complements of binarized images generated with the function *imcomplement*. For a better visualization of the results, lines of skeletonized images are fattened by applying the operation 'fatten' once. The result is shown enlarged as re-complemented image (Fig. 12) and as overlay on binarized (Fig. 13) and colour ROI images (Fig. 14).

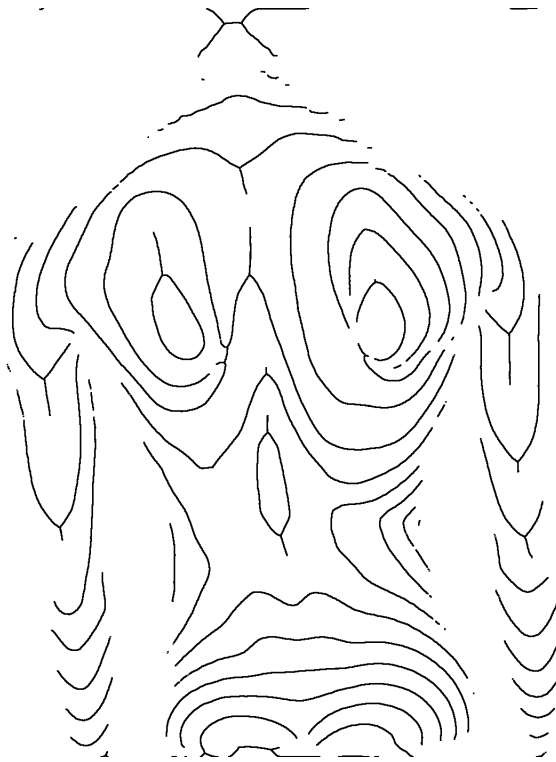


Figure 12

Result of skeletonization: segmented moiré fringes



Figure 13

Overlay of segmented moiré fringes (green) on thresholded image (black)

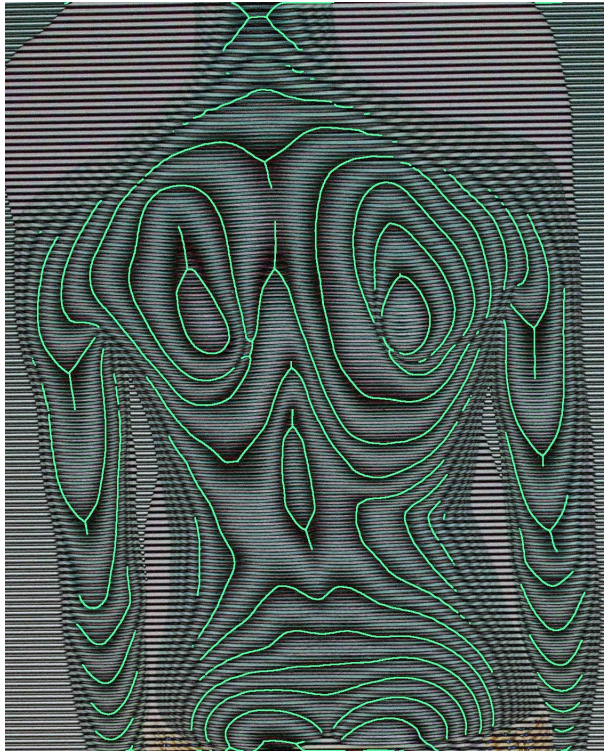


Figure 14

Overlay of segmented moiré fringes on colour ROI

3 Results and Discussion

Results show that the segmenting method is simple, fast to process and, for the most part of the image, follows the moiré stripes accurately (Fig 15-20). The average processing time of the algorithm from grayscale conversion to skeletonization is 0.146 s per image. In Table 2, the time course of the algorithm, with exclusion of manual ROI selection, is summarized in average elapsed and processing time.

The segmenting algorithm leads to a partial or sporadic segmentation of moiré stripes. Image details (i.e. parts of moiré stripes) and accuracy are lost mainly due to original fringe quality and characteristics such as (a) pale—mostly around the shoulders and the waist—, (b) convergent, (c) wider/blurred MFs, (d) image noise caused mostly by residual grating, and (e) unwanted branches generated by skeletonizing operation (Fig. 21). In terms of general usability of the code, the

necessity of re-adjusting static function parameters is challenging and—especially in higher patient populations—time consuming. To get similar results on MIs other than the 11 sample images applied in this study, function parameters need to be empirically determined and implemented in the code. Therefore, static function parameters are desired to be eliminated with automatic adaptive and dynamic parameterization. Also, for problems (a-e), a more sensible solution is required. A possible way for further research might be to improve the algorithm with (1) adaptive and dynamic function parameters based on values of low contrasted and over contrasted images in combination with adaptive thresholds, (2) applying high-pass filters for image sharpening. Another possible direction of research is (3) to combine the algorithm with segmentation approaches based on a fuzzy inference system [16]. Due to its simplicity and fast operation, an improved solution of the algorithm could also replace time demanding and complex segmentation methods.

Table 2
Time course of the segmenting algorithm

Proc. Nr.	Process	Average Time [sec]*	
		<i>Elapsed</i>	<i>Processing</i>
1	Manual selection of ROI	excl.	excl.
2	Determining object class	0.02567	0.02567
3	Duplicating image for reference	0.02593	0.00026
4	Converting in grayscale	0.02717	0.00124
5	Enhancing contrast	0.03990	0.01272
6	Increasing brightness	0.04036	0.00046
7	Refining contrast	0.04348	0.00312
8	2-D Gaussian filter	0.05781	0.01433
9	Dilation	0.06195	0.00414
10	Thresholding	0.06901	0.00705
11	Skeletonization	0.17165	0.10264
12	Saving images in <i>.png</i> files*	0.61555	0.44390
13	Visualizing results	0.73376	0.11821

*System used: CPU: Intel® Core™ i5-8300H @ 2.30 GHz,
GPU: NVIDIA GeForce GTX 1050 (4 GB), RAM: 8 GB

**Output images are saved as skeletonized moiré contours (transparent and white-backgrounded) and its overlays on binarized and input images.

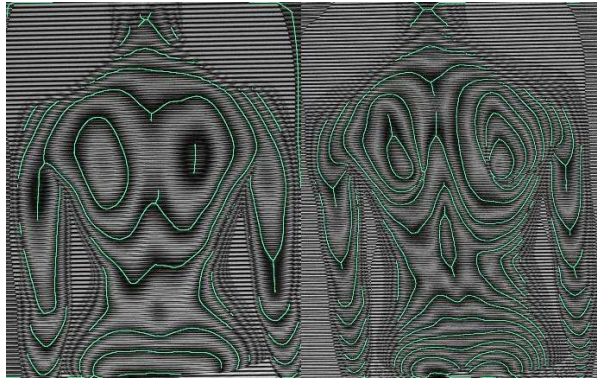


Figure 15
Segmented moiré fringes of image no. 1 (L) and 2 (R)

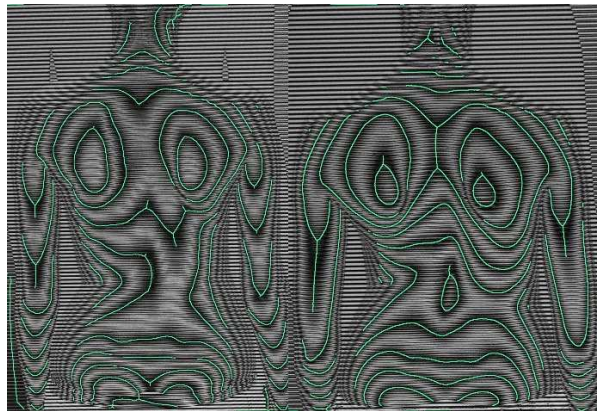


Figure 16
Segmented moiré fringes of image no. 3 (L) and 10 (R)

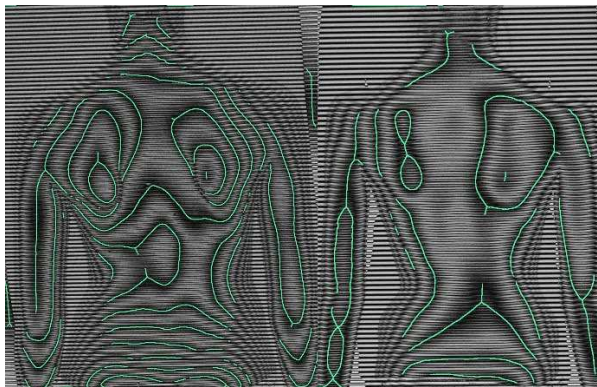


Figure 17
Segmented moiré fringes of image no. 5 (L) and 6 (R)

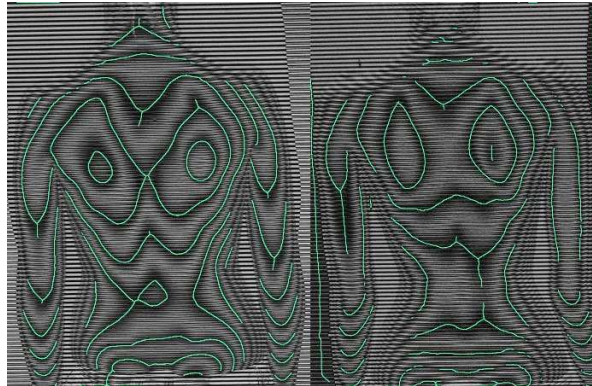


Figure 18
Segmented moiré fringes of image no. 7 (L) and 9 (R)

