Study of Direct Bonding of Ceramic and Metallic Materials with Zn4Al Solder

Roman Koleňák, Igor Kostolný

Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava, Paulínska 16, 917 24 Trnava, Slovak Republic
roman.kolenak@stuba.sk, igor.kostolny@stuba.sk

Abstract: The aim of this work was to evaluate the direct bonding of Al2O3, SiC ceramics and Cu substrates. Joints were fabricated by using 40 kHz frequency ultrasound. The Zn4Al solder wetted all materials studied and joints of good quality were produced. The shear strength attained with Al2O3 ceramics was 81 MPa. The strength with SiC ceramics was slightly lower at 65 MPa. In a copper substrate, we observed shear strengths of 84 MPa.

Keywords: soldering; ceramic; metallic; microstructure; strength

1 Introduction

Zn-based solders belong to the group of solders applicable for higher application temperatures. Currently, soldering technology at these temperatures is widely used and imparts irreplaceable properties to the resulting product for high thermal conductivity and reliability. These solders are mainly used in electronics, as well as, in the automotive, space, aviation and power industries.

In the study [1], we investigated the direct bonding of SiC ceramics with ultrasound assistance. The ceramic SiC substrates were soldered in the air with Zn8.5Al1Mg solder at a temperature of 420°C. The shear strength of joints increased with longer periods of ultrasound exposure. The highest strength (148.1 MPa) was achieved at ultrasound periods lasting for 8s. A new amorphous layer 2 to 6 nm thick was formed on the boundary between the solder and substrate. The atoms from eroded SiO_2 layers from SiC substrates quickly diffused to the solder, owing to the jet effect caused by ultrasound. The strong bond between SiC substrate and Zn-Al-Mg solder is attributed to the transfer of SiO_2 mass to Zn-Al-Mg solder by induced cavitation erosion.

Direct bonding of sapphire (a crystalline form of Al_2O_3) by ultrasound, with the application of Sn10Zn2Al solder was the subject of a study [2]. It was found that ultrasound supported the oxidation reaction between Al from the solder and

sapphire substrate. A nano-crystalline α-Al_2O_3 layer (2 nm thick) was formed in the Sn-Zn-Al/sapphire boundary during soldering in the air at a temperature of 230°C. The shear strength of joints measured 43 to 48 MPa, which is a relatively high value when compared to other Al_2O_3 ceramic joints fabricated with active Sn solders and the addition of Ti and/or lanthanides [3, 4, 5].

The aim of our work was to study the direct bonding of Al_2O_3, SiC ceramics and copper substrate. Contrary to previous studies, the close-to-eutectic solder based on Zn-Al, (actually Zn4Al) was used. This solder is used for fluxless soldering of aluminum and its alloys. Ultrasonic soldering with direct ultrasound action was employed through the layer of molten solder.

2 Experimental

Zn solder with 4 wt% of Al was used in the experiments. The solder was manufactured in cast state in a high vacuum of 10^{-4} Pa. The procedure was as follows: the calculated charges of alloy components were inserted into a graphite boat. The boat with the charge was placed into a horizontal tube resistance vacuum furnace so that the boat was situated in the heating zone. The tube could be flushed with Ar, owing to a flange on its edge and an outlet on its end.

For Zn-based solders it is more suitable to prepare them in overpressure of Ar, due to evaporation. The charge was exposed to temperature above 450°C. Homogenization of individual components took place at this temperature.

Experiments used the substrates of the following materials:

- Metallic substrate of Cu with 4 N purity in the form of rings, in dimensions Ø 15 x 1.5 mm
- Ceramic Al_2O_3 substrate, with 2N5 purity in the form of Ø 15 x 2 mm rings (manufacturer Glynwed, GmbH, designation Degussit Al23),
- Ceramic SiC substrate in the form of Ø 15 x 3 mm rings (manufacturer CeramTec, GmbH, des. Rocar® SiC).

The combinations of materials shown in Fig. 1 were used for more detailed analysis.

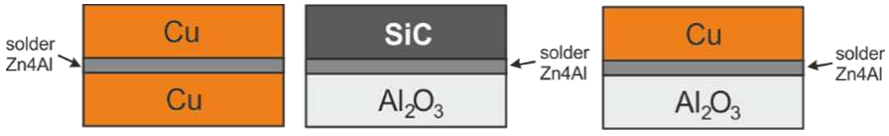


Figure 1
Analyzed combinations of Cu/Cu, SiC/Al_2O_3 and Cu/Al_2O_3 materials

Soldering was performed by Hanuz UT2 ultrasonic equipment with the parameters given in Table 1. The solder was activated by use of an encapsulated ultrasonic transducer consisting of a piezo-electric oscillating system and a titanium sonotrode with an Ø 3 mm end diameter. The scheme of ultrasonic soldering through the layer of molten solder is shown in Fig. 2. The soldering temperature was 20°C above the liquid temperature of the solder. Soldering temperature was checked by a continuous temperature measurement on the hot plate, using a NiCr/NiSi thermocouple.

Table 1
Soldering parameters

Ultrasound power	[W]	400
Working frequency	[kHz]	40
Amplitude	[μm]	2
Soldering temperature	[°C]	415 °C
Time of ultrasound activation	[s]	5

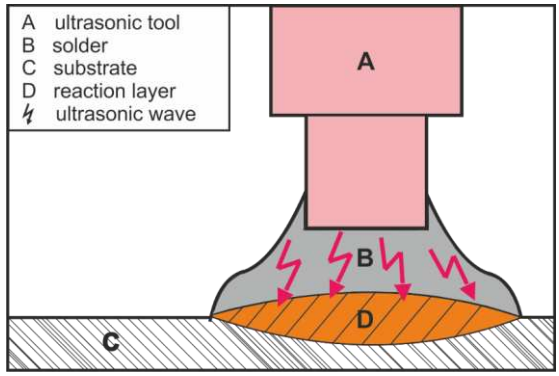


Figure 2
Ultrasonic soldering through the layer of molten solder

Soldering procedures took place in with the substrate heated at the soldering temperature deposited with a solder layer. Active ultrasound then acts upon the molten solder in the air without use of protective atmosphere for 5 s. After ultrasonic activation, the excessive layer of molten solder and the formed oxides are removed from the substrate surface. Both soldered substrates were prepared in the same way. The substrates with a deposited layer of molten solder were applied to each other so as to maintain contact during the molten phase. This assembly is then centered and the desired joint is achieved by slight compression. A graphic illustration of this procedure is shown in Fig. 3.

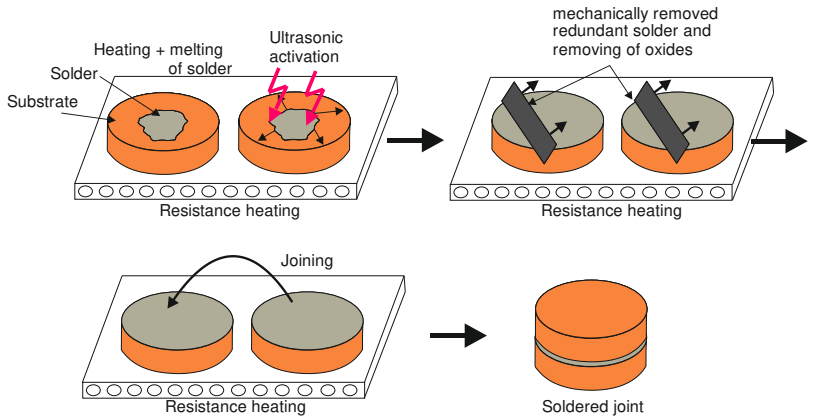


Figure 3
Procedure of joint fabrication by ultrasonic soldering

Metallographic preparation of specimens from the soldered joints was achieved by standard metallographic procedures used for preparation of specimens. The SiC emery papers with a granularity of 240, 320 and 1200 grains/cm^2 were used for grinding. The polishing was done by use of diamond suspensions with grain sizes: 9 μm, 6 μm and 3 μm. The final polishing was done using a type OP-S (Struers)polishing emulsion with a granularity of 0.2 μm.

Solder microstructure was observed by the aid of following:

- Neophot 32 light optical microscope, supplemented by a NIS-Elements, type E image analyzer
- Qualitative and semi-qualitative chemical analysis of the solder was performed by JEOL 7600 F equipment with a Microspec WDX-3PC X-ray micro-analyzer

X-ray diffraction analysis was used to identify the phase composition of the solder. It was applied on 10 x 10 mm solder specimens using a PANalyticalX´Pert PRO XRD diffractometer.

The DSC analysis of Zn4Al solder was performed on Netzsch STA 409 C/CD equipment in the Ar shielding gas with 6N purity.

A shear test was carried out to determine the shear strength of joints. Measurements were done on two ceramic (Al_2O_3 a SiC) and five metallic materials (Al, Ni, Ti, Cr-Ni steel, Cu) soldered using Zn4Al solder. The shear strength was determined on the versatile LabTest 5.250SP1-VM equipment. A shearing jig was used to change the direction of axial loading forces acting on the test specimen. This shearing jig ensured a uniform loading of specimens by shear in the plane of solder and substrate boundary (Fig. 4). The dwell time on soldering temperature during specimen fabrication was 30 s and the time of ultrasound acting was 5 s.

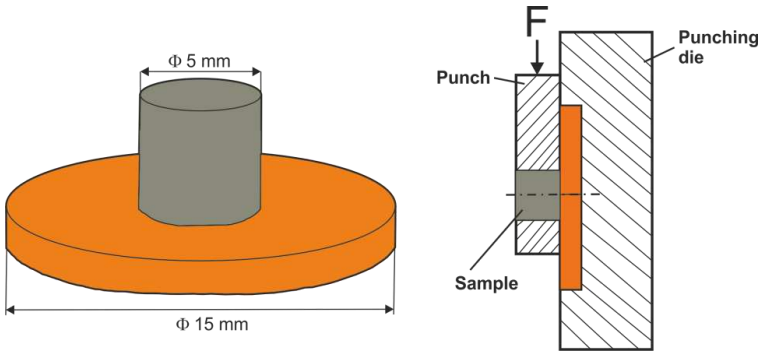


Figure 4
Test specimen for shear test and the scheme of specimen in the jig during the shear strength test [6]

3 Experimental Results

Analysis of ZnAl4 solder

Two solid solutions (Zn) and (Al) with a limited solubility occur in the binary Al-Zn system (Fig. 5). Owing to limited solubility, the eutecticum and eutectoid mixture of these solid solutions occurred in the system.

The matrix of Zn4Al solder (Fig. 6) was composed of great grains of the solid solution (Zn) with concentration of 98.68wt% Zn. A fine eutecticum, formed of solid solutions (Zn) + (Al) was segregated along the grain boundaries. The quantitative analysis of solder is given below in Fig. 6.

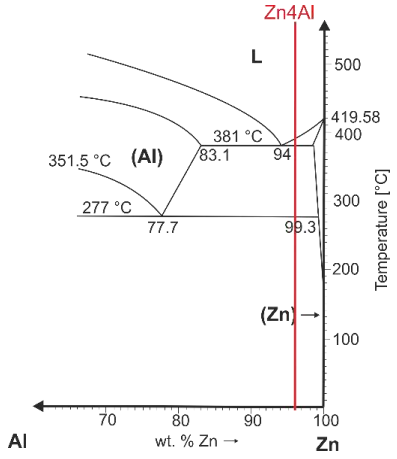
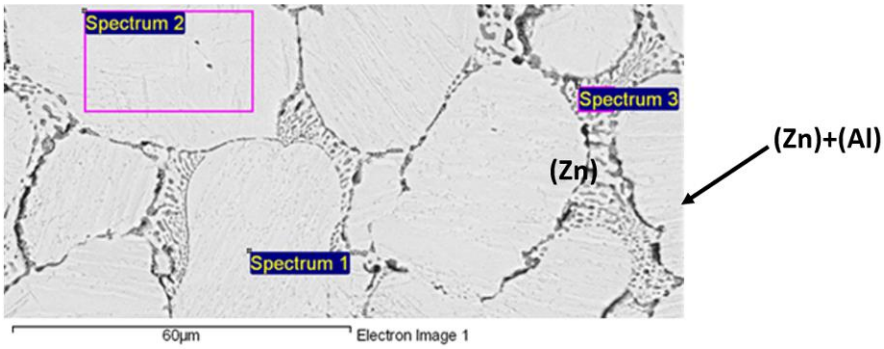


Figure 5
Binary Al-Zn diagram [7]



Spectrum	Al wt. %	Zn wt. %
Spectrum 1	1.10	98.90
Spectrum 2	1.32	98.68
Spectrum 3	16.53	83.47

Figure 6
Microstructure of Zn4Al solder

The ZnAl4 solder shows a narrower fusion interval, while it is of close-to eutectic composition. By the DSC analysis (Fig. 7a), a temperature of 277.6°C starts the onset of eutectoid transformation. The following reaction takes place at this temperature: Alrichfcc + hcp (Zn) / Zn-richfcc. The eutectic (Zn + 6 wt.% Al), segregated along the grain boundaries of Zn matrix of the solder starts to melt at 380.7°C. The solid solution (Zn) attains its fully liquid state at 385.9°C – Fig. 7b.

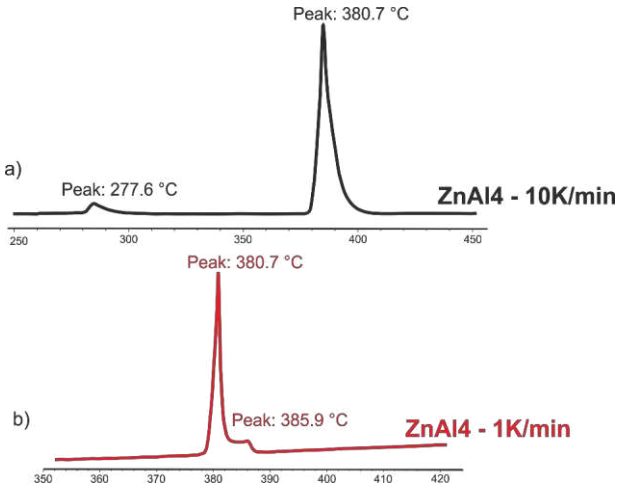


Figure 7
DSC analysis of Zn-4Al solder at heating rate: a) 10K/min b)1K/min

Mechanical tests of type Zn-Al solders were performed. The dimensions of test pieces were designed and calculated. Fig. 8 shows the dimensions of test specimens and the actual test piece. Three pieces for each type of alloy were used for experimental assessment. The results of tensile strength tests of soldered type Zn-Al alloys are documented in Fig. 9.

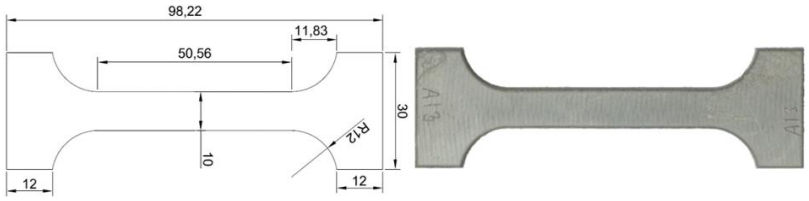


Figure 8
Dimensions of test specimen and a real view on a specimen of Zn4Al solder

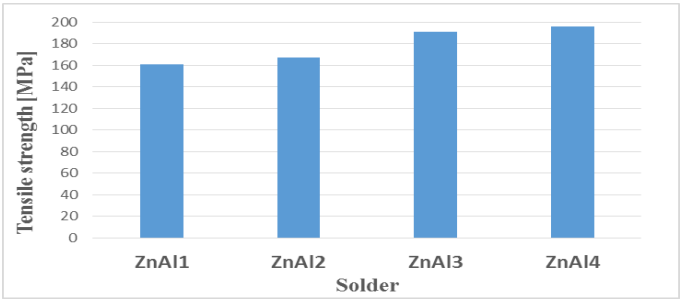


Figure 9
Tensile strength of soldered alloys type Zn-Al in dependence on Al content

Generally, Al content slightly increases the strength of Zn-Al solder. By increasing Al content, the strength of Zn-Al solder increases – Fig. 15. The variance in tensile strength between ZnAl1 and ZnAl4 solder is around 35 MPa.

Analysis of Joints in Al_2O_3 and SiC Ceramic Materials Soldered with Zn4Al

Fig. 10 shows the difference in the transition zone of Al_2O_3/Zn4Al and SiC/Zn4Al joints. A wide transition zone up to 70 μm in width was formed on the boundary with SiC, where an increased amount, mainly of carbidic and silicon particles, was identified. Due to ultrasound erosion, the solder penetrated the grains of the ceramic materials and carbidic particles from ceramics and silicon particles, infiltrated in the grain boundaries of SiC ceramics and displaced into the solder. The concentration profiles in the SiC/Zn4Al boundary zone are shown in Fig. 11.

No distinct transition layer is observable in the boundary with Al_2O_3 ceramics – Fig. 10b. The character of solder matrix in this boundary remains unchanged. No new transition phases were identified. The bond with Al_2O_3 ceramics is formed by solder adhesion with Al_2O_3 ceramics.

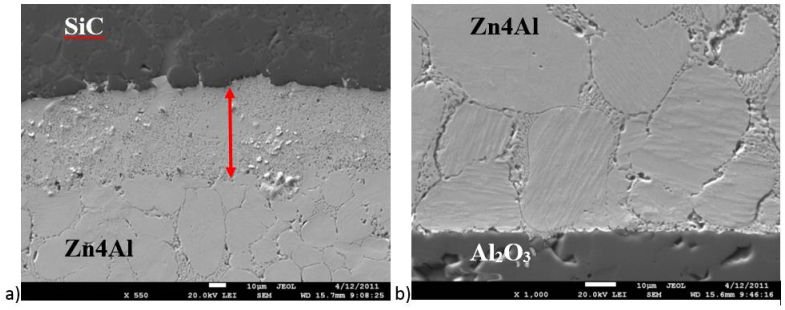


Figure 10
Microstructure in the boundary of a) SiC/Zn4Al, b)Al_2O_3/Zn4Al bonds

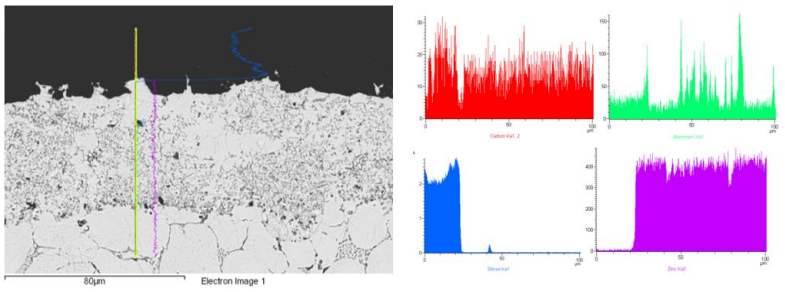


Figure 11
Microstructure in the boundary of SiC/Zn4Al bond and the concentration profiles of C, Al, Si and Zn elements

Analysis of Bond between Cu and Zn4Al Solder

A wide zone of two new intermetallic phases was formed in the boundary of the Cu/Zn4Al/Cu bond (Fig. 12). The $CuZn_4$ and Cu_5Zn_8 phases were identified. The concentration profiles of Cu, Zn and Al elements are shown in Fig. 13.

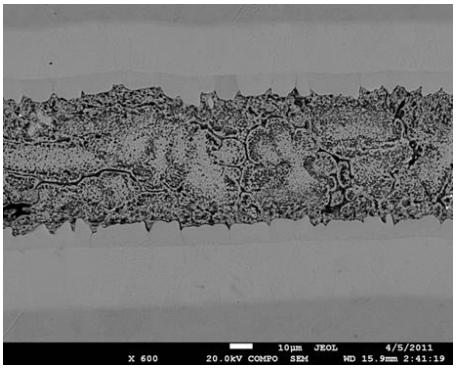


Figure 12
Microstructure of Cu/Zn4Al/Cu bond boundary

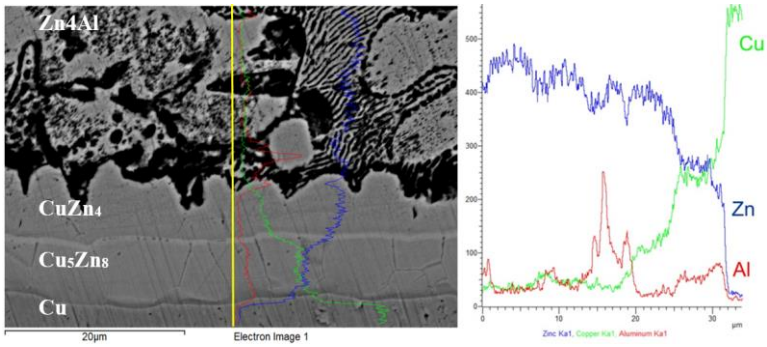


Figure 13
Line EDX analysis of Cu/Zn4Al soldered joint and concentration profiles of Cu, Zn, Al elements in Cu/Zn4Al boundary

The Results of Shear Strength of Soldered Joints

Research within this study was primary oriented toward soldering Al_2O_3 and SiC ceramic substrates and Cu substrates. The experiments to determine the shear strength of soldered joints were also extended to other metallic materials such as Al, Ni, Ti and CrNi steel, in order to prove the wider applicability of Zn4Al solder.

Measurement was performed on 4 specimens of each material. The results of average shear strength of joints are documented in Fig. 14. The shear strength of Al_2O_3 ceramics attained 81.0 MPa. With SiC ceramics, a slightly lower strength of 65.0 MPa was observed. With the copper substrate, a shear strength of 84.0 MPa was measured. The highest shear strength was achieved with aluminum -- 174.5 MPa.

For a more exact identification fractured surfaces in the boundary of Cu/Zn4Al and eventually Al_2O_3/Zn4Al bonds were also identified (Figs. 15, 16).