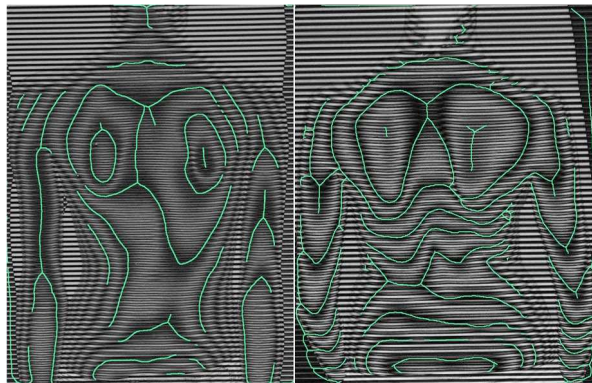


Figure 19
Segmented moiré fringes of image no. 4 (L) and 8 (R)

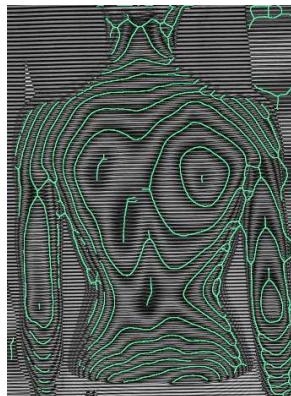


Figure 20
Segmented moiré fringes of image no. 11

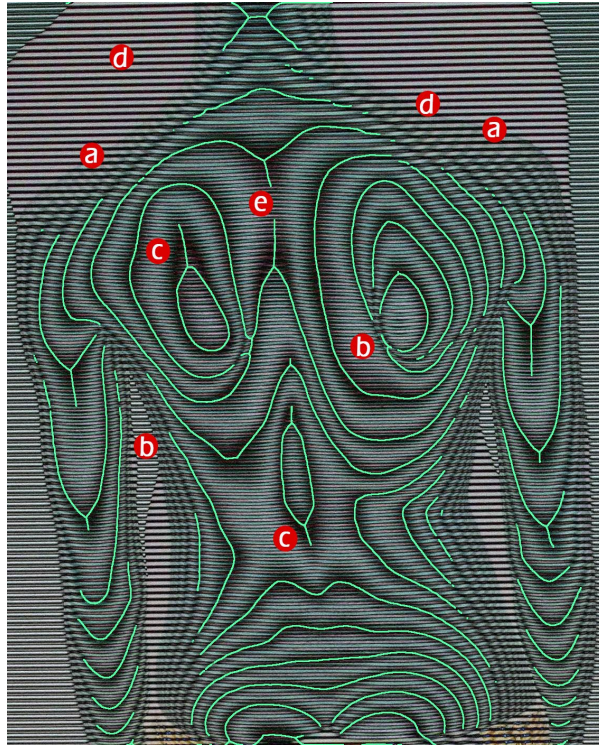


Figure 21

Problems to be handled in moiré fringe segmenting algorithm
(a-c) pale, convergent, and wider/blurred moiré fringes
(d) noise of residual grating
(e) unwanted branches

Conclusion

In this study, an initial phase of research on MF segmentation of scoliotic spines is conducted, presenting an algorithmic sequence of filtering and morphological operations with static function parameters in MATLAB® environment. The applicability of the algorithm is confirmed by a simple, fast to process and, for the most part of sample images, accurate MF segmentation. The results indicate that the algorithm introduced constitutes a suitable base for further research on segmenting MFs with adaptive and dynamic function parameters and adaptive image processing solutions and replacing time demanding and complex image processing techniques.

Acknowledgment

We would like to express our gratitude to Salus Orthopedtechnika Kft. (István Joó, Katalin Prommer, Ferenc Marlok) for providing us sample images and granting permission to our study.

References

- [1] J. W. S. Rayleigh. On the manufacture and theory of diffraction gratings. *Philosophical Magazine*, 47:81-93, 1874
- [2] H. Takasaki. Moiré topography. *Applied Optics*, 9(6):1467-72, 1970
- [3] H. Takasaki. Moiré topography. *Applied Optics*, 12(4):845-52, 1973
- [4] Takei T., Yokosawa S., Innami T., Miyazawa T., Kuwana T., Takagi M., Furukawa A. Application of Moiré Topography to Forensic Odontology. *The Journal of Nihon University School of Dentistry*, Vol. 27(2), 1985, 87-104
- [5] Wood, J. D.; Wang, R.; Weiner, S.; Pashley, D. H. Mapping of tooth deformation caused by moisture change using moiré interferometry. *Dent. Mater.*, Vol. 19(3), 2003, 159-166
- [6] Jelen K., Kusová S. Pregnant women: Moiré contourgraph and its semiautomatic and automatic evaluation. Dedicated to Professor Antonín Dolezal on his 75th birthday anniversary. *Neuro endocrinology letters*, Vol. 25, 2004, 52-56
- [7] Á. Antal. Optikai úton generált Moiréfelületek hibaanalízise és identifikálása mérés technikai alkalmazásokkal [dissertation on the internet]. Budapest, Hungary, University of Technology and Economics, 2009 [cited 2021 February 24] Available from: <https://repozitorium.omikk.bme.hu/bitstream/handle/10890/851/ertekezes.pdf?sequence=1&isAllowed=y>
- [8] F. Porto, J. L. Gurgel, T. Russomano, P. T. V. Farinatti. Moiré Topography: From Takasaki Till Present Day. In Grivas T. B., Editor: *Recent advances in scoliosis*, pp. 103-118, 2012
- [9] M. Batouche, R. Benlamri. A computer vision system for diagnosing scoliosis. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, 3, pp. 2623-2628, 1994
- [10] H. S. Kim, S. Ishikawa, Y. Ohtsuka, H. Shimizu, T. Shinomiya, M. A. Viergever. Automatic scoliosis detection based on local centroids evaluation on Moiré topographic images of human backs. *IEEE Trans Med Imaging*, 20(12):1314-1320, 2001
- [11] T. Ikeda, H. Terada. Development of the moiré method with special reference to its application to biostereometrics. *Opt. Laser Technol.*, 12:302-306, 1981
- [12] H. Kim, H. Ushijima, S. Ishikawa, Y. Otsuka, H. Shimizu, T. Shinomiya, M. A. Viergever. Scoliosis detection based on difference of apexes position and angle on Moiré topographic images. *International Congress Series*, 1268:1294, 2004

- [13] P. Balla, G. Manhertz, Á. Antal. Diagnostic moiré image evaluation in spinal deformities. *Optica Applicata*, 46:375-385, 2016
- [14] H. Zhi, R. B. Johansson. Interpretation and classification of fringe patterns. *Optics and Lasers in Engineering*, 17(1):9-25, 1992
- [15] MATLAB (2018) 9.5.0.944444 (R2018b). Natick, Massachusetts: The MathWorks Inc.
- [16] H. W. Wing, Y. S. Kin. Moiré Fringe Segmentation Using Fuzzy Inference System. In: H. Ibrahim, S. Iqbal, S. S. Teoh, M. T. Mustaffa, Editors: *9th International Conference on Robotic, Vision, Signal Processing and Power Applications*. Lecture Notes in Electrical Engineering, 398. Springer, Singapore, 2017

A Systematic Approach for Identification of SOPTD Processes using a Relay Feedback with a Fractional Order Integrator

**Rahul B. Gaikwad, Raghu Raja Pandiyan K, R.
Ranganayakulu, G. Uday Bhaskar Babu**

Department of Chemical Engineering, National Institute of Technology Warangal, Telangana, India – 506004, grahul6@student.nitw.ac.in, raghuraj@nitw.ac.in, rayalla.ranga@student.nitw.ac.in, udaybhaskar@nitw.ac.in

Abstract: In this work, a new systematic identification approach is proposed to obtain second-order plus time delay (SOPTD) models by relay feedback with hysteresis and fractional order integrator. A relay with hysteresis and fractional order integrator is used to generate sustained oscillation at process output for model identification. The addition of a fractional order integrator helps improve the position frequency point obtained by the Describing function (DF) method and thus leads to accurate model. The proposed approach has an additional degree of freedom for estimating parameters. In addition, the proposed relay test was performed in the presence of measurement noise. The proposed method was applied to overdamped, underdamped and critically damped transfer function models. The performance of the proposed approach is evaluated by comparing the Integral absolute error (IAE) criteria in the frequency domain, Nyquist plot, and step response. Compared with the literature method, the proposed approach reduced IAE for Overdamped, Overdamped, Underdamped, and critically damped processes by 77.68%, 68.34%, 98.57%, and 95.78%, respectively. The simulation results show that the proposed approach identifies satisfactory models compared to existing techniques.