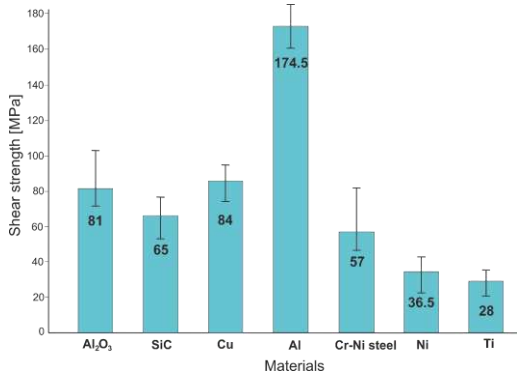


Figure 14
The results of shear strength measurements in joints fabricated by use of Zn4Al solder

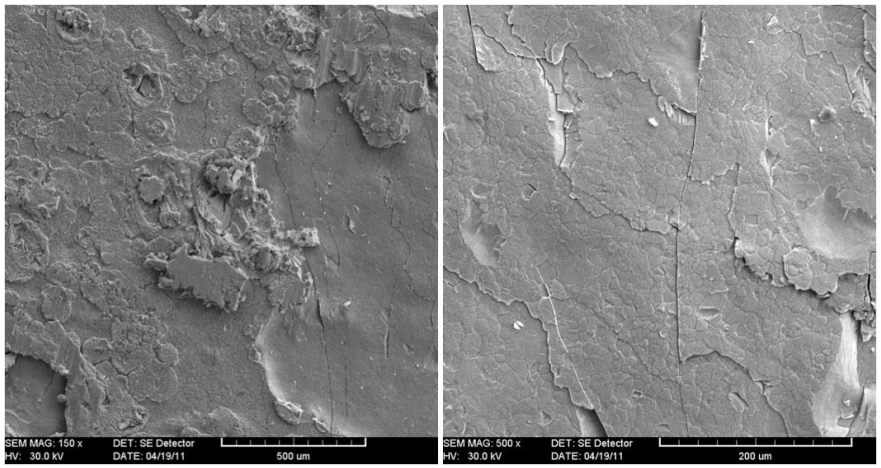


Figure 15
Fractured surface of Zn4Al/Cu joint

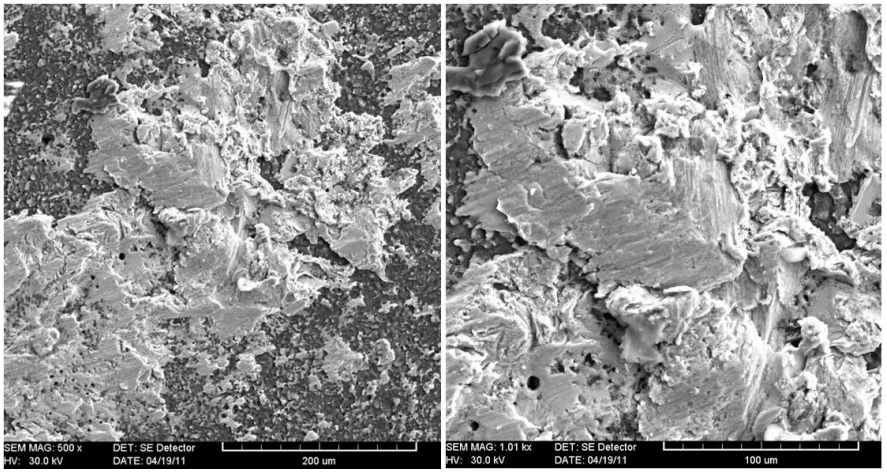


Figure 16
Fractured surface of Zn4Al/Al_2O_3 joint

Formation of a typical ductile failure by shear mechanism was documented in Cu/Zn4Al bond (Fig. 15). Fig. 16 shows the fractured surfaces of Al_2O_3/Zn4Al bonds. Fracture morphology evidently shows a visible motion of the shearing tool with a ductile fracture.

After the shear test, 100% coverage of Cu substrate with Zn4Al solder remained. To the contrary, in the case of Cu substrate the solder was detached from the ceramic substrate. Partial coverage of Al_2O_3 substrate was observed, as documented in Fig. 16.

4 Discussion

The results achieved by direct bonding of ceramic and metallic materials proved that Zn-based solder ensures the wettability at application of ultrasound activation, by which the Zn4Al solder becomes suitable for practical soldering applications.

For a comparison of results from the shear strength measurement we also present the results from similar studies, while it should be taken into account that different works make use of different test methods, the shape of test pieces and loading rates during testing. They also use different compositions of soldering alloys and soldering parameters.

For example in the case of the application of Zn-Al (Zn14Al) solders in work [8], with Al_2O_3/Zn14Al/Cu bonds, a shear strength of 80 MPa was observed at ultrasound power of 200 W. In study [1], the joint of SiC ceramic substrate, soldered with Zn8.5Al1Mg solder, attained a shear strength of 148.1 MPa at 8s. of ultrasound action.

In study [2], sapphire was soldered with ultrasound activation by use of Sn10Zn2Al solder. The shear strength of joints attained 43 to 48 MPa. More studies dealt with Sn-Ag-Ti solders. Also, new metallic, ceramic and non-metallic materials were tested.

For example in study [3], the Al_2O_3/Sn-Ag-Ti/Al_2O_3 bond showed a shear strength of 24 MPa. In work [4], the following strengths were achieved by soldering: Cu/Cu (14.3MPa), ITO/ITO (6.8 MPa) and ITO/Cu (3.4 MPa). Similarly, in study [8], the attained shear strengths of joints were as follows: alumina/alumina (13.5 MPa), copper/copper (14.3 MPa) and alumina/copper (10.2 MPa).

Conclusions

The aim of work was oriented toward the direct bonding of Al_2O_3, SiC ceramic substrates and a copper substrate. We examined the behavior of Zn-based solder, alloyed with Al, for the wetting of Al_2O_3 ceramics and other ceramic materials and the formation of a strong bond with them. Due to this, several analyses of the transition zones for the bonds and the measurements of shear strengths were performed. The following results were achieved:

- DSC analysis revealed that the solder has a smaller melting interval. At a temperature of 380.7°C, the eutecticum (Zn + 6wt% Al) segregated along the zinc matrix of the solder starts to melt. The solid solution (Zn) attains a fully liquid state at 385.9°C
- The matrix of Zn4Al solder is formed of great grains of solid solution (Zn) with a Zn concentration of 98.68 wt%. A fine eutecticum, formed of solid solutions (Zn) + (Al), is segregated along the grain boundaries.

- In SiC/Zn4Al bond boundary a transition zone was formed up to 70 μm in width, where increased amounts, mainly of carbidic and silicon particles, were identified.
- No distinct transition layer was observed in the Al_2O_3/Zn4Al bond boundary. The character of solder matrix in this boundary is unchanged. The Al_2O_3/Zn4Al bond is formed due to the adhesion of solder with Al_2O_3 ceramics.
- A wide zone of two new intermetallic phases was formed on the Cu/Zn4Al boundary (Fig. 14), where $CuZn_4$ and Cu_5Zn_8 phases were identified.
- The shear strength of 81.0 MPa was obtained with Al_2O_3 ceramics. With SiC ceramics, a slightly lower strength of 65.0 MPa was observed. With copper substrate, a shear strength of 84.0 MPa was measured. The highest shear strength was achieved with aluminum at 174.5 MPa.

After the shear test, 100% coverage of Cu substrate with Zn4Al solder remained. In contrast, for the case of the Cu substrate, the solder was detached from the ceramic substrate, whereas, partial coverage of Al_2O_3 substrate was observed.

Acknowledgement

The contribution was prepared with the support of APVV–0023–12: Research of new soldering alloys for fluxless soldering with application of beam technologies and ultrasound and VEGA 1/0455/14: Research of modified solders for fluxless soldering of metallic and ceramic materials. The authors thank Ing. Marián Drienovský, PhD. for DSC analysis; doc. Ing. Maroš Martinkovič, PhD. for shear strength measurements and Ing. Ivona Černičková, PhD. for EDX analysis.

References

[1] Chen, X., Yan, J., Ren, S., et al., Microstructure, Mechanical Properties, and Bonding Mechanism of Ultrasonic-assisted Brazed Joints of SiC Ceramics with ZnAlMg Filler Metals in Air. In *Ceramics International*, Vol. 40, 2014, pp. 683-689

[2] Cui, W., Yan, J., Dai, Y., Li, D., Building a Nano-Crystalline α-alumina Layer at a Liquid Metal/Sapphire Interface by Ultrasound. In *Ultrasonics Sonochemistry*, Vol. 22, 2015, pp. 108-112

[3] Koleňák, R., Šebo, P., Provazník, M., Koleňáková, M., Ulrich, K. Shear Strength and Wettability of Active Sn3.5Ag4Ti(Ce,Ga) Solder on Al_2O_3 Ceramics. In *Materials and Design*, Vol. 32, 2011, pp. 3997-4003

[4] Chang, S. Y., Tsao, L. C., Chiang, M. J., et al., Active Soldering of Indium Tin Oxide (ITO) With Cu in Air Using an Sn3.5Ag4Ti(Ce,Ga) Filler. In *Journal of Materials Engineering and Performance*, Vol. 12, No. 4, 2003, pp. 383-390

[5] Chang, S. Y., Chuang, T. H., Yang, C. L., Low Temperature Bonding of Alumina/Alumina and Alumina/Copper in Air Using Sn3.5Ag4Ti(Ce,Ga) Filler. In *Journal of Electronic Materials*, Vol. 36, No. 9, 2007, pp. 1193-1199

[6] M. Martinkovič [patent inventor], R. Koleňák [patent inventor]: The Form on the Production of Experimental Soldered Joint – Patent 288180. Industrial Property Office of the Slovak Republic [patentee]. Date of patent accordance: 12.02.2014

[7] Murray, J. L. The Al-Zn (Aluminium-Zinc) System. In Bulletin of Alloy Phase Diagrams, Vol. 4, Is. 1, 1983, pp. 55-73

[8] Ji, H., Chen, H., Li, M., Microstructures and Properties of Alumina/Copper Joints Fabricated by Ultrasonic-assisted Brazing for Replacing DBC in Power Electronics Packaging. In 15th International Conference on Electronic Packaging Technology, IEEE, 2014, pp. 1291-1295

Study of the Stabilization of Uncertain Nonlinear Systems Controlled by State Feedback

Amira Gharbi[1,2], Mohamed Benrejeb[1] and Pierre Borne[2]

[1]Laboratoire de Recherche en Automatique, LARA. Ecole Nationale d'Ingénieurs de Tunis, BP 37 Le Belvédère 1002 Tunis, Tunisia

[2]Centre de Recherche en Informatique, Signal et Automatique de Lille, CRIStAL. Ecole Centrale de Lille. Cité scientifique, BP 48-59651 Villeneuve d'Ascq Cédex, France

E-mail: amira.gharbi@enit.rnu.tn, mohamed.benrejeb@ec-lille.fr, pierre.borne@ec-lille.fr

Abstract: The control of a process by poles placement is one of the most used forms of feedback control. It allows not only to stabilize a process, but also to control its dynamic. Furthermore, the optimal controls with quadratic criteria of linear systems in fact lead to the pole placement. In this work, we present an approach to the stabilization of nonlinear systems in presence of uncertainties using poles placement by state feedback and the determination of attractors by diagonalization of the characteristic matrices linearized around operating points and using aggregation techniques.

Keywords: aggregation techniques; attractors; comparison systems; state feedback control; uncertain nonlinear

1 Introduction

The control of complex nonlinear process appears generally difficult, particularly in the case of ill-defined or imprecise models and when these processes are subject to unidentified noises or disturbances for which the only available information is the amplitudes of the uncertainties resulting in the definition of the model. A great number of works have been presented related to this problem [1-6]. For a nonlinear process in continuous time, whose evolution is described by a set of differential equations, the most commonly used model is represented in the state space.

However, starting from a set of given differential equations, several representations can be used and the choice of the model can affect the accuracy of the expected results.

In the presence of uncertainties in modeling, that increase the complexity of the stability study, it is not always possible to obtain a control law ensuring the stability of the process with respect to a chosen objective. It is then necessary to estimate the maximum deviation from this target, an operation which can be performed by determining an attractor corresponding to the vicinity of the purpose for which the local stability cannot be guaranteed [7-15].

Linear system stability study generally leads to necessary and sufficient conditions and doesn't depend, generally, on the system representation. The task is different for nonlinear systems with or without uncertainties, for which only sufficient conditions can be proposed; then the determination of their stability domains and attractors depends on the choice of both the description of the studied system and the used stability method [16-18].

Process control through poles placement is an usual feedback control used for linear systems [19]. It doesn't allow only to stabilize the studied process, but also imposes its dynamics. For nonlinear systems with uncertainties, the approach is more complex.

In the case of large scale systems, generally described in the state space, stability conditions are obtained, either directly for the whole system or separately for the various subsystems.

In this paper, the determination of the state feedback is based on a specific state space description of the linearized process and the determination of the attractor, when the process is submitted to uncertainties, is achieved by using aggregation techniques and the Borne-Gentina stability criteria, with the use of vector norms and of comparison systems [20-27].

The aim of this work is to present an approach to the study of stability of nonlinear systems and the estimation, by overvaluation, of the attractor. In Section 2, we propose an attractor determination method by diagonalization of the linearized characteristic matrix around an operating point when the control law is achieved by poles placement and by the use of the aggregation technique for stability study. The determination of attractor for a third order nonlinear complex system is presented, in Section 3, to illustrate the efficiency of the proposed approach.

2 Proposed Attractor Determination Method

In this section the poles placement is determined on a linearized model of the initial system without uncertainties.

2.1 Determination of State Feedback Gain *L*

Let us consider the system (S) described by

$$\dot{x}(t) = A(.)x(t) + B(.)u(t) + B'(.) \tag{1}$$

with $A \in R^{n\times n}$, $B \in R^n$, $x \in R^n$, $u \in R$ and $B' \in R^n$ characterizing the influence of uncertainties.

By linearization of the system (1) without uncertainties, around the operating point x_0, it comes the correspondent linearized model (2)

$$\dot{x}(t) = A(0)x(t) + B(0)u(t) \tag{2}$$

assumed to be controllable.

The state feedback control law of (2) is defined in the form

$$u(t) = -Lx(t) \tag{3}$$

such that

$$L = \begin{bmatrix} l_0 & l_1 & \dots & l_{n-1} \end{bmatrix}, \quad L \in R^n \tag{4}$$

Note P_c the matrix of change of base such that

$$x = P_c x_c \tag{5}$$

which enables to describe the linearized system (1) without uncertainties in the controllable canonical form

$$\dot{x}_c(t) = A_c x_c(t) + B_c u(t) \tag{6}$$

with x_c the new state vector of the process, $A_c = P_c^{-1} A P_c$ and $B_c = P_c^{-1} B$.

After substituing (3) in (1), it comes for the process without uncertainties

$$\dot{x}_c(t) = A_c x_c(t) - B_c L x_c(t) \tag{7}$$

or

$$\dot{x}_c = P_c^{-1} A P_c x_c + P_c^{-1} B L P_c x_c \tag{8}$$

then

$$\dot{x}_c(t) = H_c\, x_c(t) \tag{9}$$

with

$$H_c = A_c - B_c L_c \tag{10}$$

and

$$LP_c = L_c \tag{11}$$

such that

$$L_c = -\left[l_{c_0} \quad l_{c_1} \quad \ldots\ldots l_{c_{n-1}} \right] \tag{12}$$

L_c is the state feedback gain in the controllable base in which , the matrices A_c and B_c are written in the canonical controllable form. The characteristic polynomial of matrix A , $\mathrm{P}_A(\lambda)$.

$$\mathrm{P}_A(\lambda) = \det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_0 \tag{13}$$

is invariant by change of base. Then, we have $\mathrm{P}_{A_c}(\lambda) = \mathrm{P}_A(\lambda)$.

The matrix A_c , being in the companion canonical form, we can easily calculate the characteristic polynomial of the closed loop system characteristic matrix, noted $\mathrm{P}_{H_c}(\lambda)$,

$$\mathrm{P}_{H_c}(\lambda) = \det(\lambda I - (A_c - B_c L_c)) \tag{14}$$

By the choice of L_c , we can impose the coefficients of the characteristic polynomial such that

$$\begin{aligned} \mathrm{P}_{H_c}(\lambda) &= \mathrm{P}_{A-BL}(\lambda) \\ &= \lambda^n + \alpha_{n-1}\lambda^{n-1} + \alpha_2\lambda^2 + \alpha_1\lambda + \cdots + \alpha_0 \end{aligned} \tag{15}$$

This enables to impose the poles of the system, poles we choose real and distinct.

Once L_c determined, a simple calculation of $L = L_c P_c^{-1}$ allows to determine the state feedback into the initial base.

It comes for the closed loop initial model the characteristic matrix

$$H(x) = (A(x) - B(x)L) \tag{16}$$

the linearised closed loop system is described as following

$$\dot{x}(t) = H(0)x(t) \tag{17}$$

with

$$H(0) = A(0) - B(0)L \tag{18}$$

A suitable choice of the gain vector L enables to make the poles, of this linear closed loop system, real and distinct.

In practice, a first determination of the attractor can be achieved directly on the initial representation. Another one obtained by the use of the change of basis, which diagonalizes the linearized system at the origin, can lead to different and, very often, better results. With this change of base, the representation of the initial nonlinear system is generally diagonal dominant in the neighborhoods of the origin which enables, with a convenient definition of the comparison system, a better estimation of the attractor.

Let now P be the change of variables which diagonalizes the linearized closed loop model characterized by $H(0)$.

It comes, the corresponding diagonal characteristic matrix $H_d(0)$ such that

$$H_d(0) = P^{-1}H(0)P \tag{19}$$

By using the new state vector x_d , $x_d = \left[x_{d1}, x_{d2} \ldots, x_{dn}\right]^T$, such that

$$x(t) = Px_d(t) \tag{20}$$

$B_d^{'}$, characterizing the uncertainty in the new base, is defined by

$$B_d^{'} = P^{-1}B' \tag{21}$$

it comes for the initial non linear system

$$\dot{x}_d(t) = H_d(.)x_d(t) + B_d^{'}(.) \tag{22}$$

where $H_d = \left\{a_{d_{ij}}(.)\right\}$ is defined by

$$H_d(.) = P^{-1}H(.)P \tag{23}$$

After applying the change of base allowing to diagonalize the linearized system to the initial one's (1), we propose, in this paper, to study the stability and to determine the attractor of the initial system, controlled by the same state feedback law (3).

2.2 Proposed Attractor Determination

For the vector norm $p(x_d) = \left[|x_{d1}|, |x_{d2}| \ldots, |x_{dn}|\right]^T$ (Appendix A), the overvaluing system of the perturbed system is described by [13].

$$\frac{d}{dt} p(x_d) \le M(.) p(x_d) + N(.) \tag{24}$$

where $M(H_d(.)) = \{m_{i,j}(.)\}$ is obtained by replacing the off-diagonal elements of $H_d(x)$ by their absolute values such as

$$\begin{cases} m_{i,i}(.) = a_{d_{i,i}}(.) & \forall i = 1, 2, \ldots n \\ m_{i,j}(.) = \left| a_{d_{i,j}}(.) \right| & \forall i \ne j \end{cases} \tag{25}$$

and $N(.)$ defined by

$$N(.) = \left| B_d^{'}(.) \right| \tag{26}$$

With $M = \max M(.)$ and $N = \max N(.)$, it comes the linear comparison system

$$\dot{z} = Mz + N \tag{27}$$

such that

$z(t_0) \ge p(x_d(t_0))$ implies $z(t) \ge p(x_d(t)), \forall t \succ t_0$

If M is the opposite of an M-matrix, we can have an estimation by overvaluation of the attractor defined by

$$p(x_d(t)) \le -M^{-1}N \tag{28}$$

or

$$p(P^{-1}x(t)) \le -M^{-1}N \tag{29}$$

Then, we have

$$\lim_{t \to +\infty} z(t) = -M^{-1}N \tag{30}$$

and

$$\lim_{t \to +\infty} p(x_d(t)) \le -M^{-1}N \tag{31}$$

It comes the attractor D_1 of system (22) defined by

$$D_1 = \left\{ x_d \in R^n ; p(x_d) \le -M^{-1}N \right\} \tag{32}$$

In the domain D_1, according to the limitations that appear on the state variables, it is possible to the choose a new nonlinear model which enables to determine a better estimation of the attractor as it appears in the application of Section 3

3 Attractor Characterization of a Third Order Nonlinear Complex System

Let us consider the third order system (S) described by

$$(\mathrm{S}): \dot{x}(t) = A(x,t)x(t) + B(x,t)u(t) + B'(.) \tag{33}$$

with

$$A(x(t)) = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{34}$$

$a_{11} = 7$

$a_{12} = -2.9 - 0.1e^{-x_2^2}$

$a_{13} = -5$

$a_{21} = 0.1\sin x_1 + 6\cos x_3$

$a_{22} = 1.05 - 2.05\cos x_3$

$a_{23} = 5 - 3\cos x_3$

$a_{31} = -12 - 0.1\sin x_1$

$a_{32} = 0$

$a_{33} = -2 + 0.02\sin x_1$

and

$$B(x(t)) = \begin{bmatrix} -2 \\ -\cos x_3 \\ 2 \end{bmatrix} \tag{35}$$

$$B'(x) = \begin{bmatrix} -0.2\,\mathrm{sat}\, x_1 \\ b_2'(.) \\ 0.1e^{-x_2^2} \end{bmatrix} \tag{36}$$

such that

$$\begin{aligned} &\mathrm{sat}\, x_i = x_i, \text{if } |x_i| < 1, \text{else, } \mathrm{sat}\, x_i = \mathrm{sign}\, x_i, \\ &\qquad \text{and, } |b_2'(.)| \leq 0.15 \end{aligned} \tag{37}$$

By linearization of the system without uncertainties, around the operating point $x_0 = 0$, we obtain the linear model characterized by the following $A(0)$ and $B(0)$

$$A(0) = \begin{bmatrix} 7 & -3 & -5 \\ 6 & -1 & 2 \\ -12 & 0 & -2 \end{bmatrix} \tag{38}$$

and

$$B(0) = \begin{bmatrix} -2 \\ -1 \\ 2 \end{bmatrix} \tag{39}$$

Then, by putting the linearized system in controllable canonical form, it comes

$$\dot{x}_c = A_c x_c + B_c u \tag{40}$$

The characteristic polynomial of the linearized system can be written as

$$\det(\lambda I - A(0)) = \lambda^3 - 4\lambda^2 - 61\lambda - 110 \tag{41}$$

and we have

$$A_c = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 110 & 61 & 4 \end{bmatrix}; \quad B_c = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{42}$$

In order to impose a choosen dynamic to the process, the state feedback gain *L*, of system (17) with (18), (39) and (40), is chosen such that the poles of the closed loop characteristic $\mathrm{P}_{A(0)-B(0)L}(\lambda)$ are (-3), (-4) and (-5), i.e the characteristic polynomial:

$$\begin{aligned} \mathrm{P}_{A(0)-B(0)L}(\lambda) &= (\lambda+3)(\lambda+4)(\lambda+5) \\ &= \lambda^3 + 12\lambda^2 + 47\lambda + 60 \end{aligned} \tag{43}$$

corresponding to the following characteristic matrix H_c

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 60 & 47 & 12 \end{bmatrix} \tag{44}$$

The state feedback gain have to satisfy the following conditions

$$u = -\begin{bmatrix} 170 & 108 & 16 \end{bmatrix} x_c = \begin{bmatrix} l_{c_1} & l_{c_2} & l_{c_3} \end{bmatrix} x_c \tag{45}$$

Given that we have $L=L_c P_c^{-1}$, it comes the control vector gain

$$L = \begin{bmatrix} -6 & 2 & 3 \end{bmatrix} \tag{46}$$

and the matrix of the closed loop system without uncertainties linearized at the origin $H(0)$

$$H(0) = \begin{bmatrix} -5 & 1 & 1 \\ 0 & 1 & 5 \\ 0 & -4 & -8 \end{bmatrix} \tag{47}$$

which becomes diagonal for the change of base P defined by

$$P = \begin{bmatrix} 0.125 & 0 & -0.625 \\ 1.25 & 0.75 & 0 \\ -1 & -0.75 & 0 \end{bmatrix} \tag{48}$$

In this case, the initial system defined by (33) with (34) and (35), controlled by the control law (3) with (47), can be described by $H(x) = (A(x) - B(x)L)$ such that

$$H(x(t)) = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{49}$$

with

$h_{11} = -5$

$h_{12} = 1.1 - 0.1e^{-x_2^2}$

$h_{13} = 1$

$h_{21} = 0.1\sin x_1$

$h_{22} = 1.05 - 0.05\cos x_3$

$h_{23} = 5$

$h_{31} = -0.1\sin x_1$

$h_{32} = -4$

$h_{33} = 0.02\sin x_1 - 8$

Let us try to determine directly an attractor estimation D_1 of the initial model.