Keywords: SOPTD; model identification; Fractional order integrator; Describing function (DF); Relay feedback with hysteresis

1 Introduction

Most of the process industries use Proportional Integral Derivative (PID) controllers and their tuning largely depends on identification of a good process model. The process model can be identified by open-loop and closed-loop methods. In the open-loop method, introduce an excitation in each input variable one at a time of the process to get the output responses. Then, the transfer function model is to be identified using the output responses. The open loop identification method is simple, but it has some drawbacks, i.e., sensitive to disturbances, more computational time,

and sometime process output deviates from the set point. Closed-loop identification methods overcome the drawbacks of open loop identification. The relay feedback identification method based on the closed loop test has gained interest for tuning PID controllers because of its simplicity. Relay feedback is one of the promising tools for the identification of process models. The theory behind relay feedback identification is straightforward as a Ziegler-Nichols (Z-N) closed-loop test. The relay feedback method uses the relay instead of the controller (Figure 1); thus, the system generates sustained oscillations called a limit cycle. This limit cycle gives valuable process information, i.e. peak amplitude and frequency. Hence, by using the limit cycle information, the process model parameters are estimated.

Some chemical processes with higher order dynamics may not be satisfactorily described by first-order plus time delay (FOPTD) models but more accurately described by SOPTD models. The relay feedback technique was first introduced to tune the PID controller [1-3]. Using Laplace transforms used asymmetric relay for process identification in the frequency domain [4]. Developed the identification method using relay data and state-space approach to derive nonlinear equations for various lower and higher order process models [5]. The relay with hysteresis was used to generate a limit cycle at process output. Identification was carried out offline and online using the DF method [6-9]. The state-space method and relay with hysteresis identifies stable and unstable processes to estimate the unknown process model parameters [10] [11].

Time domain-based analytical expressions are emanated to assess the exact model parameters using a relay with hysteresis for non-minimum phase (NMP) processes [12]. Proposed a method based on Fourier series analysis, like a DF method using an ideal relay with a fractional order integral. A comparative study of different relay identification techniques has been conducted [13]. Nonlinear equations for non-zero set points were developed and identified as first and second-order process models [14]. DF method was used to determine the higher-order and NMP process models as first and second-order models [15]. The limit cycle information near the non-zero set point was used to derive mathematical equations for accurately identifying unknown plants [16]. Novel explicit expressions are proposed to identify stable, unstable and integrating first order plus dead time processes. An asymmetrical relay generates a smooth limit cycle at the output [17]. The frequency domain and state space approach was proposed for modeling and identifying non-minimum phase processes [18]. After the relay feedback experiment, a set of explicit expressions was derived for identifying unknown FOPTD and SOPTD models [19].

A new “shifting method” was introduced recently in the literature to estimate three points on Nyquist plot of an unknown process from limit cycle data generated by biased relay with hysteresis. Now, optimization technique is used to identify anisochronic and isochronic models by minimizing the error between identified model and actual model with reference to the three points [20] [21]. The shift method was extended by developing explicit formulas for identifying isochronic process model [22]. The shift method was modified by adding an integrator or time

delay in relay feedback loop to identify stable, unstable, higher order, integrating and NMP processes [23-25]. There were two methods namely closed-loop test with proportional controller and unbiased ideal relay feedback test along with the usage of Lamber W function for calculating unknown process parameters [26].

Although the relay feedback technique for process identification has been widely addressed, much scope still exists to improve the developments. In particular, in this paper, we have investigated and contributed to the following: Simple analytical expressions based on the DF technique are derived for identifying the SOPTD transfer function models. A single relay with hysteresis and fractional order integrator is used in a closed loop to extract process information and reduce measurement noise's effect. The additional fractional order integrator helps improve the DF method's frequency point. The proposed approach provides flexibility in the degree of freedom and thus leads to more accurate models. Since measurement noise is a critical issue in process industries, the validity of the proposed method is illustrated in noisy environments. As relay with hysteresis and fractional order integrator reduces the effect of noise, the Fourier series-based curve fitting technique is appended to obtain noise-free process output. The accurate model was identified based on the minimum IAE. Furthermore, the effect of fractional order integrator on model parameters is studied.

This paper considered four examples of SOPTD process from works of literature. The results are compared based on the integral absolute error criteria in the frequency domain, Nyquist plot, and step response between the proposed model, actual process, and methods present in literature with and without noise. MATLAB Programming/Simulink environment used for all experiments. This paper is arranged in the following sections, the proposed method is given in Section 2, mathematical expressions of Process identification are derived in Section 2.1, the Simulation study is detailed in Section 3, and finally, conclusions are presented in Section 4.

2 Proposed Identification Approach

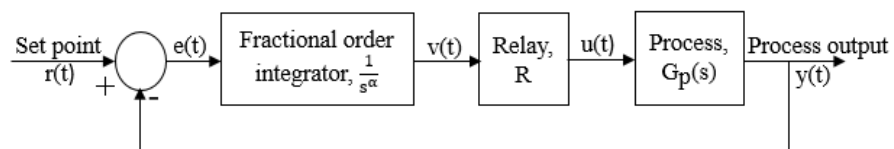


Figure 1

Block diagram of Relay feedback with FO integral

The scheme used for identification is shown in Figure 1, which consists of process $G_p(s)$, nonlinear element relay with hysteresis R and fractional order integrator. $u(t)$, $v(t)$ and $e(t)$ are the process input, relay input, and error respectively.

If $r(t)=0$ then the error signal $e(t)=y(t)$. Consider the error as a sinusoidal function as given in eq. (1)

$$e(t) = A \sin(\omega t) \quad (1)$$

Where A and ω are the Amplitude and fundamental frequency of the process output signal. The representation of eq. (1) which is used throughout the identification procedure is given by eq. (2)

$$e(t) = A \sin(L) \quad (2)$$

Where $L = \omega t$ Fractional order integral follows the Riemann-Lowville (R-L) definition [27-30]. The Describing Function of relay with hysteresis changes with the order of fractional integral, which is fixed in the case of conventional relay. The fractional order integrator is a linear element and is defined (eq. 3) as

$$\frac{1}{s^\alpha} = \frac{1}{(j\omega)^\alpha} = e^{-j\frac{\pi}{2}\alpha} \quad (3)$$

Where α is the order of fractional integrator. The signal after passing through the fractional integrator shifts their phase by $\frac{\pi}{2}\alpha$. The output of fractional integrator $v(t)$ is

$$v(t) = A \sin(L - \frac{\pi}{2}\alpha) \quad (4)$$

Then, the output of the relay is given by (5) [31].

$$u(L) = \begin{cases} -h & 0 < L < \theta_0 \\ +h & \theta_0 < L < \theta_0 + \pi \\ -h & \theta_0 + \pi < L < 2\pi \end{cases} \quad (5)$$

Where

$$\theta_0 = \sin^{-1} \left(\frac{\varepsilon + \frac{\pi}{2}}{A} \right) \quad (6)$$

Where $\pm h$ indicates relay height or amplitude and $\pm \varepsilon$ is the hysteresis width. As describing function analysis provides the tool for frequency domain analysis of nonlinear system, the DF is obtained by considering the principle harmonics of relay output signal. Therefore, relay with hysteresis is approximated with gain as given in (7)

$$N = \frac{1}{\pi\alpha} \int_0^{2\pi} u(L)(\sin L + j \cos L) dL \quad (7)$$

The Describing function of relay with hysteresis and fractional integrator is obtained by solving eq. (7) using eq. (5). The resulting describing function N is given by eq. (8)

$$N = \frac{4h(\sqrt{A^2 - \varepsilon^2} - j\varepsilon)}{\pi A^2} e^{-j\frac{\pi}{2}\alpha} \quad (8)$$

The condition to obtain sustained oscillations during identification is

$$NG_p(j\omega) = -1 \quad (9)$$

2.1 Identification Procedure for SOPTD Process

Consider the SOPTD model given in eq. (10)

$$G_p(s) = \frac{ke^{-\theta s}}{(\tau_1 s + 1)(\tau_2 s + 1)} \quad (10)$$

Equation (11) gives the frequency domain representation of above equation with $s=j\omega$ is

$$G_p(j\omega) = \frac{ke^{-\theta j\omega}}{(\tau_1 j\omega + 1)(\tau_2 j\omega + 1)} \quad (11)$$

The unknowns to be identified are: process gain (k), time constants (τ_1, τ_2) and time delay (θ). Substitute eq. (11) and eq. (8) in eq. (9) to obtain the condition for sustained oscillation

$$\frac{4hke^{-j\omega\theta}(\sqrt{A^2 - \varepsilon^2} - j\varepsilon)}{\pi A^2(\tau_1 j\omega + 1)(\tau_2 j\omega + 1)} e^{j\frac{\pi}{2}\alpha} = -1 \quad (12)$$

Equate the magnitude and phase angles on both sides of eq. (12) to get the unknown parameters. The equation obtained by equating the magnitude is given in (13)

$$\frac{4hk}{\pi A \sqrt{(\tau_1^2 \omega^2 + 1)} \sqrt{(\tau_2^2 \omega^2 + 1)}} = 1 \quad (13)$$

The resulting equation (14) after simplifying eq. (13) in terms of τ_1 and τ_2 is

$$\tau_1 + \tau_2 = \sqrt{\frac{1}{\omega^2} \left[\left(\frac{4hk}{\pi A} \right)^2 - 1 \right] + 2\tau_1\tau_2 - (\omega\tau_1\tau_2)^2} \quad (14)$$

Equation (15) is obtained by equating the phase angles in eq. (12)

$$-\theta\omega - \tan^{-1}(\tau_1\omega) - \tan^{-1}(\tau_2\omega) - \tan^{-1}\left(\frac{\varepsilon}{\sqrt{A^2 - \varepsilon^2}}\right) - \frac{\pi}{2}\alpha = -\pi \quad (15)$$

Rearranging the above equation for τ_1 and τ_2 results in eq. (16)

$$\tau_1\tau_2 = \frac{1}{\omega^2} \left[1 - \frac{\omega(\tau_1 + \tau_2)}{\tan\left[\phi - \omega\theta - \tan^{-1}\left(\frac{\varepsilon}{\sqrt{A^2 - \varepsilon^2}}\right)\right]} \right] \quad (16)$$

Where

$$\theta = t_2 - t_1 \quad (17)$$

$$\phi = \pi - \alpha \frac{\pi}{2} \quad (18)$$

2.2 The Systematic Approach for Identification of SOPTD Model Parameters is given as follows:

Step 1: Choose the parameters h and ε before performing the relay test.