If the comparison system of the process is in the form (27), according to (49), the minimal overvaluing matrix relatively to the regular vector norm $p(x) = \left[|x_1|, |x_2|, |x_3|\right]^T$ is

$$M(H(x(t))) = \begin{bmatrix} h_{11} & |h_{12}| & |h_{13}| \\ |h_{21}| & h_{22} & |h_{23}| \\ |h_{31}| & |h_{32}| & h_{33} \end{bmatrix} \tag{50}$$

and $N(B^{'})$ is

$$N(B^{'}) = \begin{bmatrix} 0.2 \\ 0.15 \\ 0.1 \end{bmatrix} \tag{51}$$

In this case, the comparison system can be described by

$$\dot{z} = \begin{bmatrix} -5 & 1.1 & 1 \\ 12.1 & 1.1 & 5 \\ 0.1 & -4 & -7.98 \end{bmatrix} z + \begin{bmatrix} 0.2 \\ 0.15 \\ 0.1 \end{bmatrix} \tag{52}$$

For this comparison system, the matrix M is not the opposite of an M-matrix because of one of the diagonal elements is positive. Then we cannot conclude concerning the determination of an attractor.

By the use of change of variables P, $H_d(x)$ becomes such that

$$H_d(x(t)) = \begin{bmatrix} h_{d11} & h_{d12} & h_{d13} \\ h_{d21} & h_{d22} & h_{d23} \\ h_{d31} & h_{d32} & h_{d33} \end{bmatrix} \tag{53}$$

with

$$h_{d11} = -0.25\cos x_3 - 0.08\sin x_1 - 2.75$$
$$h_{d12} = -0.15\cos x_3 - 0.06\sin x_1 + 0.15$$
$$h_{d13} = 0$$
$$h_{d21} = 0.33\cos x_3 + 0.15\sin x_1 - 0.333$$
$$h_{d22} = 0.2\cos x_3 + 0.1\sin x_1 - 4.2$$
$$h_{d23} = -0.0833\sin x_1$$
$$h_{d31} = 0.2e^{-x_2^2} - 0.05\cos x_3 - 0.016\sin x_1 - 0.15$$
$$h_{d32} = 0.12e^{-x_2^2} - 0.03\cos x_3 - 0.012\sin x_1 - 0.09$$
$$h_{d33} = -5$$

The comparison system of the process, corresponding to the vector norm $p(x_d) = \left[|x_{d1}|, |x_{d2}|, |x_{d3}|\right]^T$, is in the form (27), with

$$N = \max |\, P^{-1}B^{'}(.)| \leq \begin{bmatrix} 1 \\ 1.4667 \\ 0.733 \end{bmatrix} \tag{55}$$

According to (49), the minimal overvaluing matrice relatively to the regular vector norm is the following

$$M = \begin{bmatrix} -2.42 & 0.36 & 0 \\ 0.813 & -3.9 & 0.0833 \\ 0.216 & 0.132 & -5 \end{bmatrix} \quad \text{and } N = \begin{bmatrix} 1 \\ 1.4667 \\ 0.733 \end{bmatrix} \tag{56}$$

It is trivial that the following conditions

$$\begin{cases} -2.42 < 0 \\ (-2.42 \times -3.9) - (0.813 \times 0.36) > 0 \\ \det(M) < 0 \end{cases} \tag{57}$$

are satisfied, M is then the opposite of an M-matrix (Appendix B),

and we have

$$\lim_{t \to +\infty} z(t) = -M^{-1}N \tag{58}$$

and

$$\lim_{t \to +\infty} p(x_d(t)) \leq -M^{-1}N \tag{59}$$

It comes an estimation, by overvaluation, of the attractor defined by $p(x_d(t)) \leq -M^{-1}N$, or

$$p(x_d(t)) \leq \begin{bmatrix} 0.4848 \\ 0.4810 \\ 0.1802 \end{bmatrix} \tag{60}$$

The attractor D_1 is finally defined by

$$\begin{cases} |4x_2 + 4x_3| \le 0.4848 \\ |-5.333x_2 - 6.6667x_3| \le 0.4810 \\ |-1.6x_1 + 0.8x_2 + 0.8x_3| \le 0.1802 \end{cases} \quad (61)$$

In D_1 we have $|x_1| \le 0.1732$, $|x_2| \le 0.9643$ and $|x_3| \le 0.8431$

A new description of the system (S) can be defined, in D_1

As $|x_1| \le 0.1732$ it comes, $\text{sat}\, x_1 = x_1$, then this value can be introduced in the definition of $H(x(t))$

Hence the description

$$H(x(t)) = \begin{bmatrix} -5.2 & 1.1 - 0.1e^{-x_2^2} & 1 \\ 0.1\sin x_1 & 1.05 - 0.05\cos x_3 & 5 \\ -0.1\sin x_1 & -4 & 0.02\sin x_1 - 8 \end{bmatrix} \quad (62)$$

and

$$B'(x) = \begin{bmatrix} 0 \\ b_2'(.) \\ 0.1e^{-x_2^2} \end{bmatrix} \quad (63)$$

By the use of change of variables P, $H_d(x)$ becomes such that

$$H_d(x(t)) = \begin{bmatrix} h'_{d11} & h'_{d12} & h'_{d13} \\ h'_{d21} & h'_{d22} & h'_{d23} \\ h'_{d31} & h'_{d32} & h'_{d33} \end{bmatrix} \quad (64)$$

with

$$h'_{d11} = -0.25\cos x_3 - 0.08\sin x_1 - 2.75$$
$$h'_{d12} = -0.15\cos x_3 - 0.06\sin x_1 + 0.15$$
$$h'_{d13} = 0$$
$$h'_{d21} = 0.33\cos x_3 + 0.15\sin x_1 - 0.333$$
$$h'_{d22} = 0.2\cos x_3 + 0.1\sin x_1 - 4.2$$
$$h'_{d23} = -0.0833\sin x_1$$
$$h'_{d31} = 0.2e^{-x_2^2} - 0.05\cos x_3 - 0.016\sin x_1 - 0.11$$
$$h'_{d32} = 0.12e^{-x_2^2} - 0.03\cos x_3 - 0.012\sin x_1 - 0.09$$
$$h'_{d33} = -5.2$$

the comparison system corresponding to the vector

$p(x_d) = \left[|x_{d1}|, |x_{d2}|, |x_{d3}|\right]^T$, is in the form (26), with

$$N = \max \mid P^{-1}B^{'}(.) \mid \leq \begin{bmatrix} 1 \\ 1.4667 \\ 0.2 \end{bmatrix} \tag{65}$$

Then, in D_1, the comparison system of the process is on the form (27). According to (64), the minimal overvaluing matrices relatively to the regular vector norm are the followings

$$M = \begin{bmatrix} -2.9025 & 0.0605 & 0 \\ 0.1393 & -3.9828 & 0.0144 \\ 0.0897 & 0.0783 & -5.2 \end{bmatrix}$$
$$\text{and } N = \begin{bmatrix} 1 \\ 1.4667 \\ 0.2 \end{bmatrix} \tag{66}$$

As the following conditions

$$\begin{cases} -2.9025 < 0 \\ (-2.9025 \times -3.9828) - (0.1393 \times 0.0605) > 0 \\ \det(M) < 0 \end{cases} \tag{67}$$

are satisfied, M is, then, the opposite of an M-matrix.

It comes an estimation, by overvaluation, of the attractor defined by $p(x_d(t)) \leq -M^{-1}N$

or

$$p(x_d(t)) \leq \begin{bmatrix} 0.3525 \\ 0.3808 \\ 0.0503 \end{bmatrix} \tag{68}$$

The attractor D_2 is finally defined by

$$\begin{cases} |4x_2 + 4x_3| \leq 0.3525 \\ |-5.333x_2 - 6.6667x_3| \leq 0.3808 \\ |-1.6x_1 + 0.8x_2 + 0.8x_3| \leq 0.0503 \end{cases} \tag{69}$$

The obtained attractors D_1 and D_2 are given in the Figure 1, for which a trajectory in the state space is simulated for $b_2'(.) = 0.15\sin t$.

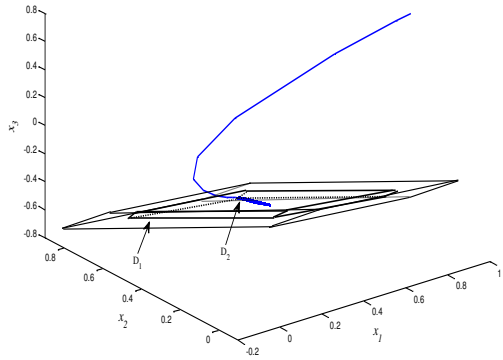


Figure 1
Evolution of the state vector towards the attractors D_1 and D_2 (in bold)

Conclusion

An efficient technique for determination of attractors characterizing the precision of a control law is defined in this paper using the concept of vector norm, associated to the definition of comparison systems obtained by the use of the Borne and Gentina stability approach. The proposed approach for determination of the control law by state or output feedback in presence of uncertainties is based on a local linearization and control of the system. Process control through poles placement of the linearized system is used in the feedback control. This method enables to test the accuracy of a controlled system by providing an estimation by overvaluation of the error. The proposed method is applied with success for a third order nonlinear complex system to illustrate the efficiency of the proposed approach.

Appendices

Appendix A. Vector Norms Definition

Definition1: Let $\mathrm{E} = \mathrm{R}^{\mathrm{n}}$ and $\mathrm{E}_1, \mathrm{E}_2 \ldots \mathrm{E}_{\mathrm{k}}$ be subspaces of the space $\mathrm{E}, \mathrm{E} = \mathrm{E}_1 \cup \mathrm{E}_2 \ldots \cup \mathrm{E}_{\mathrm{k}}$

Let x be an n vector defined on E and $x_i = P_i x$ the projection of x on E_i, where P_i is a projection operator from E into E_{i}, p_i a scalar norm (i=1,2,…, k) defined on the subspace E_i and p denotes a vector norm of dimension k and with its component

$p_i(x) = p_i(x_i)$, $p(x): R^n \to R_+^k$

Let y be another vector in space E, with $y_i = P_i y$, we have the following properties

$$\begin{cases} p_i(x_i) \geq 0, \forall x_i \in \mathrm{E_i} \forall i = 1,2,\ldots,\mathrm{k} \\ p_i(x_i) = 0 \leftrightarrow x_i = 0, \forall i = 1,2,\ldots,\mathrm{k} \\ p_i(x_i + y_i) \leq p_i(x_i) + p_i(y_i), \forall x_i, y_i \in \mathrm{E_i} \forall i = 1,2,\ldots,\mathrm{k} \\ p_i(\boldsymbol{\lambda} x_i) = |\boldsymbol{\lambda}| p_i(x_i), \forall x_i \forall i = 1,2,\ldots,\mathrm{k}, \forall \boldsymbol{\lambda} \cup R \end{cases}$$

If k-1 of the subspaces E_i are insufficient to define the whole space E, the vector norm is surjective.

If in addition the subspaces E_i are in disjoint pairs, $\mathrm{E}_i \cap \mathrm{E}_j = \varnothing$, $\forall i \neq j = 1,2,\ldots,\mathrm{k}$, the vector norm p

is said to be regular.

Appendix B. Overvaluing and comparison systems

Let the differential equation $\dot{x} = A(x,t)x$. The overvaluing system is defined by the use of the vector norm $p(x)$ of the state vector x and the use of the right-band derivation $D^+ p_i(x_i)$ proposed by [28, 29] $D^+ p_i(x_i)$ is taken along the motion of x in the subspace E_i and $D^+ p(x)$ along the motion of x in E.

Definition 2: The matrix $M(x,t)$ defines an overvaluing system of S with respect to the vector norm p if and only if the following inequality is verified for each corresponding component: $D^+ p(x) \leq M(x,t)\, p(x)$

If for the same system we can define a constant overvaluing matrix M, we have $M \geq M(x,t)$ and we have $z(t) \geq p(x(t))$ for $t \geq t_0$ as soon as this property is satisfied at the origin t_0

When an overvaluing matrix $M(x,t)$ of a matrix $A(x,t)$ is defined with respect to a regular vector norm p we have the following properties:

- The off- diagonal elements of matrix $M(x,t)$ are non negative.
- If we denote by $\mathrm{Re}(\lambda_M)$ the real part of the eigenvalue of the maximum real part of $M(x,t)$ the following inequality is verified

  $$\mathrm{Re}(\lambda_A) \leq \mathrm{Re}(\lambda_M) = \lambda_M \quad \forall t, x \in \tau \times \square^n,$$

  whatever the eigenvalue λ_A of matrix $A(x,t)$

- When all the real parts of the eigenvalues of $M(x, t)$ are negative this matrix is the opposite of an M-matrix and it admits an inverse whose elements are all non positive.
- When due to perturbations and/or uncertainties it is not possible to define an homogeneous overvaluing system we can define a non homogeneous overvaluing system of the form $D^{+}p(x) \le M(x,t)\,p(x) + N(x,t)$, where all the elements of vector norm nonnegative and when M and N are constant, we can define the comparison system $\dot{z} = Mz + N$

Remark 1. With $M(.) = \left\{ m_{ij}(.) \right\}$ the verification of the Kotelyanski lemma by the matrix $M(.)$ prove that $M(.)$ is the opposite of an M-matrix

$$m_{1,1} \prec 0, \begin{vmatrix} m_{1,1} & m_{1,2} \\ m_{2,1} & m_{2,2} \end{vmatrix} \succ 0, \ldots, (-1)^k \begin{vmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,k} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,k} \\ \vdots & \vdots & \cdots & \vdots \\ m_{k,1} & m_{k,2} & \cdots & m_{k,k} \end{vmatrix} \succ 0$$

Remark 2. A less conservative approach consists to use a vector norm of size $k=n$, for example $p(x) = \left[|x_1|, |x_2|, \ldots, |x_n| \right]^T$

Remark 3. If $M(.)$ is an overvaluing matrix of a matrix $A(.)$, $M(.) + M^{*}$ where the elements of M^{*} are all non negative is also an overvaluing matrix of $A(.)$. This property can be used to simplify the determination of an overvaluing matrix of $A(.)$ when some elements of $A(.)$ are ill defined or subject to uncertainties.

References

[1] G. Bartolini, A. Pisano, and E. Usai: Global Stabilization for Nonlinear Uncertain Systems with Unmodeled Actuator Dynamics, IEEE Transactions on Automatic Control, Vol. 46(11), pp. 1826-1832, 2001

[2] M. B. Radac, R. E. Precup, E. M. Petriu and S. Preitl: Experiment-based Performance Improvement of State Feedback Control Systems for Single Input Processes. Acta Polytechnica Hungarica, Vol. 10(3), pp. 5-24, 2013

[3] J. K. Huusom, N. K. Poulsen and S. B. Jorgensen: Iterative Feedback Tuning of Uncertain State Space Systems. Brazilian Journal of Chemical Engineering, Vol. 27(3), pp. 461-472, 2010

[4] Y. Xia, P. Shi, G.P. Liu, D. Rees, J. Han: Active Disturbance Rejection Control for uncertain multivariable Systems with Time-Delay, IET Control Theory and Applications, Vol. 1(1), pp. 75-81, 2007

[5] H. Wu: Adaptive Stabilizing State Feedback Controllers of Uncertain Dynamical Systems with Multiple Time Delays, IEEE Transactions on Automatic Control, Vol. 45, No. 9, pp. 1697-1701, 2000

[6] N. Luo, M. de la Sen: State Feedback Sliding Mode Control of a Class of Uncertain Time Delay Systems, IEE Proceedings D- Control Theory and Applications, Vol. 140(4), pp. 261-274, 1993

[7] J. C. Gentina, P. Borne and F. Laurent: Stabilité des systèmes continus non linéaires de grande dimension. RAIRO, Aôut, pp. 69-77, 1972

[8] L. T. Grujic and D. D. Siljak: Asymptotic Stability and Instability of Large Scale Systems. IEEE Trans. on Auto. Control, Vol. 18(6), 1973

[9] L. T. Grujic, J. C. Gentina and P. Borne: General Aggregation of Large Scale Systems by Vector Lyapunov Functions and Vector Norms, Inernational. Journal. of Control, Vol. 24(4), pp. 29- 550, 1976

[10] M. Benrejeb and P. Borne: On an Algebraic Stability Criterion for Non-Linear Process. Interpretation in the frequency domain. Measurement and Control International Symposium MECO, Athens, pp. 678-682, 1978

[11] L. T. Grujic, J. C. Gentina, P. Borne C. Burgat, and J. Bernussou: Sur la stabilité des systèmes de grande dimension. Fonctions de Lyapunov vectorielles. RAIRO, Vol. 12(4), pp. 319-348, 1978

[12] J. C. Gentina, P. Borne C. Burgat, J. Bernussou and L. T. Grujic: Sur la stabilité des systèmes de grande dimension. Normes vectorielles, Vol. 13(1), pp. 57-75, 1979

[13] P. Borne: Nonlinear System Stability. Vector Norm Approach, System and Control Encyclopedia. Pergamon Press, Lille, France, 5, pp. 3402-3406, 1987

[14] P. Borne and M. Benrejeb: On the Representation and the Stability Study of Large Scale Systems. International Journal of Computers Communications and Control, Vol. 3(5), pp. 55-66, 2008

[15] M. Xiaowu, W. Jumei and M. Rui: Stability of Linear Switched Differential Algebraic Equations with Stable and Unstable Subsystems. International Journal of Systems Science, Vol. 44(10), pp. 1879-1884, 2013

[16] M. Benrejeb, P. Borne and F. Laurent: Sur une application de la représentation en flèche à l'analyse des processus. RAIRO Automatique, Vol. 16(2), pp. 133-146, 1982

[17] M. Benrejeb and M. Gasmi: On the Use of an Arrow form Matrix for Modeling and Stability Analysis of Singularly Perturbed Non-Linear Systems. Systems Analysis Modelling and Simulation, Vol. 40(4): pp. 509, 2001

[18] M. Benrejeb: Stability Study of Two Level Hierarchical Nonlinear Systems. Plenery lecture. Large Scale Complex Systems Theory and Applications IFAC Symposium, Lille, Vol. 9(1): pp. 30-41, 2010

[19] P. Borne, J. P. Richard and M. Tahiri: Estimation of Attractive Domains for Locally Stable or Unstable Systems. Systems Analysis Modeling and Simulation, Vol. 78, pp. 595-610, 1990

[20] D. D. Siljac: Stability of Large Scale Systems under Structural Perturbations, IEEE Trans. On Syst. Manand Cyber, Vol. 2(5), 1972

[21] P. Borne, J. P. Richard and N. E. Radhy: Stability, Stabilization, Regulation using Vector Norms, Nonlinear Systems, 2. Stability and Stabilization. Chapman and Hall, Chapter 2, pp. 45-90, 1996

[22] A. Gharbi, M. Benrejeb, and P. Borne: On Nested Attractors of Complex Continuous Systems Determination. Proceedings of the Romanian Academy, Series A, Vol. 14(2): pp. 259-265, 2013

[23] A. Gharbi, P. Catalin, M. Benrejeb and P. Borne: New Approach for the Control and the Determination of Attractors for Nonlinear Systems. 2nd International Conference on Systems and Computer Science (ICSCS), Villeneuve d'Ascq, France, August 26-27, 2013

[24] A. Gharbi, M. Benrejeb, and P. Borne: Tracking Error Estimation of Uncertain Lur'e Postnikov Systems. 2nd International Conference on Control, Decision, Metz, France, November 2, 3, 2014

[25] A. Gharbi, M. Benrejeb, and P. Borne: Determination of Nested Attractors for Uncertain Nonlinear Systems, 6th Multi Conference on Computational Engineering in Systems Application, Marrakech, Marocco, March, 24, 26, 2015

[26] A. Gharbi, M. Benrejeb, and P. Borne: Error Estimation in the Decoupling of Ill-defined and/or Perturbed Nonlinear Processes. 19th International Conference on Circuits, Systems Communications and Computers (CSCC2015), Zakynthos Island, Greece, July 16-20, 2015

[27] A. Gharbi, M. Benrejeb, and P. Borne: A Taboo Search Optimization of the Control Low of Nonlinear Systems with Bounded Uncertainties. International Journal Of Computers Communications & Control, Vol. 11(2): pp. 158-166, 2016

[28] M. N. Rosenbrock. A Lyapunov Function for some Naturally Accuring Linear, Homogeneous Time Dependent Equations. Automatica, 1963

[29] IW. Sandeberg. Some Theorems on the Dynamic Response of Non Linear Transistor Network. Bell Syst. Tech. J. Vol. 48(35), 1969

Breast Tumor Computer-aided Diagnosis using Self-Validating Cerebellar Model Neural Networks

Jian-sheng Guan[1,4], Lo-Yi Lin[2], Guo-li Ji[1], Chih-Min Lin[3,4,5,*], Tien-Loc Le[5], Imre J. Rudas[6]

[1]Department of Automation, Xiamen University, Xiamen 361000, China
jsguan@xmut.edu.cn, glji@xmu.edu.cn

[2]School of Medicine, Taipei Medical University, Taipei 100, Taiwan
b101098025@tmu.edu.tw

[3]School of Information Science and Engineering, Xiamen University, Xiamen 361000, China, cml@saturn.yzu.edu.tw

[4]College of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen 361024, China, jsguan@xmut.edu.cn

[5*] Corresponding author, Department of Electrical Engineering and Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan 320, Taiwan, cml@saturn.yzu.edu.tw, s1038505 @mail.yzu.edu.tw

[6]Óbuda University, H-1034 Budapest, Hungary, e-mail: rudas@uni-obuda.hu

Abstract: Breast cancer is becoming a leading cause of death among women in the world. However, it is confirmed that early detection and accurate diagnosis of this disease can ensure a long survival of the patients. This study proposes a self-validation cerebellar model articulation controller (SVCMAC) neural network which can yield high accuracy of predication and low false-negative rate for breast cancer diagnosis. With its self-validation unit, the SVCMAC neural network has higher classification accuracy than the conventional CMAC neural network. The parameters of the receptive-field basis function and the weights are all updated first by training data, and the most suitable parameters are then chosen through the self-validation algorithm to retrain the neural network for better performance. Experimental results provide evidence that the SVCMAC neural network has a higher classification accuracy when compared with the BP neural network, LVQ neural network and CMAC neural network.

Keywords: cerebellar model articulation controller; breast cancer diagnosis; self-validation

1 Introduction

Breast cancer is the leading type of cancer in women, accounting for 25% of all cases worldwide. In 2012, it resulted in 1.68 million cases and 522,000 deaths [1]. It is more prevalent in developed countries [2] and is about 100 times more common among women than men [3]. Belgium has the highest rate of breast cancer, followed by Denmark and France [4]. Therefore, prompt detection, early diagnosis and active prevention can minimize the risk of unneeded suffering from this disease.

A palpable breast lump, whether benign or malignant, is a cause of anxiety to the patient. Therefore, accurate pathological diagnosis is crucial for further treatment and estimation of an outcome [5]. The key issue in the diagnosis of breast cancer is to determine whether the lump is benign or malignant, especially for patients who are not suitable for surgery. Fine needle aspiration cytology (FNAC) has become popular as a valuable tool in the preoperative assessment of breast masses and it shows high accuracy, sensitivity, and specificity. Differentiating benign from malignant lesions is one of the major goals of FNAC. However, the accuracy of FNAC, with visual interpretation ranges from 65% to 98% and is dependent on the doctor's knowledge and experience [6]. Human error would easily cause missed, incorrect or delayed diagnoses. In view of such a situation, computer-aided diagnosis technology has been widely applied to reduce a false-negative rate of breast tumor, and increase the rate of true positive [7].

There are currently many computer-aided diagnostic methods for breast cancer. Peng [8] developed a breast cancer diagnosis system from a multi-agent and probabilistic neural network, and used changes in breast tissue resistance to enhance the diagnostic accuracy. Wang [9] used the Learning Vector Quantization (LVQ) neural network model for breast cancer diagnosis and obtained a higher diagnostic accuracy. Jin [10] improved the Back Propagation (BP) neural network for breast cancer diagnosis with additional momentum and adaptive rate. A fuzzy cerebellar model neural network is designed to classify breast nodules with 92.31% accuracy [11]. A decision tree method was used for breast cancer detection with 94.74% classification accuracy [12]. A rule induction algorithm was derived from the approximate classification method and applied to a breast

cancer detection problem achieving 94.99% accuracy [13]. A neuro-fuzzy technique was proposed, and the accuracy was 95.06% [14]. A method combining association rules with neural networks (AR+NN) was proposed for the breast cancer diagnosis problem, and the classification accuracy obtained was 97.4% [15].

This study proposes a diagnostic method called self-validation cerebellar model articulation controller (SVCMAC) neural network, to distinguish between benign and malignant breast tumors according to intelligent classification. The proposed SVCMAC is a computational model of the human cerebellum [16]. Compared with a neural network, the SVCMAC has the advantages of fast learning, good generalization capability, and simple computation. Moreover, it learns the correct output response to each input vector by modifying the contents of the selected memory locations. Thus, this study used the SVCMAC neural network to classify breast lumps for computer-aided breast tumor diagnosis.

2 Breast Cancer Diagnosis

Currently, fine needle aspiration cytology (FNAC) is performed as a pre-operative test to evaluate a suspicious breast lump, where a needle is inserted into the body, and a small amount of tissue is aspirated for examination under a microscope. Then the tissue was identified as benign or malignant through observation and analysis on cell morphological changes and cell qualitative changes [17]. With FNAC becoming more reliable in diagnosing malignancy, the use of frozen-section histology had been reduced by about 80% [18]. However, FNAC has also resulted in missed diagnosis and misdiagnosis because the tissue structure is not observed. Medical research has found a close relationship between tumor characteristics and pathological features, such as lump thickness, uniformity of cell size and cell shape, which were revealed in microscope images of the nucleus of breast tumor tissues.

The Wisconsin breast cancer dataset (WBCD) was collected by Wolberg at the University of Wisconsin-Madison hospitals [19]. The dataset consists of 699 samples taken from fine needle aspirates of human breast tissue, and each sample has nine features: lump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis. The measurements are assigned as an integer value between 1 and 10. Each sample has its class label, which is either benign or malignant.

However, 16 instances were discarded because they contained unavailable value "?", the remaining 683 samples comprised 239 (35%) malignant and 444 (65%) benign cases. The SVCMAC neural network can classify the breast tumor according to these nine pathological features to reduce misdiagnoses.

3 CMAC and SVCMAC Neural Networks

The Cerebellar Model Articulation Controller (CMAC) is a type of neural network developed from a model of the mammalian cerebellum. The CMAC was first proposed as a function modeler for robotic controllers by Albus in 1975, but has been extensively used in reinforcement learning and also for automated classification in the machine learning community [20]. The CMAC has an associative memory network, and employs error correction learning to drive its memory formation process.

The conventional CMAC, shown in Fig. 1, in general, is trained by presenting pairs of input points and output values, and adjusting the weights in the activated cells by a proportion of the error observed at the output. An input "X" given by the set of vectors so that $X = \{x_1, x_2, x_3, x_4 \ldots x_n\}$ is mapped to the desired output vector "Y" given by $Y = \{y_1, y_2, y_3, y_4 \ldots y_n\}$. The mapping function "$F$" can be given by the following equation:

$$y_i = F(x_1^i, x_2^i, \cdots, x_p^i) \tag{1}$$

CMAC can be either a single-input or multiple-input system which utilizes the given mapping function "F" to compute the output. In the case of a multiple-input system where the inputs x_1, $x_2 \ldots x_n$ are simultaneously considered, hashing techniques are utilized to generate address table. Hashing can be defined as a technique for obtaining the address table when two or more values are considered as inputs to the CMAC network. The sum of all the weights that the address pointers are pointing towards is equal to the output of the CMAC. After this, the output of CMAC is compared against the pre-determined output value; the training algorithms such as least mean square, gradient decent and back propagation are then executed, and the weights of CMAC are updated till the required minimum error has been achieved at the output.

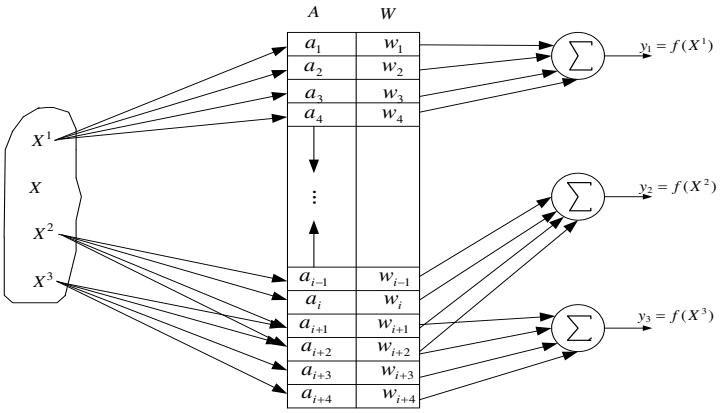


Figure 1
Structure of a CMAC network

This study proposed an advanced self-validation cerebellar model neural network as a classifier. Benign or malignant breast nodules are classified according to the pathological features extracted. Fig. 2 shows an SVCMAC neural network, which is composed of input space, association memory space, receptive- field space, weight memory space, output space, and a self-validation unit. The signal propagation and the basic function in each space are described as follows.

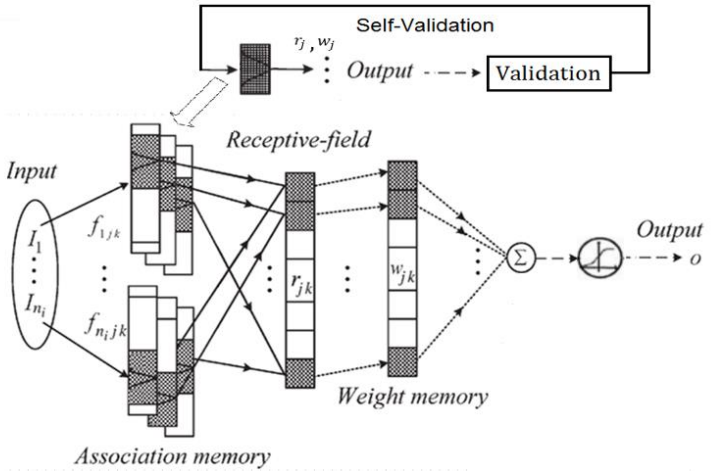


Figure 2
Structure of an SVCMAC network

1）Forward Computation

a) The Input space: $\boldsymbol{I}=[I_1, \cdots, I_i, \cdots, I_{n_i}]^T \in \Re^{n_i}$, where n_i is the number of input state variables, each input state variable, I_i can be quantized into discrete regions (called elements or neurons), according to a given control space. The number of elements n_e is referred to as the resolution.

b) The Association memory space (*Membership function*): Several elements can be accumulated as a block. In this space, each block acts as a receptive-field basis function. The Gaussian function is used here as a receptive-field basis function, which can be represented as

$$\begin{aligned} &f_{ijk}(F_{ijk}) = exp(-F_{ijk}^{\ 2}), \\ &for\ i=1,2,\cdots,n_i,\ j=1,2,\cdots,n_j,\ and\ k=1,2,\cdots,n_k \end{aligned} \tag{2}$$

where $F_{ijk}=\dfrac{I_i - m_{ijk}}{v_{ijk}}$, m_{ijk} is a mean parameter and v_{ijk} is a variance parameter.

c) The Receptive-field space (*hypercube*): The multi-dimensional receptive-field function is defined as

$$\begin{aligned} r_{jk} &= \prod_{i=1}^{n_i} f_{ijk}(F_{ijk}) = \prod_{i=1}^{n_i} exp\left(-\left[\frac{I_i - m_{ijk}}{v_{ijk}}\right]^2\right), \\ &for \quad j=1,2,\cdots,n_j,\ and\ k=1,2,\cdots n_k \end{aligned} \tag{3}$$

where r_{jk} is associated with the j^{th} layer and the k^{th} block.

d) The Weight memory space $\boldsymbol{w}$: Each location of a receptive-field to a particular adjustable value in the weight memory space can be expressed as

$$\begin{aligned} \boldsymbol{w} &= [w_{11}, \cdots, w_{1n_k}, w_{21}, \cdots, w_{2n_k}, \cdots, w_{n_j 1}, \cdots, w_{n_j n_k}]^T \in \Re^{n_j n_k} \\ &= [w_1, \cdots, w_l, \cdots, w_{n_l}]^T \in \Re^{n_l} \end{aligned} \tag{4}$$

where w_{jk} denotes the connecting weight value of the output associated with the j^{th} layer and the k^{th} block.

e) Output space: The output of the SVCMAC is the algebraic sum of the activated weighted receptive-field and is expressed as

$$y_{net} = u_{SVCMAC} = \boldsymbol{w}^T \boldsymbol{r} = \sum_{j=1}^{n_j} \sum_{k=1}^{n_k} w_{jk} r_{jk} = \sum_{l=1}^{n_l} w_l r_l \tag{5}$$

$$y_o = \frac{1}{1+e^{-y_{net}}} \tag{6}$$

where y_{net} is the output of the SVCMAC neural network and y_o is the output of the classification.

2） Backward Parameter Adjustment

a) Back propagation (BP) employed to adjust the parameters is the steepest decent algorithm that has been designed to minimize the error of an objective function defined as:

$$e = y_{ref} - y_o \tag{7}$$

$$E = \frac{1}{2} \cdot e^2 \tag{8}$$

where y_{ref} is the previously known value for the testing set;

y_{ref} = 0 for a malignant lump, and y_{ref} = 1 for a benign lump. m_{ijk} and v_{ijk} of the Gaussian function, and output weight w_{jk} are updated respectively as

$$w_{jk}(N+1) = w_{jk}(N) + \Delta w_{jk} \tag{9}$$

$$m_{ijk}(N+1) = m_{ijk}(N) + \Delta m_{ijk} \tag{10}$$

$$v_{ijk}(N+1) = v_{ijk}(N) + \Delta v_{ijk} \tag{11}$$

The updating laws (9) to (11) perform error BP with the following chain rules:

$$\begin{aligned} \Delta w_{jk} &= -\eta_w \cdot \frac{\partial E}{\partial w_{jk}} = -\eta_w \cdot \frac{\partial E}{\partial y_o} \cdot \frac{\partial y_o}{\partial y_{net}} \cdot \frac{\partial y_{net}}{\partial w_{jk}} \\ &= \eta_w \cdot \left(y_{ref} - y_o \right) \cdot y_o \cdot \left(1 - y_o \right) \cdot r_{jk} = \eta_w \cdot e_p \cdot r_{jk} \end{aligned} \tag{12}$$

$$\Delta m_{ijk} = -\eta_m \cdot \frac{\partial E}{\partial m_{ijk}} = -\eta_m \cdot \frac{\partial E}{\partial y_o} \cdot \frac{\partial y_o}{\partial y_{net}} \cdot \frac{\partial y_{net}}{\partial r_{jk}} \cdot \frac{\partial r_{jk}}{\partial m_{ijk}}$$
$$= 2 \cdot \eta_m \cdot e_p \cdot w_{jk} \cdot r_{jk} \cdot \frac{x_i - m_{ijk}}{v_{ijk}^{\ 2}} \tag{13}$$

$$\Delta v_{ijk} = -\eta_v \cdot \frac{\partial E}{\partial v_{ijk}} = -\eta_v \cdot \frac{\partial E}{\partial y_o} \cdot \frac{\partial y_o}{\partial y_{net}} \cdot \frac{\partial y_{net}}{\partial r_{jk}} \cdot \frac{\partial r_{jk}}{\partial v_{ijk}}$$
$$= 2 \cdot \eta_v \cdot e_p \cdot w_{jk} \cdot r_{jk} \cdot \frac{(x_i - m_{ijk})^2}{v_{ijk}^{\ 3}} \tag{14}$$

Where,

$$e_p = \left(y_{ref} - y_o \right) \cdot y_o \cdot \left(1 - y_o \right) \tag{15}$$

Among these, η_w, η_m and η_v are positive learning rates for w_k, m_{ijk} and v_{ijk}, respectively.

b) Self-validation algorithm

The training process will be terminated on any one of the following three conditions. First, the error of the sample is not decreased for six consecutive iterations; second, the error is equal to zero; and third, the maximum number of epochs is reached.

The data used in this study are divided into three categories, training data, validation data and test data. The training data with initial values of w_{jk}, m_{ijk} and v_{ijk} being random are employed to train the SVCMAC neural network, while the validation data test and verify the neural network. After training, w_{jk}, m_{ijk} and v_{ijk} are updated and serve as a self-validation unit to start a new SVCMAC training epoch with the same training data set and validation data in order to improve the accuracy rate. After 100 such iterations, the value of three parameters that attain the highest accuracy rate on the validation data set are saved, and finally employed to classify the test data. Fig. 3 summarizes the SVCMAC training procedure.

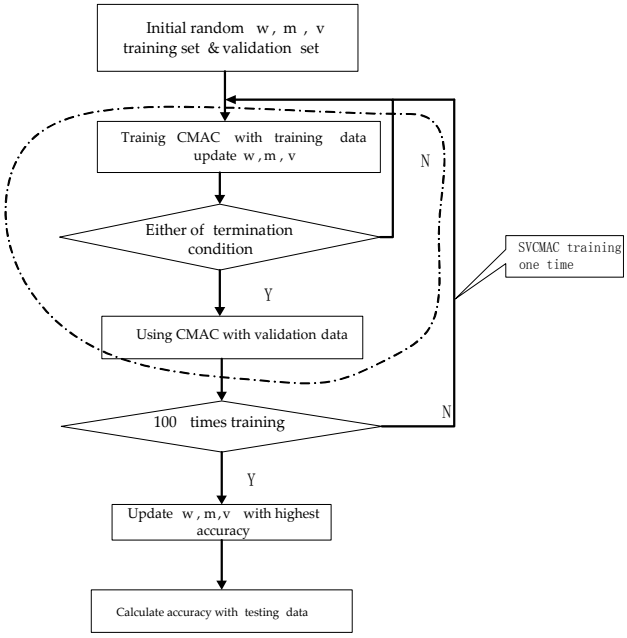


Figure 3
The procedure of the SVCMAC training

4 Performance Evaluation

The performance evaluation of the proposed method is carried out using the random data selection approach. The dataset is randomly divided into three subsets; that is, 50% of the data are for training, 30% for validation, and 20% for testing. The performance of the proposed SVCMAC neural network classification method is evaluated with sensitivity, specificity and accuracy tests. Sensitivity, specificity and accuracy terms are commonly used statistics, which uses the True positive (TP), true negative (TN), false negative (FN), and false positive (FP) terms. TP is number of true positives, denoting cases in the 'positive' class that are correctly classified as positive; FP, number of false positives, denoting cases in the 'positive' class that are misclassified as positive, and should be in the 'negative class'; TN, number of true negatives, denoting cases in the 'negative'

class that are correctly classified as negative; and FN, number of false negatives, denoting cases in the 'negative' class that should be classified as positive. Thus, sensitivity, specificity and accuracy are described in the following equations:

$$Accuracy(ACC) = (TP + TN) / (TN + TP + FP + FN) \times 100\% \quad (16)$$

$$Sensitivity(SEN) = TP / (TP + FN) \times 100\% \quad (17)$$

$$Specificity(SPE) = TN / (TN + FP) \times 100\% \quad (18)$$

TP: Malignant case identified as malignant.

FP: Benign case identified as malignant.

TN: Benign case identified as benign.

FN: Malignant case identified as benign.

5 Experiment and Results

The WBCD with nine attributes and 683 records was used in the experiment. The parameters for the SVCMAC neural network, including weight (w), mean (m), and variance (v), are also randomly initialized. The samples are randomly selected for training and testing datasets. In the end, there were 341 samples in the training dataset; 205 data were used for validation and the remaining 137 samples were for testing. The training will run 1000 iterations but can stop any time when the accuracy rate is 100%. The updated w, m and v of validation data serve as the self-validation unit of the SVCMAC, which provide new values to the next SVCMAC training epoch. After 100 iterations, the w, m, v were updated with the highest accuracy rate of validation data set.

Back Propagation neural network (BP), Learning Vector Quantization Neural network (LVQ), CMAC neural network and SVCMAC neural network were all employed to train and identify breast cancer from the same data sets. The training set contained 541 sets of data, 352 benign and 189 malignant cases; and the testing set comprised 142 sets of data, 92 benign and 50 malignant cases. It should be mentioned that the SVCMAC neural network needs the validation set only. The original training set is divided into two sets, the training set containing 341 sets of data with 220 benign and 121 malignant cases, and the validation set comprising 200 sets of data with132 benign and 68 malignant cases. Table 1 shows the simulation results of BP, LVQ, CMAC and SVCMAC.

Table 1
Classification with 80% training data and 20% testing data

Methods		BP	LVQ	CMAC	SVCMAC
Acc	Highest	97.81%	97.08%	96.35%	98.12%
	Lowest	87.59%	93.43%	93.43%	94.53%
	Avg	93.72%	94.96%	95.03%	96.5%
Sen	Highest	96.23%	92.68%	96.52%	98.5%
	Lowest	69.57%	84%	87.5%	93.35%
	Avg	84.7%	87.5%	91.49%	95.16%
Spe	Highest	100%	100%	98.89%	99.1%
	Lowest	94.05%	97.7%	93.4%	96.65%
	Avg	97.63%	99%	97.78%	97.85%

All the algorithms are executed with 10 folds.
Avg=average

Analyzing the diagnosis of BP, LVQ, CMAC and SVCMAC revealed that the SVCMAC neural network has higher diagnosis accuracy than the LVQ, BP and CMAC. As shown in Table 1, the SVCMAC neural network has higher sensitivity, which can assist physicians in making early and correct diagnosis on breast cancer, and can reduce misdiagnoses at the same time. In summary, the SVCMAC which has a higher diagnostic accuracy and lower false-negative rate is a reliable method for the computer-aided diagnosis of breast cancer.