Step 2: The parameter h is usually chosen as a symmetrical value. In the present work, it is fixed to ± 1 and ε is considered 2.5% of h .

Step 3: Perform relay test by choosing different values for α in the range 0.1-1.8.

Note: Any value of α beyond 2 results in an unstable response [32].

Step 4: Set $\alpha=0.1$, conduct relay experiment and note A , T and θ from the sustained oscillation along with IAE (eq. 19).

$$IAE = \int_0^{\omega_{cr}} \left| \frac{G_m(j\omega) - G(j\omega)}{G_m(j\omega)} \right| d\omega \quad (19)$$

Where ω_{cr} is the critical frequency of actual model, $G(j\omega)$ is the actual model and $G_m(j\omega)$ is the identified model.

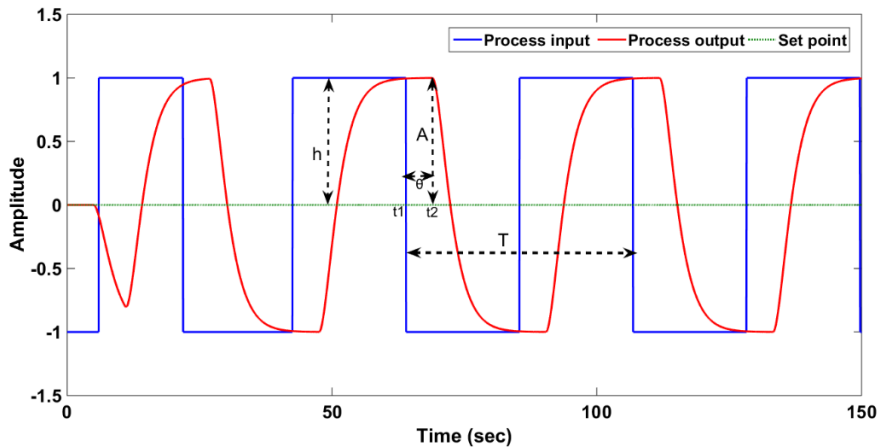


Figure 2
Process input-output signals

A sample input output signal diagram is shown in Figure 2, T is the critical time period of process output, t_1 is the time at which process input crosses set point $r(t)$ and t_2 is the time at peak amplitude of the process output.

Step 5: The process gain (k) is usually predefined and here it is chosen to be equal to the gain present in the actual process.

Step 6: Identify τ_1 and τ_2 using equations (14) and (16) after substituting h , ε , k , A , θ , and ω . Where $\omega = \frac{2\pi}{T}$ obtained in steps 4 and 5.

Step 7: Now, the second order model parameters are identified for $\alpha=0.1$

Step 8: Repeat steps 4 to 6 by varying α from 0.2 to 1.8 and identify the model parameters for each value of α .

Step 9: Finally, choose the optimum value for α and accurate second order model identified based on minimum IAE.

3 Simulation Results and Discussion

The simulations for model identification have been carried out on different second order systems Viz., underdamped, overdamped and critically damped systems. The identification of model parameters according to the novel systematic approach is delineated with the plots of model parameter variation with respect to fractional order (α) of the integrator. The value of α for which optimum model is identified is characterized through α versus IAE plots. Further, step response is observed to compare the exactness of identified model with the actual model.

3.1 Example 1

Consider the overdamped SOPTD model [4] given in eq. (20)

$$G_1(s) = \frac{e^{-2s}}{(10s+1)(s+1)} \quad (20)$$

The relay test is initiated by setting $h=\pm 1$ and $\varepsilon =\pm 0.025$. Now, the test is performed by considering $\alpha=0.1$ and then the model parameters are identified according to the systematic procedure. The relay test is repeated for different values of α (0.2-1.8) and the second order model parameters are identified for each value of α . The trends of variation of the identified parameters τ_1 , τ_2 and θ for each value of α is shown in Figure 3. It is observed that the variation of identified τ_1 and τ_2 are very close to actual values and are equal to the actual values at $\alpha=1.325$. The delay also becomes equal to the actual θ at $\alpha=1.325$. The best model is identified from the set of identified models based on minimum IAE for $\alpha=1.325$, which is evident from Figure 4. The critical period and amplitude are obtained as $T=67.80$ and $A=0.9275$.

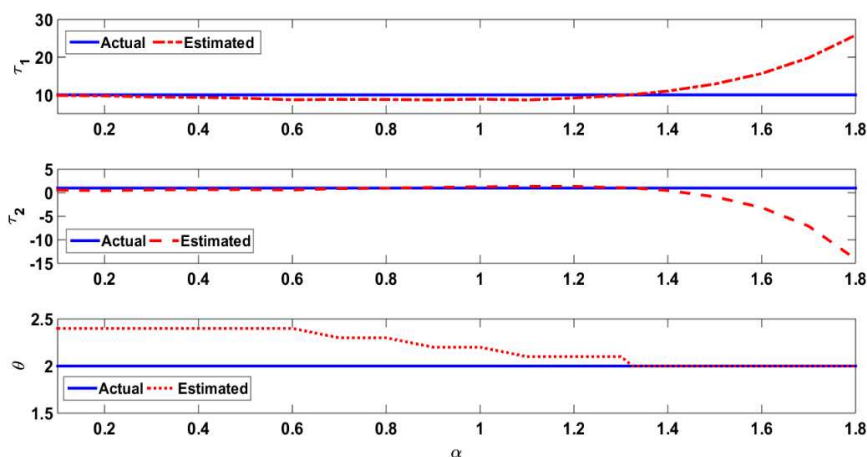


Figure 3
Trends of τ_1 , τ_2 and θ for variation in α

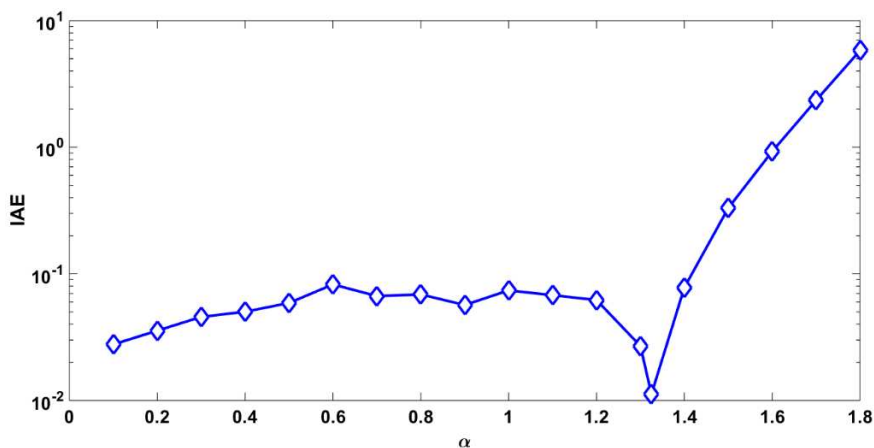


Figure 4
alpha vs IAE graph

The identified model, actual model and other models used for comparison are listed in Table 1 along with IAE. The model identified through proposed method gives low IAE compared to other models. The efficiency of the proposed identification method is proved under noisy environment in presence of measurement noise of 20 dB. The noise effect is achieved using a random additive noise with zero mean and 0.00013526 variance. The noisy process output and noise free limit cycle output obtained by curve fitting technique are as shown in Figure 5. The identified model with measurement noise is given in Table 1 and it proves that the proposed method is efficient with low IAE even under the influence of noise.