Table 2
Classification accuracies for each fold

Folds	Number of training data	Number of test data	Correct classified	Miss classified	Correct classification rate (%)
1	541	142	136	6	95.7%
2	541	142	138	4	97.2%
3	541	142	137	5	96.4%
4	560	123	120	3	97.2%
5	560	123	119	4	96.7%

Moreover, a five-fold cross validation test was applied and the average values were calculated for performance measurements. In the cross validation test, the first three folds contained 541samples for the training dataset with the remaining 142 samples for testing; and in the last two folds, the training set contained 560 samples with the remaining 123 samples for testing. Table 2 lists the accuracy

values calculated for each fold with correct classification rates and the number of misclassified samples. As can be seen, the best performance is obtained in the second and fourth folds with calculated accuracy exceeding 97%.

Conclusions

This study proposed a SVCMAC neural network algorithm for application in breast cancer diagnosis. Compared with current breast cancer diagnosis approaches, the proposed SVCMAC neural network classifier achieves a higher accuracy rate. The textural feature method overcomes the natural drawbacks of FNAC. Hence, the SVCMAC neural network is most suitable for classifying WBCD data and is very helpful to oncologists in making the ultimate diagnosis decision.

Acknowledgment

This work was supported by Natural Sciences and Engineering Research Council of Canada, National Natural Science Foundation of China (Nos. 61573296, and 61473329), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No.20130121130004), and the Fundamental Research Funds for the Central Universities in China (Xiamen University: No. 2013121025, No. 201412G009, and No. 2014X0234), the National Science Council of the Republic of China under Grant NSC-95-2221-E-155-014-MY3.

References

[1] World Cancer Report 2014, World Health Organization. 2014, Chapter1.1

[2] World Cancer Report 2014, World Health Organization. 2014, Chapter5.2

[3] World Health Organization (2008) World Cancer Report, IARC Press

[4] http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics

[5] A. N. Pandya and N. P. Shah, "Breast Fine Needle Aspiration Cytology Reporting: a Study of Application of Probabilistic Approach," Indian Medical of Gazette, Vol. 147, No. 2, pp. 54-59, 2013

[6] A. Muratli, N. Erdogan, S. Sevim, I. Unal and S. Akyuz "Diagnostic Efficacy and Importance of Fine-Needle Aspiration Cytology of Thyroid Nodules," Journal of Cytology, Vol. 31, No. 2, pp. 73-78, 2014

[7] R. A. Castellino, "Computer-aided Detection (CAD): an Overview," Cancer Imaging, Vol. 5, No. 1, pp. 17-19, 2005

[8] Y. Peng and Y. Q. Chen, "Diagnosis System of Breast Cancer based on Probabilistic Neural Network," Journal of Hefei University of Technology, Vol. 36, No. 6, pp. 685-687, 2013

[9] B. Wang, X. X. Du and M. Jin, "Application of Breast Tumor Diagnosis based on Learning Vector Quantization Neural Network," Journal of Computer Simulation, Vol. 29, no. 8, pp. 171-174, 2012

[10] Q. Jin and P. Z. Gao, "Study on Breast Cancer Diagnosis based on Artificial Neural Network," Journal of Computer Simulation, Vol. 28, No. 6, pp. 235-238, 2011

[11] C. M. Lin, Y. L. Hou, T. Y. Chen and K. H. Chen, "Breast Nodules Computer-aided Diagnostic System Design Using Fuzzy Cerebellar Model Neural Networks," *IEEE Transactions on Fuzzy Systems*, Vol. 22, No. 3, pp. 693-699, 2014. C. M. Lin, Y. L. Hou, T. Y. Chen and K. H. Chen, "Breast Nodules Computer-aided Diagnostic System Design using Fuzzy Cerebellar Model Neural Networks," IEEE Transactions on Fuzzy Systems, Vol. 22, No. 3, pp. 693-699, 2014

[12] J. R. Quinlan, "Improved Use of Continuous Attributes in C4. 5," Journal of Artificial Intelligence Research, Vol. 4, pp. 77-90, 1996

[13] H. J. Hamilton, N. Shan and N. Cercone, "RIAC: a Rule Induction Algorithm based on Approximate Classification," Technical Report CS 96-06, University of Regina, 1996. H. J. Hamilton, N. Shan and N. Cercone, "RIAC: a Rule Induction Algorithm based on Approximate Classification," Technical Report CS 96-06, University of Regina, 1996

[14] D. Nauck and R. Kruse, "Obtaining Interpretable Fuzzy Classification Rules from Medical Data," Artificial Intelligence in Medicine, Vo1. 6, No. 2, pp. 149-169, 1999

[15] M. Karabatak and M. C. Ince, "An Expert System for Detection of Breast Cancer based on Association Rules and Neural Network," Expert Systems with Applications, Vol. 36, No. 2, part 2, pp. 3465-3469, 2009

[16] W. Li, Y. Li, W. Yang, Q. Zhang, D. Wei and W. Li, " Brain Structures and Functional Connectivity Associated with Individual Differences in Internet Tendency in Healthy Young Adults," Neuropsychologia, Vol. 70, pp. 134-144, 2015

[17] A. O. Daramola, M. O. Odubanjo, F. J. Obiajulu and N. Z. Ikeri "Correlation between Fine-Needle Aspiration Cytology and Histology for Palpable Breast Masses in a Nigerian Tertiary Health Institution," *International Journal of Breast Cancer*, article id 742573, 2015

[18] MZ. Rahman, AM. Sikder and SR. Nabi, "Diagnosis of Breast Lump by Fine Needle Aspiration Cytology and Mammography," Mymensingh Medical Journal, Vol. 20, No. 4, pp. 658-664, 2011

[19] A. Marcano-Cedeño, J. Quintanilla-Domíngueza, and D. Andina, "WBCD Breast Cancer Database Classification Applying Artificial Metaplasticity Neural Network," Expert Systems with Applications, Vol. 38, No. 8, pp. 9573-9579, 2011

[20] J. S. Albus, "A New Approach to Manipulator Control: the Cerebellar Model Articulation Controller (CMAC)," *Journal of Dynamic Systems, Measurement and Control*, Vol. 97, No. 3, pp. 220-227, 1975. J. S. Albus, "A New Approach to Manipulator Control: the Cerebellar Model Articulation Controller (CMAC)," Journal of Dynamic Systems, Measurement and Control, Vol. 97, No. 3, pp. 220-227, 1975

[21] J. S. Albus, "Mechanisms of Planning and Problem Solving in the Brain," Mathematical Biosciences, Vol. 45, No. 3, pp. 247-293, 1979

[22] C. M. Lin, Y. M. Chen and C. S. Hsueh, "A Self-Organizing Interval Type-2 Fuzzy Neural Network for Radar Emitter Identification," International Journal of Fuzzy Systems, Vol. 16, No. 1, pp. 20-30, 2014

[23] C. M. Lin, C. F. Tai and C. C. Chung, "Intelligent Control System Design for UAV Using a Recurrent Wavelet Neural Network," Neural Computing and Applications, Vol. 24, No. 2, pp. 487-496, 2014

[24] C. M. Lin and H. Y. Li, "TSK Fuzzy CMAC-based Robust Adaptive Backstepping Control for Uncertain Nonlinear Systems," IEEE Transaction on Fuzzy Systems, Vol. 20, No. 6, pp.1147-1154, 2012. C. M. Lin and H. Y. Li, "TSK Fuzzy CMAC-based Robust Adaptive Backstepping Control for Uncertain Nonlinear Systems," IEEE Transaction on Fuzzy Systems, Vol. 20, No. 6, pp. 1147-1154, 2012

[25] S. D. Teddy, "KCMAC: a Novel Fuzzy Cerebellar Model for Medical Decision Support," Artificial Neural Networks-ICANN 2008, Vol. 5164, pp. 537-546, 2008

[26] H. A. Khan, A. C. M. Tan and Y. Xiao, "An Implementation of Novel CMAC Algorithm for Very Short Term Load Forecasting," Journal of Ambient Intelligence and Humanized Computing, Vol. 4, No. 6, pp. 673-683, Dec 2013

[27] C. M. Lin and C. H. Chen, "Robust Fault-Tolerant Control for a Biped Robot Using a Recurrent Cerebellar Model Articulation Controller," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 37, No. 1, pp. 110-123, 2007

Ontology Evaluation Based on the Visualization Methods, Context and Summaries

Kristína Machová, Jozef Vrana, Marián Mach, Peter Sinčák

Department of Cybernetics and AI, Technical University, Letná 9, 04200 Košice, Slovakia, {kristina.machova, jozef.vrana, marian.mach, peter.sincak}@tuke.sk

Abstract: The paper focuses on the field of ontology evaluation and visualization. Ontologies represent the essential technology for the development of the Semantic web applications. This technology has been proven to be useful in a range of applications for data manipulation and administration. The paper introduces an ontology visualization approach based on descriptive vectors. It offers the design of descriptive vectors representation for an ontology domain and also the algorithm design for generation of the descriptive vectors. This approach offers quick overview of the given ontologies content. In addition, this work presents the design of methods for comparison and evaluation of various ontologies based on descriptive vectors. Moreover, it introduces a method for ontology placing in the context within an ontological space (the map). Finally, a method for administration of user navigation in the ontological space is presented.

Keywords: the Semantic web; ontology evaluation; ontology visualization; descriptive vector; key concept; user navigation

1 Introduction

The paper is associated with the Semantic web, which has become an inseparable part of our life. The current web is heterogeneous – it contains millions of information sources with various structures. Web users have sometimes big problems to process the enormous amount of information available on the Internet. The base of the web is represented by the technology of hypertext references, which enable interconnection of one source to another one. In such net, the user orientation is difficult. In this case, the semantics is presented in the web, but it is presented only in an elementary form of hypertext links. The links can be considered as instances of some form of relations, but without any exact semantic specification. They contain a lot of accessible information chunks, which are readable by people but unreadable by machines. Nowadays, many researchers try to find a way how to process the web information automatically by machines within the Semantic web searching. The Semantic web development brought up

the increased interest about the technologies used in the process of creation and using of the Semantic web applications (for example XML, RDFS for metadata and RDF, OWL for ontologies). The Semantic web assumes a new knowledge deduction from the knowledge explicitly presented in some given ontology, for which logics – another the Semantic web technology – can be used.

The Semantic web development stimulates the increased interest in the ontologies, in the developing standards and creating ontologies regarding these standards. This effort has resulted in a great number of ontologies available on the Internet these days, sustaining the premise that ontologies are useful in many areas: digital libraries, management of information, the Semantic web or knowledge based systems. For example, in the paper [22] an approach towards modelling a classical expert system using an ontology-based solution is presented. The increasing amount of accessible ontologies leads to a need of methods for effective visualization of the ontology structure in order to support ontology management and searching. The [11] presents an application of ontologies and semantic technologies to the creation of an enhanced management system. The [5] presents a tool-based semantic framework that uses ontology and requirements boilerplates to facilitate the formulation and specification of security requirements.

The increased interest in the ontologies evokes a need of their processing, indexation, searching and reusing. Within the last few years, some platforms for ontology searching like Swoogle [6] and Watson [3] were developed. The work [10] uses the Semantic web technologies to enable content representation to be independent of particular content presentation platforms. But no tool supports decision making connected to the question, which ontology is the best one from the point of view of a user and his/her domain of interest. The decision making and appropriate ontology searching can be made very complex by means of the keywords ambiguity and lack of explicit data about the domain that the ontology covers. The paper [17] introduces a domain specific language called SWSM for a model-driven development of web services. Also, the language of the ontology has to be taken into account. The work [1] proposes a semi-automatic procedure to create ontologies for different natural languages.

In this work we argue to develop an easy way enabling to obtain a general impression of what a particular ontology is about. The mentioned decision process can be supported by ontology visualization tools for the inspection of all the concepts of an ontology and their relations. This is the reason for considering ontology visualization. The aim of the paper is to propose a new tool for ontology visualization using descriptive vectors – the tool represents a novel approach to providing a quick overview of the content of given ontologies without necessity for long searching and exploration of the ontologies. This approach is an alternative to FCA (Formal Concept Analysis) [25], but our approach is more effective, because it is not so exhaustive and complex. The descriptive vector of an investigated ontology and the descriptive vector of the domain of user interests are used for user navigation within the process of searching and selection of an

appropriate ontology. This work presents a design of methods for comparison and evaluation of different ontologies based on descriptive vectors. Within the design process, the ontology visualization and evaluation are not considered separately. The ontology visualization and evaluation create design environment simultaneously serving for all needed functions in one integrated access.

2 Ontology Visualization

Because ontologies can reach extreme size and complexity, the developing of ontology visualization tools is necessary. Ontology visualization methods according to [13] can be divided into the following groups: *Indented lists, Graph and tree structures, Zoom-able techniques, Space-filling techniques, Focus + context or distortion access, 3D information landscapes*.

The *indented lists* form a simple and intuitively understandable group of visualization methods, which presents classes as nodes in indented collapsible tree. System Protégé [12] serves such representation. A disadvantage of it lies in lucidity connected with large number of classes and higher number of nested levels.

The *graph and tree structures* represent a very suitable metaphor for the ontology. The hierarchy and various types of relations between ontology objects can be well plotted by a graph or a tree. An example of this group of methods is the tool OntoViz [20], which is a plug-in for Protégé. Disadvantages of using this technique are: a lack of interactions, problems with navigation, a lack of a searching tool and low effectiveness of utilization of the space on a display.

The *zoom-able techniques* visualize nodes from lower level inside their parent nodes. An example of the technique is CropCircles [19], which visualizes ontology in the form of hierarchy of concentric circles. This group of techniques is suitable for searching ontology with the purpose of finding a concrete object. These tools do not provide understandable picture of the ontology structure.

The *space-filling techniques* aim at the best utilization of a space. They divide the space available for some representational node into rectangles. Each rectangle belongs to one descendant of this node (for examples TreeMaps for two and three dimensional space [24]). Disadvantage of them is a lack of space remaining for internal nodes. They are inadequate for an ontology structure visualization.

The *focus + context or distortion access* is based on the combination of a focusing method and a context. Usually, one node is in the centre and other nodes are situated around. Because of hyperbolic transformation, the larger distance is between those nodes, which are situated near the centre. Typical representatives of this access are 2D Hyperbolic Tree and 3D Hyperbolic Tree [21], constituting a

tree structure in two or three dimensional space. The tree root is situated in the centre and its descendants are placed around. When an actual node is changed, the tree visualization is rearranged around a new centre. It is suitable for providing a global view on ontology. Disadvantages of such methods are incompleteness of information about some of the nodes and continual redrawing of the graph.

The *3D information landscapes* methods locate ontology into a map and in this way define the context of the given ontology and its relation to other ontological documents on this map. The collocation is provided on the basis of relations between ontology and map components. The map can represent one or more domains. The examples of systems belonging to this category are File System Navigator (FSN) and Harmony Information Landscapes (HIL) [9]. The nodes are represented by three dimensional objects located on the map. Attributes of ontology documents are coded by colour and size of the given objects.

We were inspired by a metaphor of this map. Information in an ontology is usually too extensive to be visualized globally in its whole complexity. So we were motivated to design a visualization method, which allows information filtration and focusing on key concepts of the ontology. Our aim was to enhance readability and fast orientation within the ontology to improve the user navigation in the ontology space.

3 Ontology Evaluation

Ontology evaluation is a process of the determination of a measure in which a given ontology meets some defined criteria [2]. This process is often specialized to be able to identify a domain, the ontology logically belongs to. This domain may be covered by a given ontology in different measure and with different level of granularity. Known ontology evaluation methods can be divided into the following approaches: an approach based on *comparison with a gold standard, evaluation of results of applications using a given ontology, comparison with data sources,* and *human evaluation*. Another dividing of the evaluation methods according to the evaluation level is following: *lexical-data level, hierarchically-taxonomical level, semantic relation level, contextual-application level, syntactic level and architecture and design level*. Table 1 illustrates relations between these six evaluation levels (the first column) and four previously mentioned approaches to ontology evaluation (the first row). In case, that a relation exists (for example between lexical data and gold standard), the related cell is marked with "X".

The l*exical-data level* evaluation techniques focus on the concepts, instances and facts in the ontology. In [15], the evaluation technique is to describe the measures of similarity of two strings by a number from interval [0,1]. Each string from the first set is compared with each string from the second set. In principle, this

technique represents a comparison of all the names of concepts from a given ontology with the concepts of a gold standard. The gold standard can be represented by a group of strings being considered as a good representation of concepts from the given domain. The gold standard can be another ontology, concepts generated from text documents or defined by experts in the given domain.

The *hierarchically-taxonomical level* evaluation does not focus on the analysis of the objects (as previous access), but focuses on analysis of the structure of relations between these objects.

Table 1
Survey of ontology evaluation techniques - source [2]

Level	**Ontology evaluation approach**			
	Gold standard	Application based	Data comparison	Human based
lexical-data	x	x	x	x
hierarchically-taxonomical	x	x	x	x
semantic relation	x	x	x	x
contextual-application		x		x
Syntactic	x			x
architecture and design				x

The *semantic relation level* evaluation includes all types of semantic relations. Very often, it contains precision and recall computing.

The *contextual-application level* evaluation techniques focus on a context creation and evaluation in the framework of a real application. Various ontology documents can have mutual relationships between their parts or concepts. The relationships enable to connect given ontologies into one model and to create a formal and consistent domain description. This is the way, a context can be created. The ontologies are not intended for direct interactions with users. They are in the form, which is intended for reading by machines and common people (not experts) would have problems to read them. They are primarily designed for using in applications as an auxiliary source of information. Therefore, the quality of a used ontology influences the results of these applications and similarly good results of the application entitle us to presume good ontology quality. An approach for calculating the distance between two ontology concepts is described in [23]. The results are compared with a gold standard provided by an expert.

The *syntactic level* of evaluation is focused on manually created ontologies. These ontologies are written in some particular programming language. They fulfil the specifications of the used language. This fact can be utilized within the testing of the ontologies.

The *architecture and design level* is processed manually, mainly in the case when the ontology has to fulfil some predefined criteria.

4 New Evaluation Method Using Visualization

The main objective of our work is to navigate users in a large ontology space and help them to select a suitable ontology for their needs, interests or systems. The mentioned objective belongs to the field of user personalization and personalized web recommender systems [26]. To achieve the objective we decided to use ontology evaluation methods. This approach can be successful only with the aid of an effective ontology visualization method which enables to create a smart and quick picture of the content of an examined ontology. Therefore, we decided to design a combination of ontology visualization and evaluation methods based on descriptive vectors, inspired by a metaphor of 3D information maps. We have chosen this model, because it enables not only convenient user navigation in a large ontology space, but also it enables to express relations and even proximity of particular ontologies. The measure of the closeness or even of the diffusion/interleaving of ontologies in the 3D information map intuitively expresses the measure of semantic similarity. This property distinguishes our approach from other approaches.

The existing visualization techniques are too complex and thus they are inadequate for quick ontology searching and evaluation. We have designed an approach enabling reusing an ontology in a specific application even if the ontology was developed for a different purpose. Our approach enables an effective search of information within ontologies. Main steps of our design of the ontology visualization process are following:

- generating of a vector description of a domain
- generating of a vector description of an ontology
- comparing the two descriptive vectors
- visualizing in a context
- navigation in an ontology space.

4.1 Vector Description of a Domain

The aim of the vector description of a domain is to summarize all available information about the domain and to insert them into the vector in the compressed form. Each domain can be represented by one so called descriptive vector. The vector can be compared with an ontology descriptive vector to evaluate the

measure of coincidence. The concept "domain" can be defined as a field of knowledge represented by entities, their relations, attributes, their values and rules, which associate elements on the higher level of generality. Formally consistent sources of information about some domain can be just ontologies. Therefore we decided to use domain-oriented ontologies as sources of information for acquisition of the descriptive vector of a given domain. There are the following preconditions in regard to ontologies:

- the natural language used to define ontologies is English
- the syntactical properties of English are exploited in searching on ontologies
- the label plays the role of the denotation for a concept as a node in a complex network and it is not used for edges that represent relationships.

Some concept within the ontology can be represented by its label – name pair as well as by other concepts within its environment (consisting of the closest concepts). They both define concept's semantics. For example the concept "soul" accompanied by an environment "music, blacks, rhythm" has different semantics as the concept "soul" accompanied by the environment "spirit, psyche, animus". The representation of the context of a term (including its environment) is a vector of words – labels of concepts. The existing visualization techniques visualize ontology content in the form of a complex and complicated graph. That is why we have come with a solution, which can compress data from different ontologies and consequently about domain, in the form of domain descriptive vector d_i (1):

$$d_i = [(c_{i1}, w_{i1}),...,(c_{iM}, w_{iM})]. \quad (\mathbf{1})$$

Symbols w_{ik} are weights of the concepts c_{ik} with relation to the domain d_i. The number of domains is N: i Є [1,N]. Each of these vectors represents "gold standard" of the given domain. Unlike a classic gold standard, which was created manually, the gold standard within our approach was created in an automatic way by analysing contents of the related ontologies.

Within the descriptive vector of a given domain creation, all relating ontologies with respect to the domain have to be searched. At the beginning of this process, user has to enter a key term characterizing the given domain. The key term (key concept) can be an object of Class type. It cannot be an object of Individual or Property types. The Semantic web browser finds all the ontologies, containing this key term. All concepts from the nearest environment of the key term in the given ontology are selected and saved with their status. The information will be used for weight calculation for the descriptive vector of the given domain. Figure 1 illustrates examples of the ontologies, which were selected, because they contain the key term "*academic employee*" (red colour).

The nearest environment of the key term "*academic employee*" in our ontology example can be found in the left part of Figure 1 (green colour). It contains all

nodes (*owl:Thing, lecturer, PhD student*), which are related directly to the original key term. The given key concept was found also in the ontology example on the right side of Figure 1 in the form of the term "*academic*", with its nearest environment (*academy, professor in academy, researcher in academy*). Our approach considers also a partial match between the key term given by the user ("*academic employee*") and the key term found in the ontology ("*academic*"). All selected terms are inserted into the descriptive vector together with their weights. The weight of a term represents its semantic closeness to the key term. The weight is calculated in the following way. At first, initializing weights for key term (exact match, *academic employ*) and for similar terms (partial match, *academic*) are calculated.

These terms (red colour in Figure 1) are marked as *original concepts*. Each of the considered ontologies contains one original concept. The weight initialization of the original concept is calculated according to the type of the given object:

- object Class: $w_0 = 10 + G$
- object Individual: $w_0 = G$
- object Property: $w_0 = 1$.

where the object "Class" represents a group of similar "Individuals" and object "Property" represents some relation, for example some relation between classes. Intuitively, the object Class has greater weight than for example the Individual, because it contains more individuals and therefore it represents a concept, which is generally more valuable for visualization.

The coefficient $G \in [1,10]$ represents a generality level of the concept, where G = 1/10 represents minimal/maximal generality of the concept.

A *superior concept* is the concept containing a link to the original concept (the original concept is a type/subclass of the superior concept). The weights of superior concepts from the nearest environment of the original concept (*owl:Thing* for original concept *academic employee* in Fig. 1) are calculated according to (2):

$$w = \frac{w_0 + G}{l}. \qquad (\mathbf{2})$$

The parameter l is the number of words in the label of the concept. For example, "owl:Thing" has l=2 and "MusicalExpression" has l=1 (Figure 3). The way of computation of this parameter ensures the preference of concepts with smaller number of words in their labels. These concepts have higher information value. On the other hand, the Class - concept that is labelled with multiple words has lower information value and also lower weight. The parameter "l" is not taken into account into original weight w_0, because w_0 and l are two parameters used in computing the final weight w. The original concept that matches only partially to the given key concept gets the different weight as the fully matching original concept.

An *inferior concept* is the concept containing a link from the original concept (the inferior concept is a type/subclass of the original concept). The weight of inferior concepts from the nearest environment of the original concept (*lecturer, PhD student* for original concept *academic employee* in Figure1) is computed according to the formula (3):

$$w = \frac{w_0}{l}. \tag{3}$$

Both, superior and inferior concepts take into account the existing hierarchy of the ontology. The main difference between the superior and the inferior concepts is that the superior one contains "a link to" and inferior one contains "a link from" the original concept. They cannot have the same weights as the original concept, which is the core of the descriptive vector and which is more similar to the key word, even if their labels contain the same number of words.

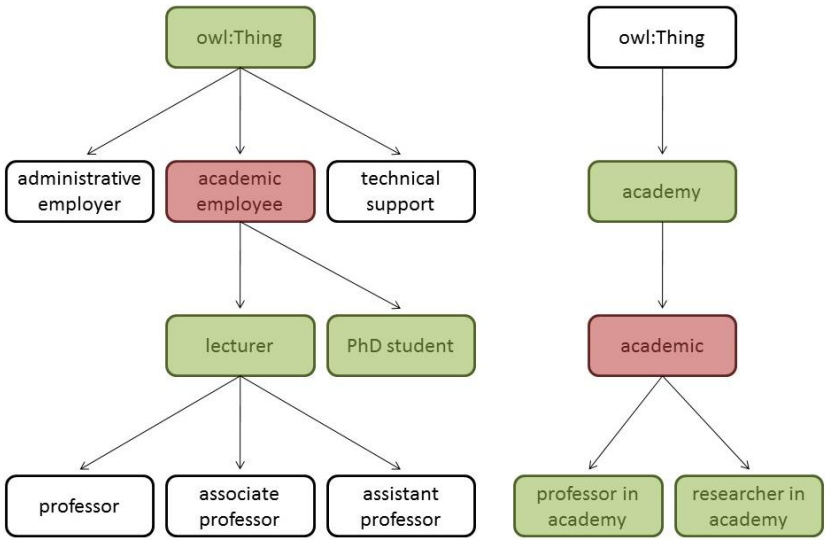


Figure 1
Tree-like graphs of two ontologies (type of using relation is "subClassOf")

For the key concept "a*cademic employee"* and *G=1* the following concepts from the ontology on the left in Figure 1 are collected into the domain descriptive vector:

[(academic employee, 11), (owl:Thing, 12), (lecturer, 11), (PhD student, 5.5), ...].

The numbers in this vector represent weights of the vector concepts. The change of the parameter *G* into value *G=10* leads to the following modification of the descriptive vector:

[(academic employee, 20), (owl:Thing, 30), (lecturer, 20), (PhD student, 10), ...].

In the case of higher occurrence of labels consisting of more words, the difference of vector modification is more significant. The descriptive vector is only a particular vector, which was created from the left ontology in Figure 1. The whole domain descriptive vector comes into being by aggregation of all particular vectors coming from all ontologies containing the given key concept. The aggregated domain descriptive vector represents summary of all particular vectors, which were derived from the related ontologies.

Various particular vectors can contain the same concept but with different weights. The concept is inserted into the aggregated vector only once with the weight, which is aggregated from all weights of the given concept coming from all particular descriptive vectors.

The next step is normalization and reduction of the domain vector. The normalization represents transformation of all weights into the interval [0,1] for the purpose of future comparison of the domain descriptive vector with an ontology vector. Consequently the concepts with lower weights than a given threshold T are eliminated from the vector (experimentally was stated the threshold T = 0.0005).

4.2 Vector Description of an Ontology

The vector description of an ontology obtains descriptive vectors of the key concepts of the ontology with their weights. The concepts on the most abstract levels of the ontology are not suitable for the role of the key concepts, because they are too general. Similarly, the concepts on the lowest levels of the ontology are too specific for common users. The most suitable and informative levels are those in the middle of the ontology taxonomy. This idea was used in a method developed in Knowledge Media Institute in Open University in Great Britain within the project NeOn [4]. The method is based on the search of n concepts, which describe the ontology in the best way – key concepts of the ontology. The method tries to maximize centrality of the concept (maximum number of appearances in all paths from the root of the ontology) and to minimize the number of words in the concept label. In addition, the method tries to maximize the density of the concept (the number of concept instances or its frequency) and the concept coverage (the number of other key concepts in the ontology, which belongs to the sub tree of the given concept). All key concepts with the highest information value represent the ontology summary. The descriptive vector of the key concept contains also concepts from its environment and it reflects only one ontology context. The relevant concepts for the inclusion into the descriptive vector of a key concept are all ancestors and all descendants of this key concept as illustrated in Figure 2.

The weights of concepts in the descriptive vector are calculated according to formulas (2) and (3) with the initializing weight value equivalent to “10” and

"G=5". For example in Figure 2 the description of the key concept "*supervisor*" (in the form of the descriptive vector) is:

[(supervisor, 15), (agent, 20), (owl:Thing, 10), (professor, 15), (senior researcher, 7.5), (assistant professor, 7.5), (associate professor, 7.5)].

The descriptive vector of each key concept must be normalized into interval [0,1] for the purpose of enabling subsequent comparison with a domain descriptive vector. The terms with weights lower than the threshold value *T=0.0005* are eliminated from the key concept vectors. After carrying out the mentioned steps, the descriptive vector of the key concept "supervisor" will be the following:

[(agent, 1), (supervisor, 0.75), (professor, 0.75), (owl:Thing, 0.5), (senior researcher, 0.375), (assistant professor, 0.375), (associate professor, 0.375)].

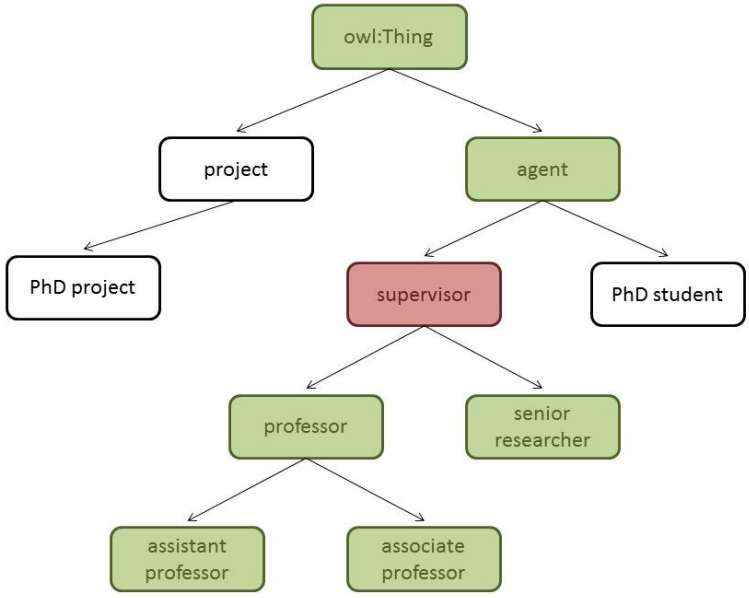


Figure 2

Key concept (red colour) of the ontology together with its relevant concepts (green colour). In this case, not only the nearest environment of the key concept is taken into account but also the rest of antecedents and descendants of the key concept, because the nearest environment would be represented by only three nodes

4.3 Comparison of Descriptive Vectors

Our approach uses the well known cosine metric of the similarity between the vectors of the ontology and the domain. According to [14] the cosine metric of the similarity is suitable for short texts. The metric expresses a cosine of the angle

between the two vector representations in the coordinate system – the domain descriptive vector and the vector of the context of the ontology. The key concepts are located in the domain space and their coordinates are determined by their similarity measure within the given domain. Each of the similarities $S(x_i, x_j)$ is calculated, where x_i is vector of the *i-th* ontology (i Є [1,M]) and x_j is vector of the *j-th* domain (j Є [1,N]). The resulting similarity matrix is following:

$$S = \begin{bmatrix} S_{11} & S_{1M} \\ S_{N1} & S_{NM} \end{bmatrix}. \quad (4)$$

The similarity matrix can be used for various purposes. For example, with the aid of the similarity matrix, the best location of the ontology key concepts in the domain space can be assigned within the visualization method. Other application possibilities are an automatic ontology evaluation and an automatic ontology searching according to user requirements and criteria. One of such criteria can be represented by a set of domains, which must be covered by the given ontology in the significant measure. Another criterion can be searching for only one domain covered by the given set of ontologies in the best way. Next possible applications can solve the problems of ontology comparison, combination of ontologies into larger information systems or key words extraction from a domain. We do not focus on the WordNet, because we utilize a content of all available ontologies within the web space.

5 Implementation of the Designed Method

From the view of our implementation of the designed method, two tools for acquisition of the ontological data were considered: Swoogle [6] and Watson [3]. Swoogle is an older tool and it cannot distinguish different versions of an ontology. The tool also administrates ontologies only on the level of documents and it cannot provide functions for access to objects into the given ontology. On the other hand, Watson can distinguish various versions of the same ontology and can manage accesses to the stored ontologies and enables their reusing, which is a very important feature. On the basis of these facts, the tool Watson was selected.

The designed method - combination of ontology visualization and evaluation based on descriptive vectors - was implemented as a system called OntoSumViz and subsequently tested. This system works in three steps: semantic content acquisition, concepts processing and cache filling.

The *semantic content acquisition* was executed with the aid of the Watson system using Java Client API. The module downloads the ontologies containing a given key word. It extracts sub-trees (which contain the key concept) from the searched ontologies and its nearest environment together with information about their types

and relationships. The sub-trees are assigned to the original concept found in the given ontology.

Within the module of *concepts processing*, the weights are initialized. In the next step, the weighting scheme is applied for modification of concepts weights, aggregation and normalization. The resulting concept descriptive vector is an input for the comparator. It compares the descriptive vectors of given domains and the descriptive vector of a key concept in an ontology and subsequently allocate the domain for the given ontology.

In the last step a *cache is filled* in order to save descriptive vectors for next reuse. The cache module checks whether the descriptive vector for the key concept given by a user occurs in the buffer. Only in the case when it was not found, the implementation OntoSumViz starts computing of a new descriptive vector. To inspect the number of the ontologies containing the given key concept, it is necessary to call the special service, which takes 11 seconds, which is an average value of the time response of the special service. It illustrates the system as extra time consuming. For the vector with 200 terms, the service has to be called 200 times. It represents a considerable delay. Therefore we decided for another solution. It is using the vectors from the cache in the role of the corpus of the ontologies.

Our implementation of the designed method is realized as a module of OntoSumViz. The novelty of our approach is the design of descriptive vectors computing. This approach offers quick overview of the given ontologies content without long searching and exploration.

5.1 Visualization of the Context within the OntoSumViz

The approach, which was described within previous sections, can be used in many ways. We use it for creating the context of the ontology. It was mentioned, that the concepts with the significant information value are situated in the middle level of the ontology. Our implementation offers these concepts to user automatically within the ontology visualization. The implementation characterizes the meaning of the concepts with the aid of their environment and in this way it makes easier the decision making process of the user about suitable ontology. A principle of gradual uncovering of the ontology content is consistent with user mental model creation.

The whole number of middle level concepts in the visualized ontology is divided into three levels of significance. In each level, the same number of concepts occurs. In the case, when twelve concepts are added on the sheet, these twelve concepts are divided into three significant levels and each level contains four concepts. The process of the visualization of these concepts is the following: at the first moment, only the four most important concepts for the decision about the

ontology suitability and the measure of interest of the given ontology for a user are visualized. Next, other four less important concepts are displayed and, at the end, the four least important concepts are visualized.

5.2 Ontology Location in the Context

Within our implementation, the ontology is represented by a set of key concepts. The user can change the level of details by closing in respectively secluding (zooming out) the used view. To specify the meanings of the key concepts, the metaphor of a geographical map is used. The map is represented by the space, which is divided into sectors (9 sectors in Figure 3). The sectors represent different domains – different possible user's interests. The ontology can (intuitively) exceed the borders of one domain. The key concept is defined by its environment on the map. Visualization of such a map and visualization of the selected ontology at the same time are illustrated in Figure 3. The implementation of the OntoSumViz is a component of the tool NeOn Toolkit [8], [16].

The domains (their names use capital letters) are represented by characterizing concepts (using small letters) in the related sectors. The terms originate from the descriptive vectors of the given domains. The descriptive vector of the domain cannot be displayed completely because of its cardinality. Due to this reason, only six the most important terms are displayed. As an example, the key concepts of the ontology "musicontology.rdf" (including their environment) are located onto map (yellow colour) on the basis of the similarity calculation between the descriptive vector of the domain (nine descriptive vectors are used, one for each domain in the example in Figure 3) and the descriptive vector representing the ontology key concepts. The ontology key concepts are situated in those domains, whose vectors are the most similar to the given key concepts vector. The ontology key concepts are situated in the positions, which are nearest to the most similar concepts of particular domains.

The ontology key concepts can be distributed to more than one domain. Figure 3 illustrates distribution of the ontology key words into 6 domains, because relations to the other three domains are marginal insignificant. Thus, it can be said, that the ontology belongs mainly to two domains: ARTIST and MUSICAL GROUP. If there is some direct relation between two concepts, then this relation is displayed by an arrow (see Figure 3). The presented implementation can be used also for displaying more ontologies on the same map.

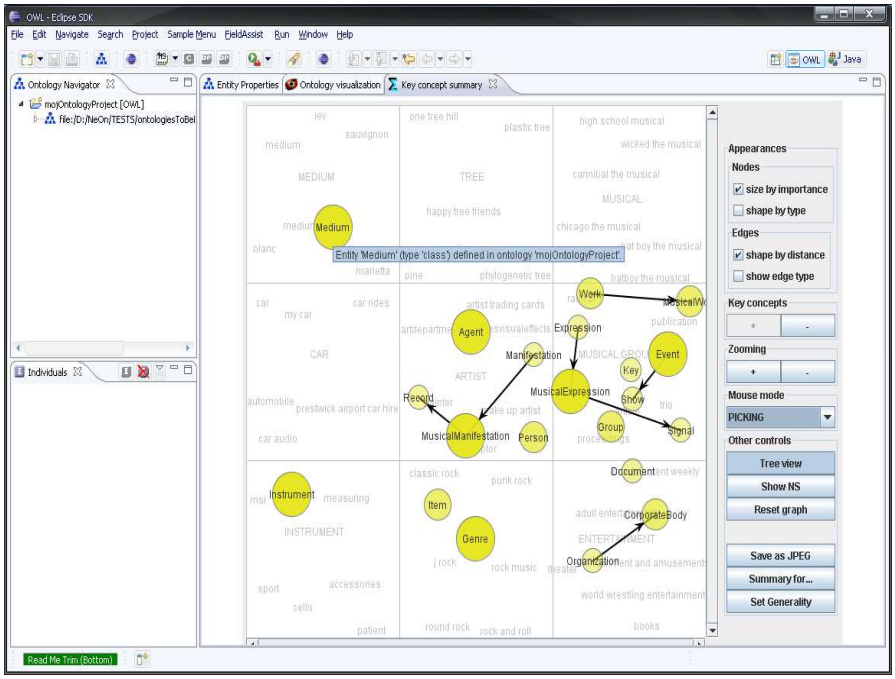


Figure 3

The screenshot of the NeOn Toolkit [16] with a map and with visualization of an example ontology using OntoSumViz (right window - control panel of the OntoSumViz)

5.3 Navigation in the Ontology

Navigation in the ontology is performed using the control panel of the tool OntoSumViz illustrated in the right window in Figure 3. The panel contains buttons, which are grouped into the following blocks: Appearances, Key concepts, Zooming, Mouse node and other controls.

Within the block "Appearances", presentation of nodes and edges of the ontology graph can be set up. The block "Nodes" contains two possibilities: "size by importance" and "shape by type". The size by importance selection enables to set a size of displayed key concepts according to their importance. The key concepts, which are displayed on the first significance level of approaching (see Section 5.1) have higher importance and therefore are of bigger size than the key concepts from other significance levels. The shape by type selection distinguishes key concepts according to types. A key concept of the type class is represented by a circle and a key concept of the type instance is represented by a square with rounded edges. The block "Edges" contains two possibilities of edge form setting: "shape by distance" (a thick link represents the relation between concepts

(subClassOf) and a thin link represents the relation between class and instance (instanceOf)) and "show edge type" (each link is signed by its name and type).

The blocks "Key concepts" and "Zooming" enable application of the metaphor of a geographical map. User can enlarge some part of the map and see this part in more details. The block "Zooming" provides traditional closing (buttons "+" and "-") and the block „Key concepts" provides contextual zooming – contextual navigation, when more detailed view of particular concepts is provided. The higher/lower level of significance can be achieved by buttons "+"/"-".

The block "Mouse node" enables manipulation with the whole graph ("transforming" – implicit setting) or moving one node of the ontology graph in the case, when two concepts (nodes) overlap ("picking" selection).

The last block "Other controls" enables access to the following menu possibilities:

1. "Tree view" button switches between the ontology displaying on the map and its displaying in the form of a graph.

2. "Show NS" button shows the whole name of the key concept.

3. "Reset Graph" button reinitializes ontology and locates it into the map in the case when some changes were performed, for example changing of the parameter generality.

4. "Save as JPEG" button enables saving of the map or the graph in the form of a picture.

5. "Summary for" button sets actual user.

6. "Set Generality" button" enables setting of parameter G from interval [1,10] (as an implicit setting is used the value 5).

6 Experimental Analysis

A set of experiments with the OntoSumViz implementation was performed with the aim to verify the designed methods. The tests were focused on the following issues:

- possibility to use the vector description of a domain as a golden standard of the given domain,
- precision of the designed methods and effectiveness of the implementation OntoSumViz within the user navigation in an ontology space,
- comparison of the decisions provided by the implementation OntoSumViz against decisions of experts.

6.1 Vector Description of a Domain as a Golden Standard

The golden standard of some domain is an etalon of the domain, which can be created by some expert in the given domain. We wanted to know, if our implementation OntoSumViz can be applied for building this golden standard and how many ontologies are necessary to be used for this golden standard building. In our case the golden standard would have the form of a set of concepts - items of the descriptive vector of the domain.

Table 2
Degree of matching between one of domain "Academic Employee", "Project" and "Object" and the seven selected reference domains ("Instrument", "PhD Project", "Student", "Education", "Music", "Supervisor" and "Entertainment")

Academic Employee

	Phd Project	Student	Instrument	Education	Music	Supervisor	Entertainment
10	0,2550	0,0997	0,0000	0,0210	0,0000	0,0577	0,0007
100	0,4308	0,0632	0,0392	0,0159	0,0019	0,2623	0,0000
200	0,2314	0,0578	0,0595	0,0338	0,0102	0,1494	0,0614
300	0,2374	0,0571	0,0595	0,0000	0,0093	0,2128	0,0526
400	0,2350	0,0512	0,0377	0,0380	0,0064	0,2113	0,0587
500	0,3200	0,0389	0,0270	0,0836	0,0069	0,2276	0,0509

Project

	Phd Project	Student	Instrument	Education	Music	Supervisor	Entertainment
10	0,8602	0,0433	0,0127	0,0000	0,0131	0,0193	0,0180
100	0,8371	0,0482	0,0038	0,0138	0,0043	0,0282	0,0088
200	0,8838	0,0251	0,0049	0,0080	0,0045	0,0091	0,0100
300	0,8475	0,0129	0,0026	0,0100	0,0021	0,0046	0,0058
400	0,8204	0,0142	0,0033	0,0030	0,0020	0,0034	0,0043
500	0,7583	0,0099	0,0021	0,0025	0,0008	0,0035	0,0036

Object

	Phd Project	Student	Instrument	Education	Music	Supervisor	Entertainment
10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
100	0,0033	0,0007	0,1003	0,0456	0,0233	0,0003	0,0330
200	0,0038	0,0031	0,0900	0,0823	0,0344	0,0014	0,0920
300	0,0050	0,0132	0,1366	0,0999	0,0477	0,0024	0,1084
400	0,0062	0,0261	0,1556	0,1182	0,0543	0,0047	0,1238
500	0,0062	0,0360	0,1625	0,1135	0,0538	0,0053	0,1264

Some experiments concerning the golden standard were performed. The experiments were carried out to find an optimal number of ontologies needed for computing the descriptive vector on the satisfied level of precision. We tried to verify also suitability of using the generated descriptive vector of a domain as a

golden standard. We have performed a series of six experiments with various numbers of ontologies used for developing a descriptive vector (MaxOnt = 10, 100, 200, 300, 400, 500 – MaxOnt is the maximal number of the used ontologies) which can be seen in rows of Table 2. The experiments showed the degree of matching (values in the cells of the table) between one of the three domains "Academic Employee", "Project", "Object" and a set of seven domains in columns of the table ("Instrument", "PhD Project", "Student", "Education", "Music", "Supervisor" and "Entertainment"). The values in this table represent particularly the cosine similarity metric between two vectors of two domains. The darker shade of colour represents higher degree of matching between the two given domains. The highest degree of matching within these experiment can be seen between the domain "Project" and "PhD Project" (e.g., in the experiment with MaxOnt=400, the similarity matching value equals to 0.8204).