Table 1
Comparison of process models

Methods	Model	IAE
Actual Process	$\frac{e^{-2s}}{(10s+1)(s+1)}$	--
Proposed model	$\frac{e^{-2s}}{(10.061s+1)(1.054s+1)}$	0.0112
Proposed with measurement noise	$\frac{e^{-1.9s}}{(10.073s+1)(1.161s+1)}$	0.0161
Method (offline) in [7]	$\frac{e^{-3.12s}}{10.24s+1}$	0.0418
Method (online) in [7]	$\frac{e^{-3.15s}}{9.81s+1}$	0.0502
Method in [4]	$\frac{e^{-2.84s}}{11.98s+1}$	0.0601

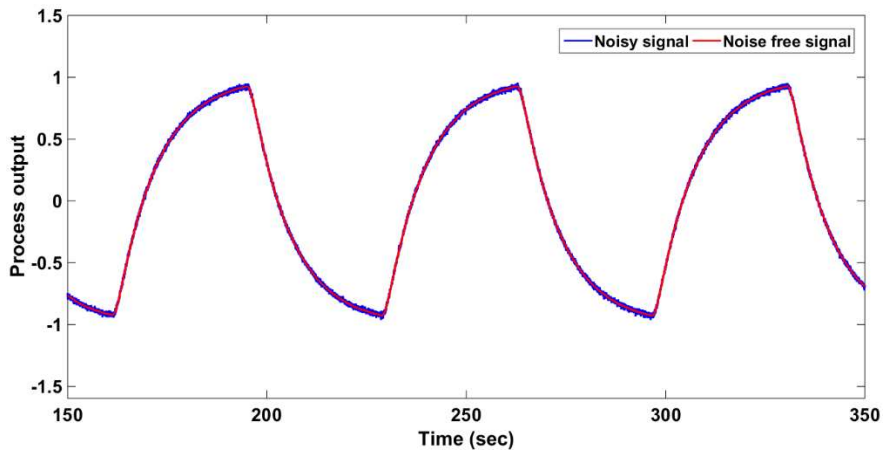


Figure 5

Noisy and noise free process output

The step response of the proposed model is shown in Figure 6. It is clear that the response with proposed model is close to the actual model compared to the other methods [7, 4]. It is observed from Figure 7 that the Nyquist plot of the proposed method is close to the actual model.

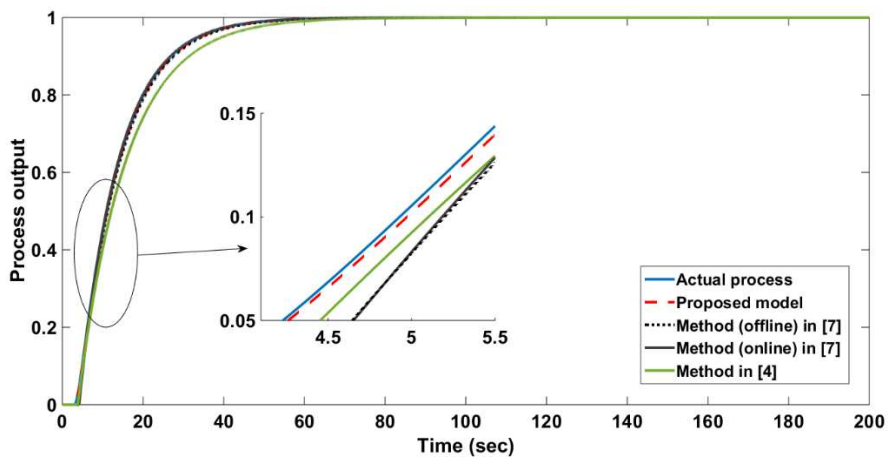


Figure 6

Step responses of the proposed model, actual process, and methods present in literature

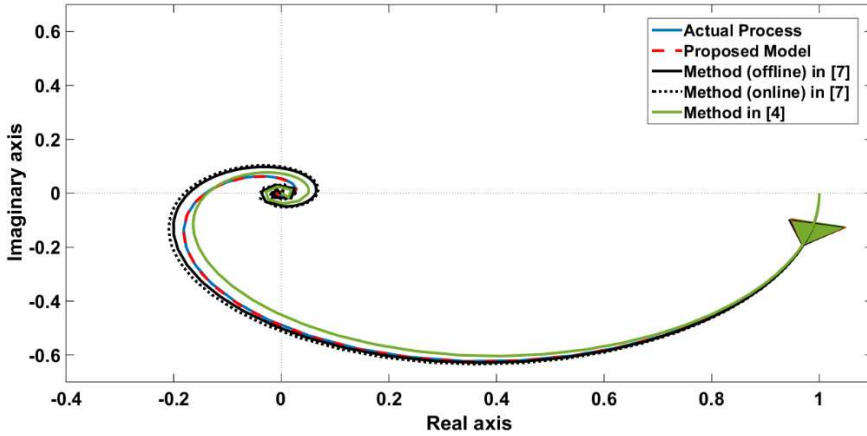


Figure 7
Nyquist plot

3.2 Example 2

The overdamped SOPTD process [9] used for identification is given in (21)

$$G_2(s) = \frac{e^{-4s}}{(10s+1)(2s+1)} \quad (21)$$

The values of $h=\pm 1$ and $\varepsilon=\pm 0.025$ are chosen to perform relay test. The model is identified according to the systematic approach given in section 2.1. The trends of variation of the identified parameters τ_1 , τ_2 and θ for each value of α is shown in Figure 8. The error between identified model and actual model (IAE) for each value of α is illustrated in Figure 9. The best model parameters are identified based on minimum IAE at $\alpha=1.15$, which is evident from Figures 8 & 9. The critical period and amplitude are obtained as $T=68.2$ and $A=0.9205$. The model identified according to the proposed method and existing model are listed in Table 2.

The model identified in presence of measurement noise (random noise with zero mean and 0.00013526 variance) is given in Table 2. The identified model under these circumstances is very close to actual model with minimum error (0.0063 lower than noise free model 0.0069) which is evident from Table 2 and Figure 10. It is also observed from the step response (Figure 11) and Nyquist plot (Figure 12) that the proposed model is in the close proximity of actual model compared to other models [9].

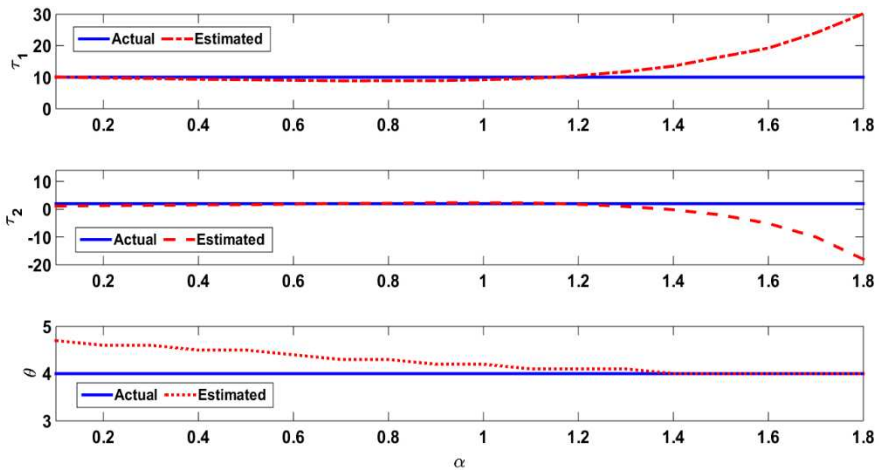


Figure 8
Trends of τ_1 , τ_2 and θ for variation in α

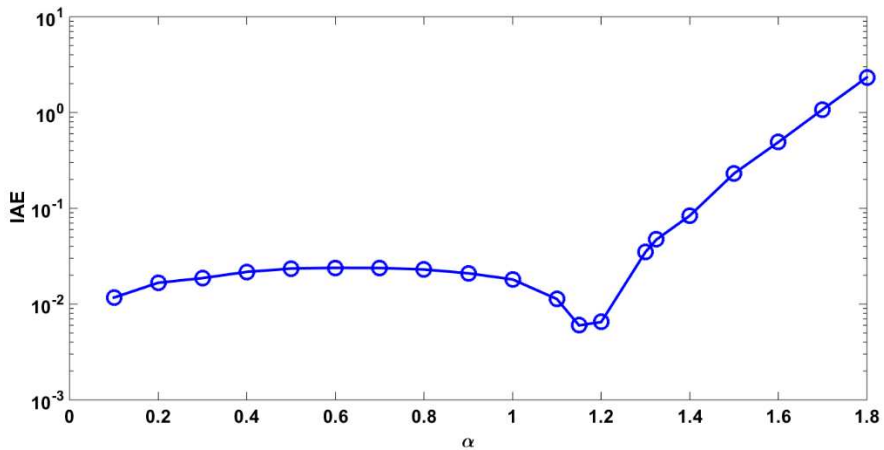


Figure 9
 α vs IAE graph

Table 2
Comparison of process models

Methods	Model	IAE
Actual Process	$\frac{e^{-4s}}{(10s+1)(2s+1)}$	--
Proposed model	$\frac{e^{-4.1s}}{(10.012s+1)(2.032s+1)}$	0.0069
Proposed with noise	$\frac{e^{-4.1s}}{(10.023s+1)(2.0156s+1)}$	0.0063

Method (offline) in [9]	$\frac{e^{-4s}}{(8.9312s+1)(2.1515s+1)}$	0.0206
Method (online) in [9]	$\frac{e^{-4.1s}}{(8.855s+1)(2.171s+1)}$	0.0218

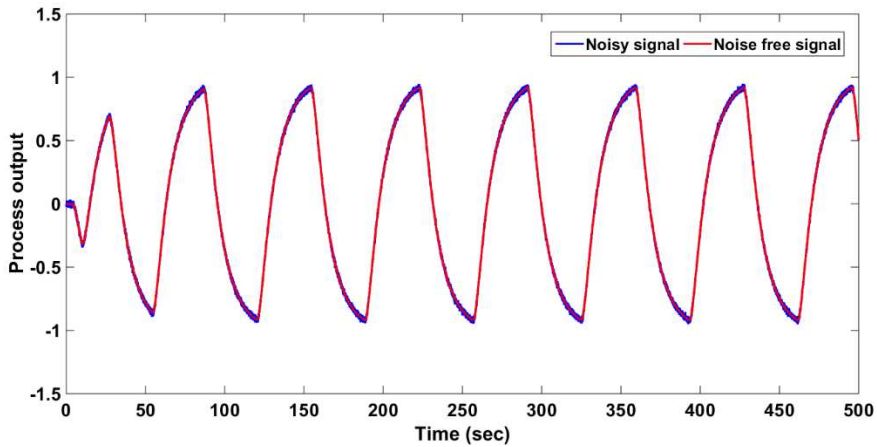


Figure 10
Noisy and noise free process output

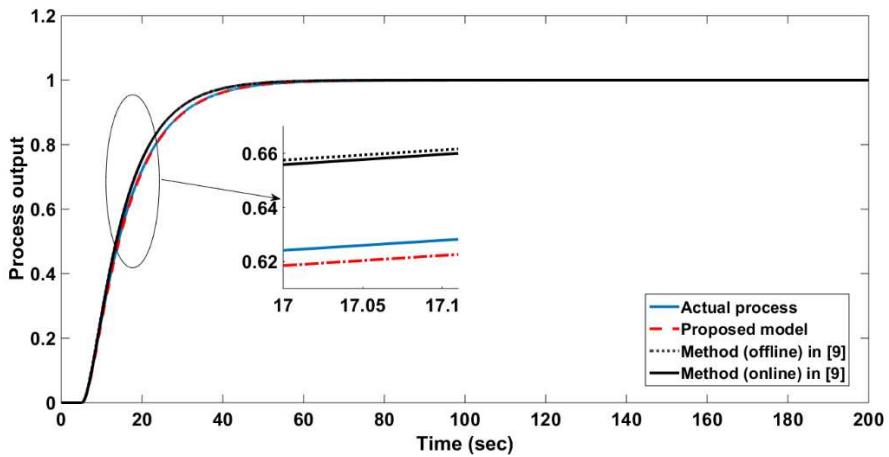


Figure 11
Step responses of the proposed model, actual process, and methods present in literature

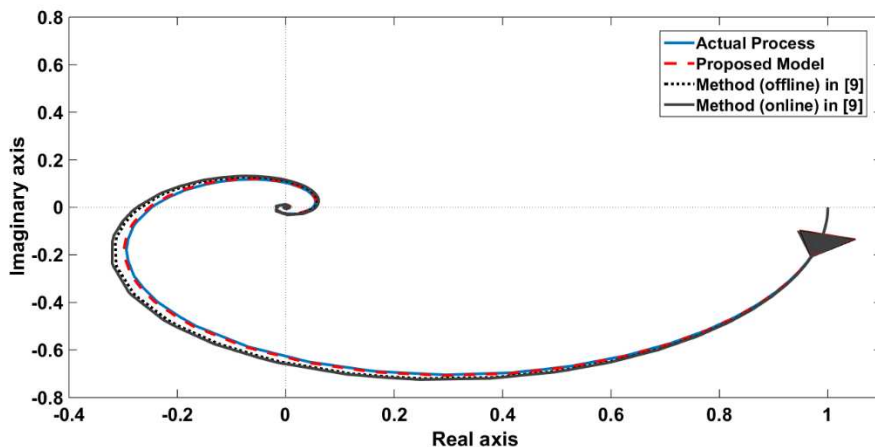


Figure 12
Nyquist plot

3.3 Example 3

The underdamped SOPTD process [2] used for simulation is in (22)

$$G_3(s) = \frac{e^{-s}}{9s^2+2.4s+1} \tag{22}$$

The model identification as per the proposed method is done with the settings: $h=\pm 1$ and $\varepsilon = \pm 0.025$. The models are identified by varying α from 0.1 to 1.8 following the systematic approach. The identifications results in complex values for the parameters τ_1 and τ_2 as the process is an underdamped system but a real value for θ . Hence, the trends (Figure 13) are plotted between $\tau_1\tau_2$ and $\tau_1 + \tau_2$ with respect to variation in α . The best model is identified at $\alpha=1.15$ (see α versus IAE in Fig. 14) and the corresponding model is given in Table 3 along with the error. The critical period and amplitude are identified as $T=23.5$ and $A=1.651$. It is observed that the proposed model is near the actual model with minimum error compared to method in [2]. The model in presence of random noise is also identified (Table 3) and it is a bit far from the actual model (illustrated in Figure 15) with a slightly high error. The exactness of the identified model to the actual model is also evident from step response and Nyquist plot shown in Figures 16 and 17.

Table 3
Comparison of process models

Methods	Model	IAE
Actual process	$\frac{e^{-s}}{9s^2+2.4s+1}$	--
Proposed model	$\frac{e^{-s}}{8.622s^2+2.487s+1}$	0.0042

Proposed model with noise	$\frac{e^{-0.9s}}{8.138s^2+2.532s+1}$	0.0330
Method in [2]	$\frac{1.0e^{-3.35s}}{3.96s+1}$	0.2948