It can be seen, that the increasing number of the ontologies, used for the domain descriptive vector calculating, causes that the values of similarity are more precise and less diffused and the position of some domain from the above given ontology triplet in some column is reinforced. At the same time, the position in the other columns is weakened. The values of parameter MaxOnt, which are higher than 300, do not cause any significant change of the results. Thus, just the value MaxOnt=300 seems to be an adequate compromise between time complexity of the calculation and precision.

6.2 OntoSumViz Implementation Testing

The main goal of the implementation OntoSumViz is to navigate users in the ontology space and to help them to select a suitable ontology for a given application or a given problem. Therefore we performed tests to compare the precision of the user navigation by experts and by the implementation OntoSumViz. Three different experts have determined the ontology key concepts (the first column of the Table 3 – "Genre", "Expression", ...) belonging to the defined domains (the first row of the Table 3 – "Artist", "Entertainment", ...). At first, semantic matching between the key concepts and domains was determined by all experts as a number from the interval [0,5]. Next, the measure of agreement of all experts was quantified by the standard deviation (Bessel modification was used). The results of the test are illustrated in Table 3.

The standard deviation shows differences among the decisions of the particular experts. Value 0 represents absolute agreement of all experts. The experiment acquired also pairs of the key words of the domain with a clear assignment for example "Instrument" – "Instrument". Such pairs can show clear agreement between experts (Table 3). There are some domains without any relation to the ontology key concepts for example "Tree" and partially "Car". They can show that concepts are not assigned in a random way. Another tested example is the

case, when one key concept can be assigned to several domains and not belonging to any one from the given domains clearly (e.g., "Genre"). The experiment proved that the agreement among experts in this case is not very high. Nevertheless, values in the Table 3 are better than we expected.

Table 3
Standard deviation of experts' agreements in the determination of belonging of the given key concepts to the selected domains

	Artist	Entertainment	Musical Group	Rock	Medium	Musical	Instrument	Car	Tree
Genre	2,645751311	2,645751311	2,645751311	2,886751346	1,154700538	1,154700538	2,081665999	0	0
Expression	2,886751346	2,645751311	2,081665999	2,081665999	2,081665999	1,154700538	2,081665999	0	0
Person	0	2,886751346	0,577350269	2,886751346	2,886751346	1,732050808	0	1,732050808	0
Signal	0,577350269	1,527525232	1,732050808	1,527525232	2,645751311	2,309401077	2,309401077	0	0
Corporate Body	1,154700538	1,527525232	1,527525232	0,577350269	2	2,081665999	0,577350269	2,886751346	0
Release Type	2,081665999	1,527525232	2,081665999	2,081665999	1,732050808	2,081665999	0,577350269	0	0
Musical Manifestation	2,081665999	1,732050808	2,081665999	2,309401077	1,527525232	0	1,732050808	0	0
Manifestation	1,154700538	0,577350269	1,732050808	2,309401077	1,732050808	2,309401077	0,577350269	0	0
Musical Expression	1,527525232	1,527525232	1,527525232	2,081665999	1,527525232	0	1,732050808	0	0
Item	0,577350269	0,577350269	0,577350269	0,577350269	2,309401077	2,886751346	1,732050808	0,577350269	0
Musical Work	2,081665999	2,081665999	1,527525232	2,309401077	1,154700538	0	0,577350269	0	0
Work	2,886751346	0,577350269	2,309401077	0,577350269	0,577350269	0,577350269	0,577350269	0,577350269	0
Group	2,645751311	1,732050808	0	1,154700538	1,732050808	0	1,154700538	0	0
Document	2,309401077	1,154700538	0	0	1,732050808	2,516611478	0,577350269	0	0
Medium	2,309401077	0,577350269	1,154700538	0,577350269	0	2,645751311	0,577350269	0	0
Record	1,154700538	2,645751311	0,577350269	0,577350269	0,577350269	0	1,527525232	0	0
Agent	0	1,154700538	2,309401077	0	0	1,732050808	1,154700538	0	0
Show	1	0	1,527525232	1,154700538	0,577350269	0,577350269	1	0	0
Instrument	0,577350269	1,527525232	1	1,154700538	0,577350269	0	0	0	0
Instant	0	0	0	0	0	0,577350269	0	1,732050808	0
TmeLineMap	0	0	0	0	0	0	0	0	0

6.3 Comparison of the Implementation OntoSumViz with Experts

For the purpose of another experiment, an arithmetic average of experts' responses was computed. The obtained average values were transformed into the interval [0,1]. This step cannot be omitted – it is necessary for the results to be comparable with the results obtained from OntoSumViz. Next, decisions of the implementation OntoSumViz were collected. They are also from the interval [0,1]. This interval represents cosine similarity metric, which has two extreme values: 0 (represents absolute dissimilarity) and 1 (represents absolute similarity). The absolute similarity is only theoretical, because the domain descriptive vectors have usually significantly higher cardinality than a descriptive vector of ontology key concepts. Finally, the differences of experts' average values and the OntoSumViz's values were calculated (Table 4). The results of this comparison are numbers from the interval [-1,1].

Table 4
Comparison between responses of the exports and the implementation OntoSumViz

	Musical	Artist	Musical Group	Entertainment	Rock	Instrument	Medium	Car	Tree	Average
Record	1	0,717933333	0,666666667	0,6	0,733333333	0,533333333	0,728533333	0	0	0,553311111
Group	1	0,5975	0,7644	0,6	0,866666667	0,458666667	0,4	0	-0,0004	0,520759259
Show	0,862966667	0,8	0,709533333	0,9906	0,733333333	0,3853	0,065966667	-0,0016	0	0,505122222
Musical Expression	0,9917	0,659566667	0,654866667	0,666666667	0,533333333	0,6	0,266666667	0	0	0,485866667
Instrument	0,9553	0,666666667	0,6659	0,465466667	0,666666667	0,6639	0,066666667	0	0	0,461174074
Genre	0,864366667	0,5988	0,583	0,6	0,472066667	0,533333333	0,133333333	0	-0,0006	0,420477778
Musical Manifestation	1	0,520533333	0,533333333	0,6	0,466666667	0,4	0,262666667	0	0	0,420355556
Expression	0,859766667	0,659366667	0,458866667	0,4	0,466666667	0,333333333	0,466666667	0	0	0,404962963
Work	0,858566667	0,665166667	0,504533333	0,533333333	0,133333333	0,732633333	0,133133333	0,133333333	0	0,410448148
Musical Work	0,99	0,531933333	0,433366667	0,666666667	0,466666667	0,265966667	0,133233333	0	0	0,387537037
Manifestation	0,733333333	0,520033333	0,6	0,466666667	0,466666667	0,333333333	0,1959	0	0	0,368437037
Person	0,1949	0,9177	0,866666667	0,327933333	0,333333333	0	0,333333333	0,2	0	0,352651852
Agent	0,5988	0,9617	0,509633333	0,266666667	0	0,261966667	-0,0034	0	0	0,288374074
Signal	0,531133333	0,066666667	0,194	0,333333333	0,266666667	0,533333333	0,6	0	0	0,28057037
Medium	0,6	0,533333333	0,128733333	0,066666667	0,066666667	0,066666667	0,7313	0	0	0,243707407
Release Type	0,466666667	0,333333333	0,333333333	0,266666667	0,466666667	0,066666667	0,2	0	0	0,237037037
Corporate Body	0,465666667	0,133333333	0,306533333	0,281933333	0,066666667	0,066666667	0,3957	0,331433333	0	0,227548148
Item	0,331133333	0,066666667	0,064466667	0,066666667	0,066666667	0,1976	0,264966667	0,064966667	-0,0024	0,124525926
Document	0,527533333	0,239566667	-0,0063	0,096433333	-0,006	0,062766667	0,1953	-0,0124	-0,0026	0,121588889
Instant	0,066666667	0	0	0	0	0	0	0,2	0	0,02962963
TmeLineMap	0	0	0	0	0	0	0	0	0	0
Average	0,932518519	0,653948148	0,615677778	0,60622963	0,563562963	0,515611111	0,280403704	0,014637037	-0,000111111	

Once again, darker shade of colour represents higher disagreement between the experts and the OntoSumViz. The shadowing illustrates ordering of the key concepts and the domains from more questionable to more clear. The last column and the last row contain "Average" values, representing classification errors. Since there are only a few negative values, it is clear that experts's decisions are systematically higher than the decisions of OntoSumViz (its highest similarity value is 0.3361 only, while the highest similarity assigned by the experts was 1.0).

The most questionable domain is the domain "Musical". The ambiguousness of the domain causes discordance among experts. This fact has an influence on different classifications by experts and by the implementation OntoSumViz. The OntoSumViz takes into account all possible contexts of the word or phrase. On the other side, experts take into account only one context of the word, usually more probable according to their experiences. Very often the OntoSumViz system prefers a domain on the higher level of generality. The vectors of more general domains have usually higher cardinality and so higher chance to match with some key concept descriptive vector.

From the point of classification, 8 concepts from 21 were classified in the same manner (the experts and OntoSumViz agreed in their classification), i.e. the overall classification error was 0.619. The error is influences by the selected domains – the most problematic domain is "Musical". After removing this domain from the test, the number of correctly classified concepts increases to 12.

Conclusions

This work provides a new insight into the ontology evaluation field according to its suitability for solving a given problem. For example, such problem can be a decision in the form of selection of an ontology for a system Magpie [7]. The system needs to load a suitable ontology for a given domain. The semantics of the Magpie (explanation of concepts for user) is based on availability of such ontology. In contrast to the hierarchical tree representation, our approach visualises and interprets concepts with the help of the context, which is represented by their environment (the neighbourhood concepts). Thus, semantics of the concepts is essential within the process of the ontology visualization. Our approach implements the principle of conceptual basis in the form of a map, which helps users to discover those domains, which the given ontology covers. In case of the necessity to visualize more than one ontology, our approach helps users to see the main differences among considered ontologies.

The main contribution of this work is the design and implementation of the method of the vector descriptions of domains, which are generated by the information contained in the related ontologies. This approach offers quick overview of the given ontologies content without long exploration. We suppose, that ontologies are more valuable for this purpose than web pages or text documents, because ontologies contain dictionary of uniformly defined concepts, defined also by their properties and relations. All the facts were taken into account during the descriptive vectors design. Another very important contribution in the visualization field is placing the represented ontology on the map and creating the environment for the ontology. The novelty of the implementation OntoSumViz is also in the combination of the ontology summarization with the conventional tree structure visualization.

We can see some possibilities for further extensions of the designed and implemented approach, for example looking through more ontologies at the same time. User could locate two different ontologies on the map and denote their mutual complementation or combination. It could be useful in the case, when the problem could not be satisfactory solved with the aid of a single ontology. In solving practical problems it is rather rare to find a single ontology, which is able to cover the whole problems and needs. For this reason it is very suitable to aggregate concepts from more sources.

Another extension could be enriching the visualization by the functionalities of the implementation to make it more helpful for those users, who require a more complex ontology view. The combination of visualization approaches could be suitable for the possibility to switch between different visualization views, for example between the contextual and the semantic view. It could be also interesting to investigate new application domains of the descriptive vectors, which have potential overlapping the field of the ontology visualization. One possible domain is the field of recognizing personality aberration from a written text [18], where

the descriptive vector for each aberration will be created from the texts written by persons suffered from this aberration. Consequently it will be compared with the descriptive vector of some new patient.

Acknowledgement

The work presented in this paper was supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/0493/16 (20%) and by the National Research and Development Project Grant 1/0773/16 (30%). This work is also the result of the project implementation Development of the Centre of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).

References

[1] Alatrish, E. S., Tošič, D., Milenkovič, N.: Building Ontologies for Different Natural Languages. In: *Computer Science and Information Systems* 11(2), 2014, 623-644

[2] Brank, D. M. J., Grobelnik, M.: A Survey of Ontology Evaluation Techniques. In *Proc. of the 8th Int. multi-conference Information Society*, 2005

[3] d'Aquin, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Sabou, M., Motta, E.: Watson: A Gateway for Next Generation Semantic Web Applications. Poster session of the *International Semantic Web Conference*, 2007

[4] d'Aquin M., Motta E., Peroni S.: Identifying Key Concepts in an Ontology, through the Integration of Cognitive Principles with Statistical and Topological Measures, Knowledge Media Institute, 2006

[5] Daramola, O., Sindre, G., Moser, T.: A Tool-based Semantic Framework for Security Requirements Specification. *Journal of Universal Computer Science*, Vol. 19, No. 13, 2013

[6] Ding, L., et al.: Swoogle: A Search and Metadata Engine for the Semantic Web, Proceedings of the Thirteenth *ACM Conference on Information and Knowledge Management*, 2004

[7] Domingue, J. B., Dzbor, M., Motta, E.: Collaborative Semantic Web Browsing with Magpie, In Proc. of the 1st *European Semantic Web Symposium* (ESWS), Greece, May 2004

[8] Dzbor, M., Motta, E., Builarabda, C., Gomez-Perez, J. M., Goerlitz, O., Lewen, H.: Analysis of User Needs, Behaviors & Requirements with Interfaces and Navigation of Ontologies. Deliverable report D4.1.1, NeOn Project Consortium, 2006

[9] Eyl, M.: The Harmony Information Landscape: Interactive, Three Dimensional Navigation trough an Information Space. Graz University of Technology, Austria, 1995

[10] Flotyński, J., Walczak, K.: Semantic Representation of Multi-platform 3D Content. In: *Computer Science and Information Systems* 11(4), 2014, 1555-1580

[11] Garcia-Moreno, C., Hernandez-Gonzalez, M. A., Minarro-Gimenez, J. A., Valencia-García, R., Almela, A.: A Semantic-based Platform for Research and Development Projects Management in the ICT Domain. *Journal of Universal Computer Science*, Vol. 19, No. 13, 2013

[12] Gruber, T. R.,: A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), 1993, 199-220

[13] Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., and Giannopoulou, E.: Ontology Visualization Methods - A survey. *ACM Comput. Surv*. 39(4), 2007, 0-4

[14] Lee, M., Pincombe, B, Welsh, M.: An Empirical Evaluation of Models of Text Document Similarity. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum, 2005, 1254-1259

[15] Maedche, A., Staab, S.,: Measuring Similarity Between Ontologies. CIKM, LNAI vol. 2473, 2002

[16] NeOn Project. [Online]. Available: http://www.neon-project.org/nw/Welcome_to_the_NeOn_Project (current August 2015)

[17] Nguyen, V. C., Qafmolla, X., Richta, K.: Domain Specific Language Approach on Model-driven Development of Web Services. *Acta Polytechnica Hungarica* Vol. 11, No. 8, 2014, 121-138, ISSN 1785-8860

[18] Ondrejka, A., Šaloun, P., Ceplakova, R.: Identification of a Personality Aberration from a Written Text. *Proc. of the 10th Workshop on Intelligent and Knowledge-oriented Technologies*, 2015

[19] Parsia, B., Wang, T., Golbeck, J.: Visualizing Web Ontologies with Cropcircles. *End User Semantic Web Interaction WS @ ISWC2005*, 2005

[20] Sintek, M.: Ontoviz tab: Visualizing Protégé Ontologies. [Online]. Available: http://protege.stanford.edu /plugins/ontoviz/ontoviz.html (current August 2015)

[21] Souza, K., X. S., Dos Santos, A. D., Evahgeista, S. R. M.: Visualization of Ontologies through Hypertrees. In Proceedings of the *Latin American Conference on Human-Computer Interaction*, Rio de Janeiro, Brazil, 2003, 251-255

[22] Sram, N., Takács, M.: An Ontology Model-based Minnesota Code. *Acta Polytechnica Hungarica* Vol. 12, No. 4, 2015, 97-112, ISSN 1785-8860

[23] Superkar, K.: A Peer-review Approach for Ontology Evaluation. *Proc. 8th Intl. Protégé Conference*, Madrid, Spain (2005)

[24] Van Ham, F., Vanwijk, J. J,: Beamtrees: Compact Visualization of Large Hierarchies. In *Proceedings of the IEEE Conference on Information Visualization*. IEEE CS Press, 2002, 93-100 (2002)

[25] Xu, B., deFréin, R., Robson, E., Ó Foghlú, M.: Distributed Formal Concept Analysis Algorithms Based on an Iterative MapReduce Framework. [Online]. Available: http://link.springer.com/chapter/10.1007%2F978-3-642-29892-9_26#page-2 (current March 2016)

[26] Zhu, T., Hu, B., Yan, J., Li, X.: Semi-supervised Learning for Personalized Web Recommender System. In: *Computing and Informatics*, 29(4), 2010: 617-627 (2010)

An Alternative Method in Optimizing Random Outcomes

Sándor Molnár[1], Ferenc Szidarovszky[2]

[1] Institute of Mathematics and Informatics, Szent István University, Páter K. u. 1, H-2100, Gödöllő, Hungary, molnar.sandor@gek.szie.hu

[2] Ferenc Szidarovszky, Department of Applied Mathematics, University of Pécs, Ifjúság u. 6, H-7624, Pécs, Hungary

Abstract: Within economic literature random outcomes can be characterized by their certainty equivalents. In this article, a general approach for their extension is first outlined and then special cases are shown. The two most simple of these cases result in the classical formula of certainty equivalent, and by increasing the degree of the approximating Taylor polynomials, more advanced formulas are derived. Additionally, a simple advanced formula is compared favorably to the classical approach in a computer study and some application models are discussed to illustrate the methodology.

Keywords: decision making; uncertainty; applications

1 Introduction

In practical decision making problems we often face random elements due to modeling, natural and economic factors. In constructing mathematical models certain elements are neglected in order to keep the model solvable. The natural and economic components are usually uncertain due to the lack of relevant data and prediction errors. Uncertainty in mathematical models is usually formulated with fuzzy or stochastic methodology, where the uncertain quantities are considered fuzzy numbers with appropriate membership functions or as random variables with certain probability density functions which are only estimated so there is no way to construct theoretically correct function forms. The fuzzy methodology constructs a fuzzy number as the solution, which is then defuzzified, for which several alternative methods are available [6, 1]. If stochastic methodology is chosen, then stochastic programming [5, 11] is a very popular approach. In order to decrease uncertainty, the variances of the objective functions are minimized in addition to optimizing the expected values of the objectives leading to multi-objective optimization problems [13, 10]. Data analytical methods also can be used to reduce variances [7]. Bayesian methodology [3] is

based on the repeated updating of the probability distributions using new sample elements. In the economic literature very often certainty equivalents are introduced and optimized instead of random objectives [12]. They are linear combinations of expectations and variances, which is the same as applying the weighting method.

There are many applications of the stochastic methodology including extractibility of natural resources [2], groundwater management [4], emission allowance prices [9], reliability engineering [8]. The many application fields show the importance of this methodology.

In this paper an alternative approach is introduced, which can replace the certainty equivalent and provides more accurate solutions. The authors of this paper could not find any earlier work deriving more advanced solutions and relating them to the root locus method. After the theoretical issues are discussed, a comparison study is reported and some particular models are described to illustrate the methodology. The last section is devoted to conclusions and future research directions.

2 The Mathematical Methodology

Consider a random variable x representing the value of an outcome. The goodness of the different values of x is characterized by a utility function $u(x)$. Introduce the notation $\bar{x} = E(x)$ and $\sigma^2 = \mathrm{Var}(x)$. Clearly the random outcome can be replaced by a deterministic value x^*, such that

$$u(x^*) = E\big(u(x)\big) = \int_{-\infty}^{\infty} u(x) f(x) dx \tag{2.1}$$

where $f(x)$ is the probability density function of x. If $u(x)$ is strictly monotonic, then

$$x^* = u^{-1}\left(\int_{-\infty}^{\infty} u(x) f(x) dx \right). \tag{2.2}$$

This formula cannot be applied in most cases, since $f(x)$ is usually unknown. We can however derive an acceptable estimate of x^* as follows. By the Taylor's formula

$$u(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \sum_{i=2}^{m} \frac{1}{i!} u^{(i)}(\bar{x})(x - \bar{x})^i + R_{m+1}(x) \tag{2.3}$$

and

$$u(x^*) = u(\bar{x}) + u'(\bar{x})(x^* - \bar{x}) + \sum_{j=2}^{n} \frac{1}{j!} u^{(j)}(\bar{x})(x^* - \bar{x})^j + R_{n+1}(x) \tag{2.4}$$

when $R_{m+1}(x)$ and $R_{n+1}(x)$ are the remainder terms. By omitting the error terms and taking expectation,

$$E\big(u(x)\big) = u(\bar{x}) + \sum_{i=2}^{m} \frac{1}{i!} u^{(i)}(\bar{x}) M_i, \tag{2.5}$$

so (2.1) implies

$$\sum_{j=1}^{n} \frac{1}{j!} u^{(j)}(\bar{x}) \Delta^j = \sum_{i=2}^{m} \frac{1}{i!} u^{(i)}(\bar{x}) M_i \tag{2.6}$$

where $M_i = E\left[(x - \bar{x})^i\right]$ is the i^{th} central moment of x and $\Delta = x^* - \bar{x}$. Notice that (2.6) gives an n^{th} degree polynomial equation for unknown Δ, from which $x^* = \Delta + \bar{x}$. In order to find the right root of (2.6) consider the root loci of equation

$$\sum_{j=1}^{n} \frac{1}{j!} u^{(j)}(\bar{x}) \Delta^j = K, \tag{2.7}$$

where *K* is the parameter. Each locus shows how the associated root of this equation varies as the value of *K* changes. In the deterministic case $x = \bar{x}$ with $\sigma^2 = 0,$ so $K = 0$ and in this case $x^* = \bar{x}$ implying that $\Delta = 0.$ Therefore, we have to select the locus which passes through the origin. The value of this locus at

$$K = \sum_{i=2}^{m} \frac{1}{i!} u^{(i)}(\bar{x}) M_i \tag{2.8}$$

gives the value of Δ.

Some special cases are presented next.

Example 2.1 *Assume* $m = n = 1$, *then we have*

$$u'(\bar{x})(x^* - \bar{x}) = 0 \tag{2.9}$$

and if $u'(\bar{x}) \neq 0$, *then* $x^* = \bar{x}$.