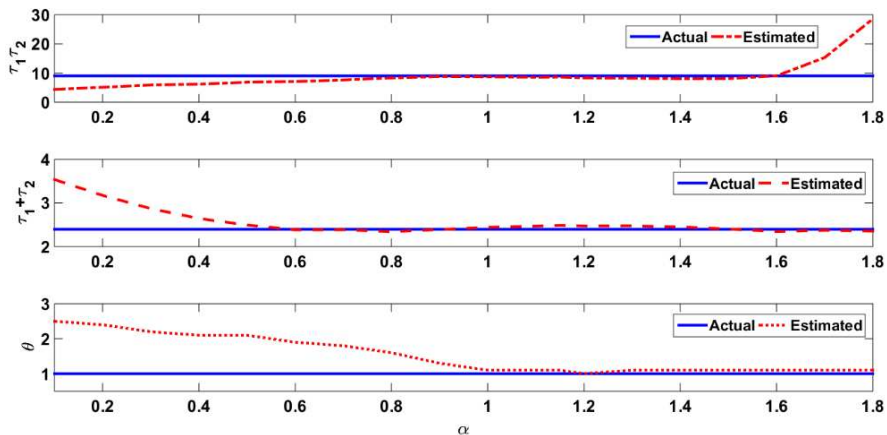


Figure 13
Trends of τ_1, τ_2 and θ for variation in α

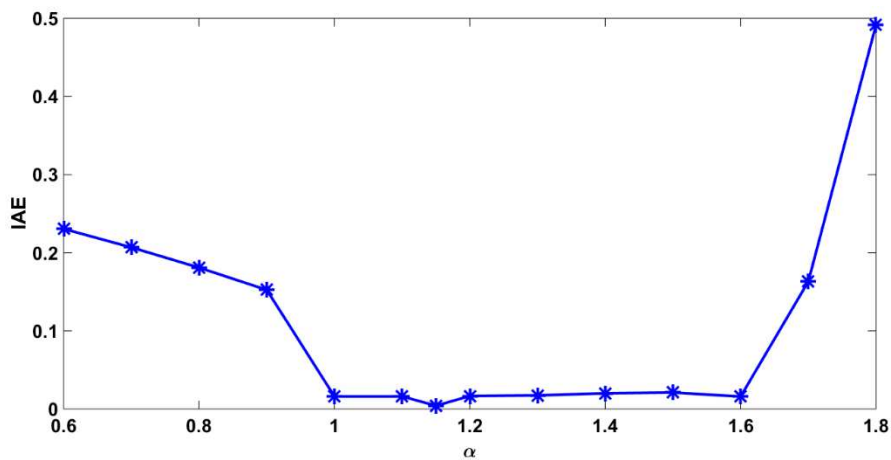


Figure 14
 α vs IAE graph

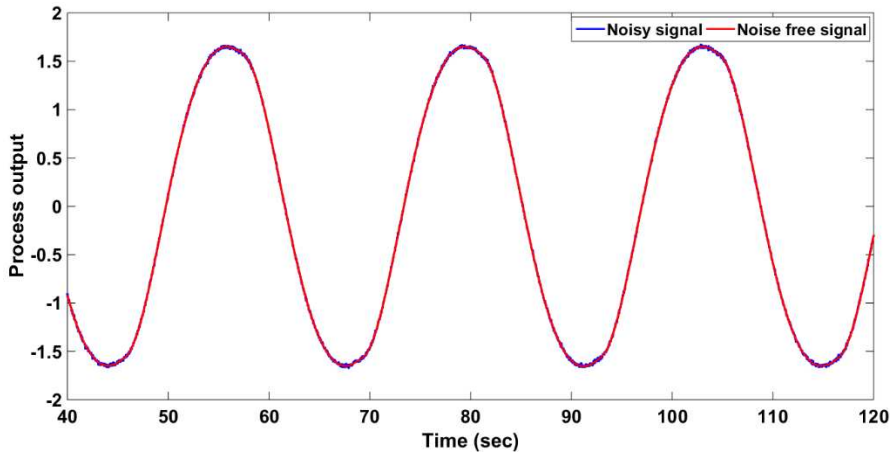


Figure 15
Noisy and noise free process output

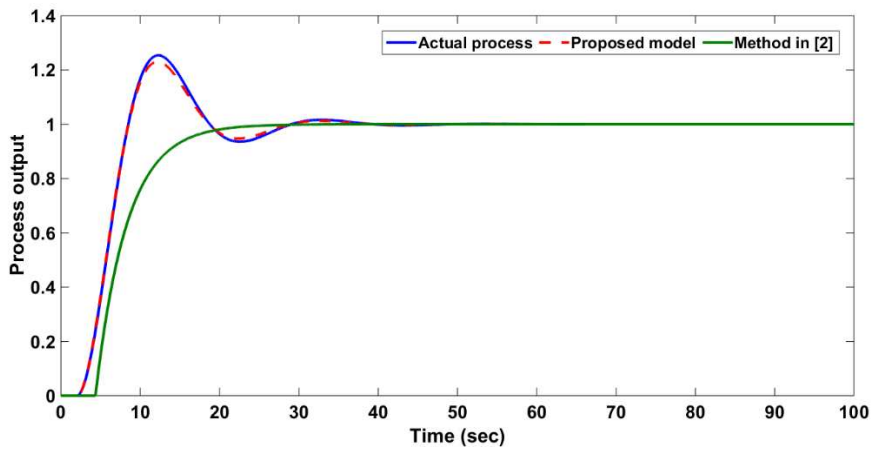


Figure 16
Step responses of the proposed model, actual process, and methods present in literature

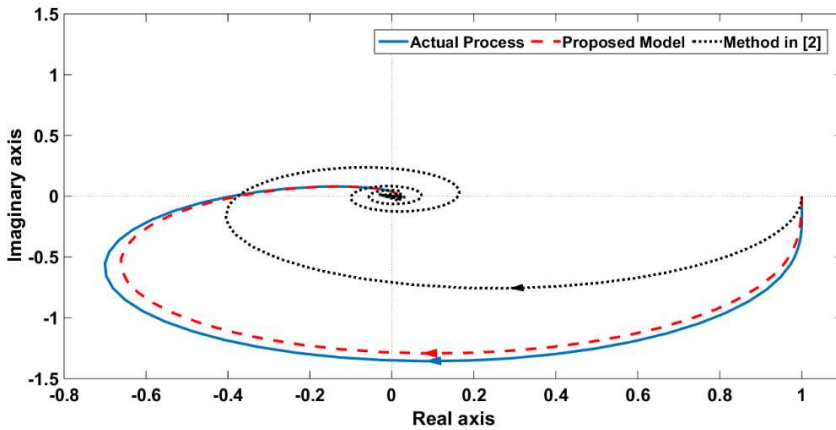


Figure 17
Nyquist plot

3.4 Example 4

Consider the following (eq. 23) critically damped process [9]

$$G_4(s) = \frac{e^{-0.01s}}{(2s+1)^2} \tag{23}$$

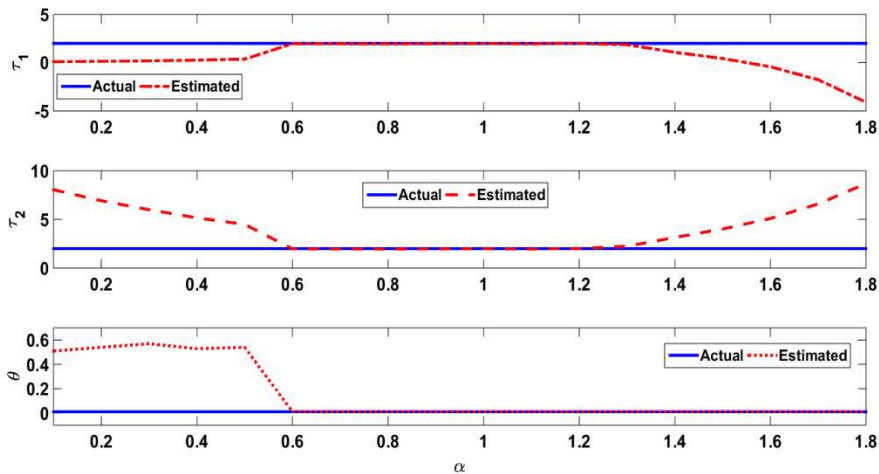


Figure 18
Trends of τ_1, τ_2 and θ for variation in α

The model identification is carried out according to the systematic approach with $h=\pm 1$ and $\varepsilon =\pm 0.025$. The variation of the identified parameters τ_1, τ_2 and θ for α is shown in Figure 18 and the IAE versus α plot is illustrated in Figure 19. It is observed that the model parameters are optimum at $\alpha =1.15$ with minimum IAE.

The corresponding critical period and amplitude are $T=16.6$ and $A=0.8459$. It is interesting to note that the model parameters are equal to actual values for a wide range of α which is evident from Figure 18. The proposed model identified according to the systematic approach along with other models is presented in Table 4. There is a slight rise in the error (Table 4) between the models identified in presence of noise compared to actual model (Figure 20). Figure 21 and Figure 22 illustrate that the proposed model is close to actual one for step change in the input and for variation in frequency.

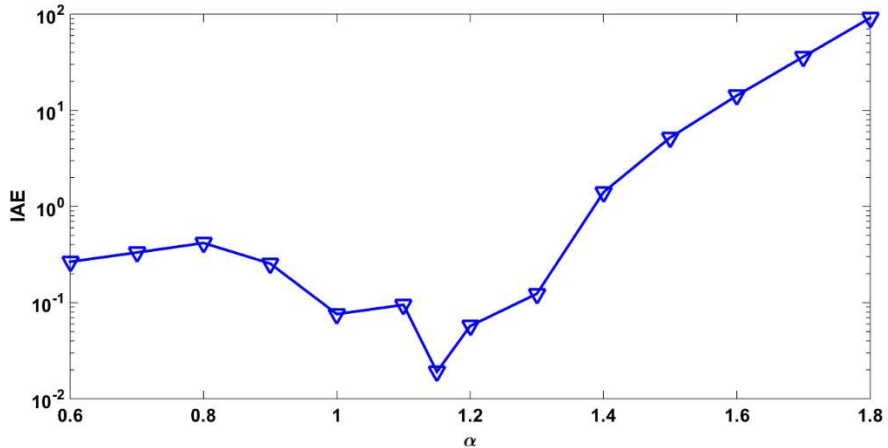


Figure 19
 α vs IAE graph

Table 4
Comparison of process models

Methods	Model	IAE
Actual Process	$\frac{e^{-0.01s}}{(2s+1)^2}$	--
Proposed model	$\frac{e^{-0.01s}}{(2.002s+1)^2}$	0.0190
Proposed with noise	$\frac{e^{-0.013s}}{(2.035s+1)^2}$	0.388
Method in [9]	$\frac{1.0084e^{-0.01s}}{(1.9962s+1)^2}$	0.118
Method in [3]	$\frac{0.8709e^{-0.013s}}{(1.897s+1)^2}$	0.465

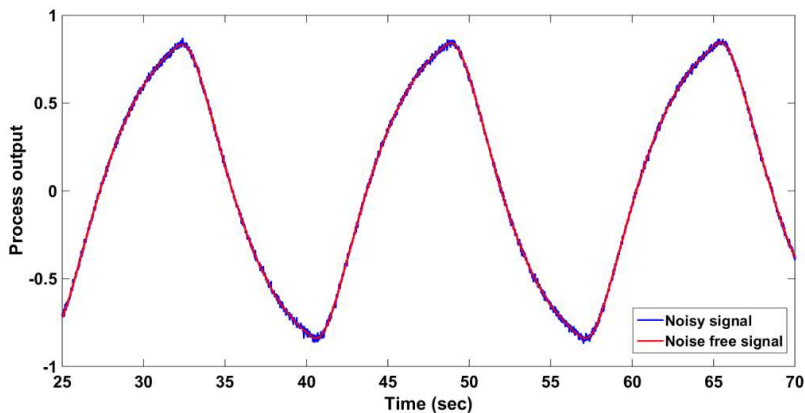


Figure 20
Noisy and noise free process output

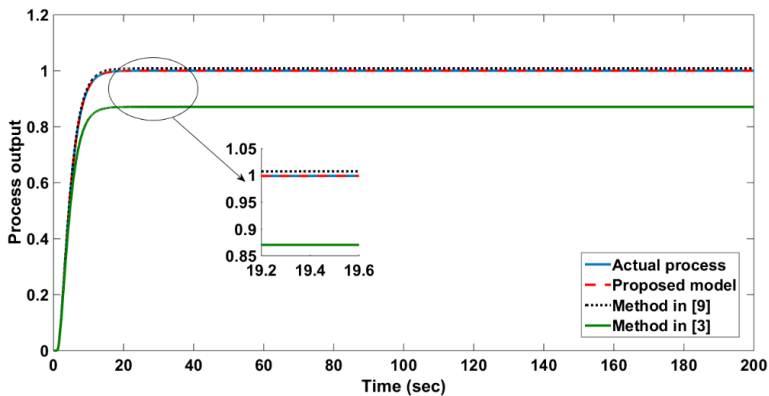


Figure 21
Step responses of the proposed model, actual process, and methods present in literature

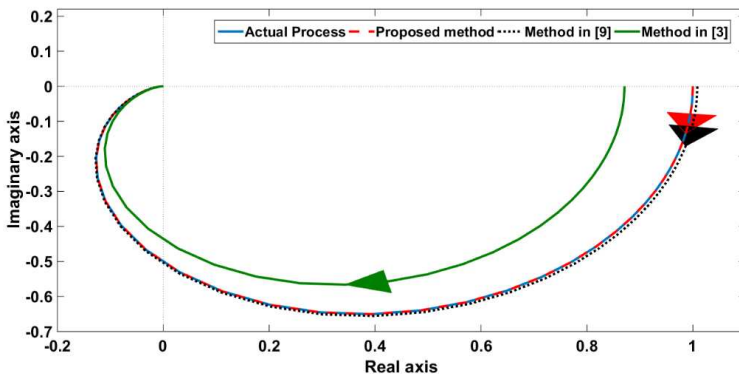


Figure 22
Nyquist plot

Conclusion

This paper proposes a new systematic identification approach for SOPTD processes using relay feedback with hysteresis and fractional order integrator. The proposed approach provides flexibility in the degree of freedom with the addition of fractional integrator and thus leads to more accurate SOPTD models. The proposed approach has an additional degree of freedom for estimating parameters, i.e., fractional order integrator. The describing function method developed the expressions to estimate the accurate model parameters. The addition of a fractional integrator helps improve the position frequency point obtained by the DF method. The proposed approach is found to be efficient under realistic conditions by estimating the model parameters in presence of measurement noise. To get the noise effect, white Gaussian noise is added to the process output, and the noisy limit cycle is processed through the curve fitting technique to obtain a clean signal. The proposed approach was applied to overdamped, underdamped, and critically damped SOPTD transfer function models and the performance is evaluated by comparing the absolute error criteria in the frequency domain, Nyquist plot, and step response. The proposed approach reduced IAE for Overdamped, Underdamped, and critically damped processes by 77.68%, 98.57%, and 95.78%, respectively compared to literature methods.

References

- [1] K. J. Astrom, T. Haggglund: Automatic Tuning of Simple Regulators with Specifications on Phase and Amplitude Margins, *Automatica*, Vol. 20, No. 5, 1984, pp. 645-651
- [2] C. L. Chen: A simple method for online identification and controller tuning, *AIChE Journal*, Vol. 35, No. 12, 1989, pp. 2037-2039
- [3] T. Thyagarajan, CC. Yu: Improved auto tuning using shape factor from relay feedback, *Industrial & Engineering Chemistry Research*, Vol. 42, No. 20, 2003, pp. 4425-4440
- [4] S. Vivek, M. Chidambaram: Identification using single symmetrical relay feedback test, *Computers and Chemical Engineering*, Vol. 29, No. 7, 2005, pp. 1625-1630
- [5] S. Majhi: Relay based identification of processes with time delay, *Journal of Process Control*, Vol. 17, No. 2, 2007, pp. 93-101
- [6] U. Mehata, S. Majhi. :Estimation of process model parameters based on half limit cycle data, *Journal system science & engineering*, Vol. 17, No. 2, 2008, pp. 13-21
- [7] R. Bajarangbali, S. Majhi: Relay Based Identification of Systems, *International Journal of Scientific & Engineering Research*, Vol. 3, No. 6, 2012, pp. 1-4
- [8] R. Bajarangbali, S. Majhi: Identification of underdamped process dynamics, *System Science & Control Engineering*, Vol. 2, No. 1, 2014, pp. 541-548

- [9] R. Bajarangbali, S. Majhi: Estimation of First and Second Order Process Model Parameters, *The National Academy of Sciences*, Vol. 88, No. 4, 2017, pp. 557-563
- [10] R. Bajarangbali, S. Majhi, S. Pandey: Identification of FOPDT and SOPDT process dynamics using closed loop test, *ISA Transactions*, Vol. 53, No. 4, 2014, pp. 1223-1231
- [11] R. Bajarangbali, S. Majhi: Identification of integrating and critically damped systems with time delay, *Control Theory & Technology*, Vol. 13, No. 1, 2015, pp. 29-36
- [12] R. Bajarangbali, S. Majhi: Identification of non-minimum phase processes with time delay in the presence of measurement noise, *ISA Transactions*, Vol. 57, 2015, pp. 245-253
- [13] Li. Zhuo, Chun Yin, Yang Quan, Chen, Jiaguo Liu: Process Identification Using Relay Feedback with a Fractional Order Integrator, *Proceedings of the 19th World Congress The International Federation of Automatic Control Cape Town, South Africa, 2014*, pp. 2010-2015
- [14] P. Ghorai, S. Majhi, S. Pandey: Modeling and Identification of Real-Time Processes Based on Nonzero Set point Auto tuning Test, *Journal of Dynamic Systems, Measurement, and Control*, Vol. 139, No. 2, 2017, pp. 1-8
- [15] P. Ghorai, S. Majhi, S. Pandey: A real-time approach for dead-time plant transfer function modeling based on relay auto tuning, *International Journal of Dynamics & Control*, Vol. 6, No. 3, 2018, pp. 950-960
- [16] S. Pandey, S. Majhi, P. Ghorai: A new modelling and identification scheme for time-delay systems with experimental investigation: a relay feedback approach, *international journal of systems science*, Vol. 48, No. 9, 2017, pp. 1932-1940
- [17] S. Pandey, S. Majhi: Limit cycle-based exact estimation of FOPDT process parameters under input/output disturbances: a state-space approach, *International Journal of Systems Science*, Vol. 48, No. 1, 2017, pp. 118-128
- [18] S. Pandey, S. Majhi: Relay-based identification scheme for processes with non-minimum phase and time delay, *IET Control Theory Appl.*, Vol. 13, No. 15, 2019, pp. 2507-2519
- [19] S. Pandey, S. Majhi: Limit cycle based identification of time delay SISO processes, *IFAC Journal of Systems and Control*, Vol. 48, No. 1, 2018, pp. 118-128
- [20] M. Hofreiter: Shifting method for relay feedback identification. *IFAC-Papers Online*, Vol. 49, No. 12, 2016, pp. 1933-1938
- [21] M. Hofreiter: Biased-relay feedback identification for time delay systems, *IFAC-Papers Online*, Vol. 50, No. 1, 2017, pp. 14620-14625

-
- [22] M. Hofreiter: Alternative identification method using biased relay feedback, IFAC-Papers Online, Vol. 51, No. 11, 2018, pp. 891-896
- [23] M. Hofreiter: Relay feedback identification with additional integral IFAC-Papers Online, Vol. 52, No.13, 2019, pp. 66-71
- [24] M. Hofreiter: Relay Feedback Identification with Shifting Filter for PID Control, IFAC Papers Online, Vol. 53, No. 2, 2020, pp. 10701-10706
- [25] M. Hofreiter: Generalized Relay Shifting Method for System Identification, IFAC Papers Online, Vol. 54, No. 1, 2021, pp. 498-503
- [26] R. Gerov, T. V. Jovanovic, Z. Jovanovic: Parameter Estimation Methods for the FOPDT Model, using the Lambert W Function, Acta Polytechnica Hungarica, Vol. 18, No. 9, 2021, pp. 141-159
- [27] I. Podlubny: Fractional differential equations, Mathematics in Science & Engineering, Academic Press, New York, 1999
- [28] E. C. De Oliveira, J. A. Tenreiro Machado: A review of definitions for fractional derivatives and integral, Mathematical Problems in Engineering, Vol. 2014, 2014, p. 6
- [29] J. Munkhammar: Riemann-Liouville fractional derivatives and the Taylor-Riemann series, Department of Mathematics, Uppsala University, Sweden, 2004
- [30] J. Liouville: Mémoire sur quelques Quéstions de Géometrie et de Mécanique, et sur un nouveau genre de Calcul pour résoudre ces Quéstions. Journal de l'école Polytechnique, Vol. 3, 1832, pp. 71-162
- [31] M. Chidambaram, V. Sathe: Relay Auto tuning for Identification and Control, Cambridge University Press, 2014
- [32] R. Caponetto, G. Maione, A. Pisano, M. R. Rapaić and E. Usai: Analysis and shaping of the self-sustained oscillations in relay controlled fractional order systems, Fractional Calculus and Applied Analysis, Vol. 16, No. 1, 2013, pp. 93-108