Example 2.2 *Let* $m = 2$ *and* $n = 1$, *then equation* (2.6) *becomes*

$$u'(\bar{x})\Delta = \frac{1}{2}u''(\bar{x})\sigma^2 \tag{2.10}$$

implying that

$$\Delta = \frac{u''(\bar{x})}{2u'(\bar{x})}\sigma^2 \tag{2.11}$$

so

$$x_1^* = \bar{x} + \alpha\sigma^2 \tag{2.12}$$

is the approximation of x^* *with*

$$\alpha = \frac{u''(\bar{x})}{2u'(\bar{x})}. \tag{2.13}$$

It is the well known certainty equivalent.

Notice that this method cannot be used if $u'(\bar{x}) = 0.$

Example 2.3 *Let now* $m = n = 2$*, then*

$$u'(\bar{x})\Delta + \frac{1}{2}u''(\bar{x})\Delta^2 = \frac{1}{2}u''(\bar{x})\sigma^2 \tag{2.14}$$

or

$$\alpha\Delta^2 + \Delta - \alpha\sigma^2 = 0. \tag{2.15}$$

If $\sigma^2 = 0$*, then there is no uncertainty, so* $\Delta = 0$ *is the solution. In general,*

$$x^* = \bar{x} + \Delta = \bar{x} + \frac{-1 \pm \sqrt{1 + 4\alpha^2\sigma^2}}{2\alpha} \tag{2.16}$$

and since at $\sigma^2 = 0$ *the solution has to be* $\bar{x}$*, the positive square root has to be considered:*

$$x_2^* = \bar{x} + \frac{-1 + \sqrt{1 + 4\alpha^2\sigma^2}}{2\alpha} \tag{2.17}$$

is the approximation of x^* *by the more accurate method.*

Similarly to the previous example this formula cannot be used if $u'(\bar{x}) = 0.$ *If* $\alpha = 0,$ *then* $\Delta = 0,$ *so* $x_2^* = \bar{x}.$

3 Comparison Study

In order to compare the accuracy of formulas (2.12) and (2.17) we conducted a simulation study. Random variable x was considered with four different density functions on $[-1,1]$ as follows:

$$f_1(x)=\frac{1}{2},\quad f_2(x)=\frac{1}{2}(x+1),\quad f_3(x)=\frac{1}{2}(1-x)\ \text{ and }\ f_4(x)=\frac{3}{4}(1-x^2)$$

where $f_1(x)$ is constant, $f_2(x)$ is increasing, $f_3(x)$ is decreasing and $f_4(x)$ is mound-shaped. Four different utility functions were chosen,

$$u_1(x)=\frac{1}{4}(x+1)^2,\ u_2(x)=1-\frac{1}{4}(x-1)^2,$$
$$u_3(x)=\frac{1}{2}(1+x^3),\ u_4(x)=\frac{1}{2}+\frac{2}{\pi}\tan^{-1}(x),$$

where $u_1(x)$ is convex, $u_2(x)$ is concave, $u_3(x)$ is convex for $x>0$ and concave for $x<0$, and $u_4(x)$ is convex for $x<0$ and concave for $x>0$. So a large variety of density and utility function types were considered, and $16=4\times 4$ cases examined. In each case the true value of x^* was determined based on equation (2.2), since both $u(x)$ and $f(x)$ were known for all cases. Table 1 shows the results. The first and second columns specify the density and utility functions, the third column shows the true value of x^*. The fourth column gives the results based on (2.12) where $\bar{x}$ and σ^2 are computed based on the given density functions. The sixth column contains the results obtained by using (2.17). The fifth and seventh columns show the errors of the obtained estimates. Among the 16 cases we can find 10, where (2.17) gives the exact answer, in 4 cases (2.17) has smaller error, and in 2 cases the formulas could not be used.

If the utility function is linear or quadratic, then with $n=m=2$, $R_{m+1}(x)=R_{n+1}(x)=0$ in equations (2.3) and (2.4), so formula (2.17) is exact. In the cases of densities f_1 and f_4, $\mathrm{E}(x)=\bar{x}=0$, furthermore $u_3'(x)=\frac{3}{2}x^2$ which is zero at $\bar{x}=0$, so formulas (2.12) and (2.17) cannot be applied. In addition $u_4''(x)=\dfrac{-4x}{\pi\left(1+x^2\right)^2}$ with zero value at $\bar{x}=0$, therefore in the cases of densities f_1 and f_4, $\alpha=0$ implying that $x_1^*=x_2^*=\bar{x}=0$ from both formulas (2.12) and (2.17).

Table 1
Simulation Results

$f(x)$	$u(x)$	x^*	x_1^*	$x^* - x_1^*$	x_2^*	$x^* - x_2^*$
$f_1 = \frac{1}{2}$	$u_1(x) = \frac{1}{4}(x+1)^2$	0.1547	0.1667	-0.0120	0.1547	0
	$u_2(x) = 1 - \frac{1}{4}(x-1)^2$	-0.1547	-0.1667	0.0120	-0.1547	0
	$u_3(x) = \frac{1}{2}(1+x^3)$	0	N/A	N/A	N/A	N/A
	$u_4(x) = \frac{1}{2} + \frac{2}{\pi} arctan(x)$	0	0	0	0	0
$f_2 = \frac{1}{2}(x+1)$	$u_1(x) = \frac{1}{4}(x+1)^2$	0.4142	0.4146	-0.0004	0.4142	0
	$u_2(x) = 1 - \frac{1}{4}(x-1)^2$	0.1835	0.1667	0.0168	0.1835	0
	$u_3(x) = \frac{1}{2}(1+x^3)$	0.5848	1.000	-0.4152	0.6667	-0.0819
	$u_4(x) = \frac{1}{2} + \frac{2}{\pi} arctan(x)$	0.2943	0.2667	0.0267	0.2679	0.0255
$f_3 = \frac{1}{2}(1-x)$	$u_1(x) = \frac{1}{4}(x+1)^2$	-0.1835	-0.1667	-0.0168	-0.1835	0
	$u_2(x) = 1 - \frac{1}{4}(x-1)^2$	-0.4142	-0.4167	0.0025	-0.4142	0
	$u_3(x) = \frac{1}{2}(1+x^3)$	-0.5848	-1.000	0.4152	-0.6667	0.0819
	$u_4(x) = \frac{1}{2} + \frac{2}{\pi} arctan(x)$	-0.2943	-0.2667	-0.0267	-0.2679	-0.0255
$f_4 = \frac{3}{4}(1-x^2)$	$u_1(x) = \frac{1}{4}(x+1)^2$	0.0954	0.1000	-0.9046	0.0954	0
	$u_2(x) = 1 - \frac{1}{4}(x-1)^2$	-0.0954	0.1000	0.0046	-0.0954	0
	$u_3(x) = \frac{1}{2}(1+x^3)$	0	N/A	N/A	N/A	N/A
	$u_4(x) = \frac{1}{2} + \frac{2}{\pi} arctan(x)$	0	0	0	0	0

4 Applications

In this section some application models are introduced.

Model 4.1 (Budget allocation) *An investment firm with budget* B *has* n *investment opportunities, where opportunity* k *gives profit* π_k *per each invested dollar with* $E(\pi_k) = \mu_k$ *and* $\mathrm{Var}(\pi_k) = \sigma_k^2$. *If the profits are independent, then*

the expectation and variance of the profit $\pi = \sum_{k=1}^{n} \pi_k x_k$ *from the total investment* $x = \sum_{k=1}^{n} x_k$ *are given as*

$$\bar{\pi} = E(\pi) = \sum_{k=1}^{n} E\left(\pi_k x_k\right) = \sum_{k=1}^{n} \mu_k x_k \tag{4.1}$$

and

$$\sigma^2 = \mathrm{Var}(\pi) = \sum_{k=1}^{n} \mathrm{Var}\left(\pi_k x_k\right) = \sum_{k=1}^{n} \sigma_k^2 x_k^2, \tag{4.2}$$

where x_k *gives the allocated investment in opportunity k.*

By assuming the utility function $u(\pi) = \pi^2$ *we have* $u'(\bar{\pi}) = 2\bar{\pi}$ *and* $u''(\bar{\pi}) = 2$ *showing that* $\alpha = \frac{1}{2\bar{\pi}}$, *the objective function becomes*

$$\begin{aligned} \bar{\pi} + \frac{-1 + \sqrt{1 + 4\alpha^2\sigma^2}}{2\alpha} &= \bar{\pi} + \frac{-1 + \sqrt{1 + 4 \cdot \frac{1}{4\bar{\pi}^2} \cdot \sigma^2}}{\frac{1}{\bar{\pi}}} = \\ &= \bar{\pi} - \bar{\pi} + \sqrt{\bar{\pi}^2 + \sigma^2} = \sqrt{\bar{\pi}^2 + \sigma^2}, \end{aligned} \tag{4.3}$$

so the firm solves the quadratic programming problem maximizing

$$\left(\sum_{k=1}^{n} \mu_k x_k\right)^2 + \sum_{k=1}^{n} \sigma_k^2 x_k^2$$

subject to

$$\begin{aligned} &x_k \geq 0 \quad (k = 1, 2, \ldots, n) \\ &\sum_{k=1}^{n} x_k = B \end{aligned} \tag{4.4}$$

Model 4.2 (Oligopoly # 1) *Consider an* n *-firm single product oligopoly without product differentiation. Let* x_k *be the output of firm* k, $c_k(x_k)$ *its cost function. The industry output is* $x = \sum_{k=1}^{n} x_k$, *and the corresponding unit price function is*

$p(x)$. *However the firms do not know the exact price function, so firm* k *believes that the price function is* $p_k(x)+\varepsilon_k$, *where* $p_k(x)$ *is the believed price function by firm k (usually different than the true price function) with a random error term* ε_k *resulting from market uncertainties. So firm* k *believes that its profit is*

$$\pi_k = x_k\left(p_k(x)+\varepsilon_k\right)-c_k(x_k), \tag{4.5}$$

which is considered as the random outcome for firm k.

By assuming that $E(\varepsilon_k)=0$, $\mathrm{Var}(\varepsilon_k)=s_k^2$, *we have*

$$\overline{\pi}_k = E\left(\pi_k\right)= x_k p_k(x)-c_k(x_k) \tag{4.6}$$

and

$$\sigma_k^2 = \mathrm{Var}(\pi_k)= s_k^2 x_k^2. \tag{4.7}$$

If the firms select exponential utility functions, $u_k\left(\pi_k\right)=e^{\pi_k\beta_k}$, *then* $\alpha_k=\dfrac{\beta_k}{2}$ *is constant for each firm, so the objective functions become*

$$\overline{\pi}_k+\frac{-1+\sqrt{1+4\alpha_k^2 s_k^2 x_k^2}}{2\alpha_k} = x_k p_k(x)-c_k(x_k)+\frac{-1+\sqrt{1+4\alpha_k^2 s_k^2 x_k^2}}{2\alpha_k}. \tag{4.8}$$

Notice that this is the profit function (4.5) *of oligopolies without uncertainty and product differentiation where the modified cost functions are*

$$c_k(x_k)-\frac{-1+\sqrt{1+4\alpha_k^2 s_k^2 x_k^2}}{2\alpha_k}. \tag{4.9}$$

Model 4.3 (**Oligopoly # 2**) *Consider now an oligopoly when the firms know the true price function* $p(x)$ *but their costs are uncertain. Assume that because of uncertain prices of labor, energy and material firm* k *believes that its cost function is* $c_k(x_k)=\gamma_k(x_k)+\varepsilon_k x_k$, *where* $\gamma_k(x_k)$ *is a known function and* ε_k *is a random variable with* $E(\varepsilon_k)=0$ *and* $\mathrm{Var}(\varepsilon_k)=s_k^2$. *Notice that* ε_k *is a random term in the marginal cost. The profit of firm k is its random outcome. The expectation and variance of the profit of firm* k *is*

$$\overline{\pi}_k = E(\pi_k)= x_k p\left(\sum_{l=1}^{n} x_l\right)-\gamma_k(x_k) \tag{4.10}$$

and

$$\sigma_k^2 = \mathrm{Var}(\pi_k) = s_k^2 x_k^2. \tag{4.11}$$

where $p\left(\sum_{l=1}^{n} x_l\right)$ *is the price function which is considered to be a public information for the firms. Similary to the previous case* (2.17) *gives the modified objective function of firm* k :

$$x_k p\left(\sum_{l=1}^{n} x_l\right) - \gamma_k(x_k) + \frac{-1+\sqrt{1+4\alpha_k^2 s_k^2 x_k^2}}{2\alpha_k}. \tag{4.12}$$

If the utility functions $u_k(\pi_k)$ *are exponential, then* α_k *is a constant for all firms.*

Conclusion

A well-known concept of the certainty equivalent is replaced by a general approach, which can be reduced to the certainty equivalent in a very special case. A simulation study showed the advantage of the new approach resulting in more accurate approximations.

The methodology was illustrated on three simple models. The more accurate formulas are based on higher order Taylor polynomials of the utility function. Methods (2.12) and (2.17) are based on the adjustment constant α which depends on the first two derivatives of the utility function. Its special form implies two important facts. If $u'(\bar{x}) = 0,$ then α cannot be determined, so these methods cannot be used as was shown in two cases of Table 1. If $u''(\bar{x}) = 0,$ then $\alpha = 0$ implying that $x^* = \bar{x},$ which was also shown in two cases of the utility function $u_4(x)$. In our comparison study we found no case when the classical certainty equivalent was better than our improved formula. We expect that by selecting higher order Taylor polynomial approximations of the utility function the accuracy of the resulting formulas can be improved even further.

Higher order formulas can be used in cases when lower order formulas cannot be used.

In our future research higher order approximations (with larger values of m and n) will be used in particular applications which will be selected from broad fields of engineering and economics.

References

[1] Bellman, R. and Zadeh, L. A. (1970): Decision Making in a Fuzzy Environment, Management Science 17(4), pp. 141-164

[2] Csábrági, A., Molnár, M. (2011): Role of Non-Conventional Energy Sources in Supplying Future Energy Needs, Bulletin of the Szent István University, Gödöllő

[3] DeGroot, M. H. (1970): Optimal Stochastic Decisions. New York: McGraw-Hill

[4] Hatvani, I. G., Magyar, N., Zessner M., Kovács, J., Blaschke, A. P. (2014): The Water Framework Directive: Can More Information Be Extracted from Groundwater Data? A Case Study of Seewinkel, Burgenland, Eastern Austria, Hydrogeology Journal 22(4), pp. 779-794

[5] Kall, P. and Wallace, S. W. (1994): Stochastic Programming. Chichester: Wiley

[6] Klir, G. J. and Yuan, B. (1995): Fuzzy Set and Fuzzy Logic: Theory and Applications. Upper Saddle River: Prentice Hall

[7] Kovács, J., Kovács, S., Magyar N., Tanos, P., Hatvani, I. G., Anda, A. (2014): Classification into Homogeneous Groups Using Combined Cluster and Discriminant Analysis, Environmental Modelling & Software 57, pp. 52-59

[8] Matsumoto, A., Szidarovszky, F. and Szidarovszky, M. (2014): Incorporating Risk Economic an Optimization Model of Reliability Engineering, International Journal of in Behavior and Organization 3(1-2), pp. 1-4

[9] Molnár, M. (2014): Opportunities for Hungary under the Stability Reserve of the EU ETS, Journal of Central European Green Innovation 2(2), pp. 105-114

[10] Molnár, S., Szidarovszky, F. (2011): Game Theory, Multiobjective Optimization, Conflict Resolution, Differential Games (in Hungarian) Budapest: Computerbooks

[11] Prekopa, A. (1995): Stochastic Programming. Dordrecht: Kluwer Academic Publishers

[12] Sargent, T. J. (1979): Macroeconomic Theory. New York: Academic Press

[13] Szidarovszky, F., Gershon, M. and Duckstein, L. (1986): Techniques for Multi-objective Decision Making in Systems Management. Amsterdam: Elsevier

[14] Szidarovszky, F. and Yakowitz, S. (1978): Principles and Procedures of Numerical Analysis. New York: Plenum Press

An Experimental and Numerical Investigation of the Mechanical Properties of Spinal Cords

Monika Ratajczak, Marek Malinowski, Romuald Będziński

University of Zielona Góra, Faculty of Mechanical Engineering,
Prof. Z. Szafrana 4, Zielona Gora, 65-516, Poland
m.ratajczak@iizp.uz.zgora.pl, m.malinowski@ibem.uz.zgora.pl,
r.bedzinski@ibem.uz.zgora.pl

Abstract: Studies concerning the mechanical properties of the spinal cord are crucial for the understanding of various related pathologies. The present study introduces the results of an analysis focused around the mechanical properties of the two types of the mammalian spinal cord: domestic pig (Sus scrofa f. domestica) and domestic rabbit (Oryctolagus cuniculus f. domesticus). The research has been conducted in an in vitro environment, and the freshly dissected cords have been subjected to uniaxial tension testing. A series of preliminary tests allowed for the selection of the optimum method for fixing the cord in the chuck of the testing machine. All preparations were tested for 3 hours after the death of the animal, while an appropriate level of hydration, and the temperature and the strain rate of 0.08s-1 have been maintained. The nonlinear response of the tested tissues has been obtained under the force-displacement conditions. On the basis of the experimental studies the mechanical properties of the samples have been described. Additionally, numerical calculations have been performed on a simplified model of the spinal cord with the use of the finite element (FE) method, which were finally compared with the actual behavior of the sample tissues. An analytical approach with the hyperelastic Ogden material model and FE model of the spinal cord were employed to derive mechanical properties of the tested spinal cords. The results demonstrate that a non-linear FE model is able to predict the mechanical behavior of the spinal cords in the uniaxial tension.

Keywords: spinal cord injury (SCI); nonlinear mechanical properties; finite element method (FEM); mechanical testing

1 Introduction

As a result of traffic accidents, falls from heights, as well as, various spinal diseases, spinal cord injuries (SCI) are reported with greater frequency. To understand the mechanisms of SCI and spinal cord compression syndromes, testing the mechanical properties of the spinal cord is necessary to facilitate the development of alternative tissue models. The early diagnosis of the type and

extent of damage is essential for the development of proper methods of treatment [1]. The forces acting in the course of injury cause high stress and strain in the spinal cord tissue, and results in primary damage and a breach of the blood-spinal cord [2, 3]. The resulting stress and deformation of the spinal cord is a major cause of neurological deficit and loss of motor and sensory function in patients with traumatic SCI [4]. The size of deformation, strain rate, size of axons and the local stress state in tissue have been proposed as the primary mechanism of damage to the spinal cord parenchyma cells during an injury [5]. Morbidities, such as degeneration, tumors, or cancers, can also cause the compression of the spinal tissue leading to its destruction. The intensity and duration of stress determine the size and potential reversibility of the spinal cord's dysfunction [6]. The mechanical tests are used to determine the mechanical properties of the spinal cord, allow for the understanding of the differences between healthy and pathological tissues, as well as the understanding of the mechanisms of an injury. The knowledge of the mechanical properties of the tissue can be used in robot-surgeon control systems, where the understanding of the deformation of the tissue is simply mandatory [7, 8, 9] and also allows one to specify the boundary conditions for the numerical analysis necessary to optimize the treatment process [4, 10, 11]. When it comes to any kind of reconstruction, the experimental results are a very important starting material for a given therapy [12, 13, 14]. The determination of the mechanical properties, as well as the understanding of the deformation response of the spinal cord is the basis for the construction of a regenerative bridge with a module fostering the nerve fiber regeneration [15]. The use of different strain rates during mechanical testing reflects the actual changes of the tissues in the aforementioned cases. The behavior of the spinal cord during a very slow strain rate can be interpreted in regard with various morbidities. The understanding of the mechanisms operating in traumatic injuries requires the use of a high-speed strain rate. On the other hand, the development of automatic surgical tools and robots [16, 17, 18, 19, 20] and virtual reality techniques [21, 22] focuses on research in the moderate-speed deformation, which is important for surgical procedures [7].

Due to the limited availability of human specimens, testing techniques are generally limited to animal autopsy samples [23, 24, 25, 26, 27, 28, 29, 30]. A pioneer in conducting research on the mechanical properties of the spinal cord was Tunturi in 1978 [23]. Three years later, Hung and colleagues examined the spinal cord of cats and puppies in vivo [24, 25, 26]. Later, the mechanical properties of the spinal cord have been studied in vitro in different species, including: humans [31], adult and neonatal rats [27, 28], and cows [32]. In 2013 Luna et al. conducted in vivo and in vitro experiments on the spinal cords of lampreys [15]. The latest data come from 2014, where the mechanical properties of denticulate ligaments in pigs have been studied at different sections of the cervical region [33]. The results of all tests have been summarized in Table 5. According to Kiwerski et al., the heaviest neurological sequelae occur in the thoracic spine injuries; patients admitted with a paralysis are the largest group

here, amounting to 71%, and 17% are without neurological disorders [1]. The aim of the study was to investigate the mechanical properties of the spinal cords of domestic pigs *(Sus scrofa f. domestica)* and domestic rabbits *(Oryctolagus cuniculus f. domesticus)*, the thoracic segments in the state of uniaxial tension, to be more precise. In the study, a moderate strain rate of $0.08s^{-1}$ has been maintained. The spinal cords characteristics in the force elongation scenario have been obtained. An analysis of the behavior of individual tissue structures under the influence of a tensile force has been conducted, up to the moment of complete rupture of the sample. On the basis of own research and appropriate literature [4, 32, 34, 35, 36, 37], the finite element analysis of the tensile test of the spinal cord has been carried out in the ANSYS 5.7 software. A computer simulation has been used to explain certain behavior of the spinal cord structures in experimental studies employed on the sectional preparations. Selection of animal species was dictated by the anatomical similarity of the porcine spinal cord to the human one, as well as the DNA compliance of 94% [38, 39, 40]. In the case of rabbits, a relatively small number of studies have so far been performed on their cords [32, 36, 37] compared to other species.

In the paper the experimental measurement of the mechanical properties of two types of the mammalian spinal cord were reported and numerical approach was presented and finite element modeling of mechanical parameters were derive. Uniaxial tension was conducted on spinal cords samples in testing machine ZWICK. The FE model of the spinal cord including only the gray and white matters (the first model) and the in the second model the dura mater and the pia mater were added. The hyperelastic constants (parameters) were calculated and numerical solutions were performed. The comparison of measurements and FE modeling results reveals the relationship between the mechanical properties and the studied structure of selected spinal cords.

2 Materials and Methods

Preparations were obtained right after slaughter. Intact spinal cords together with denticulate ligaments and dural sacs were dissected from healthy animals. The tissues were dissected from pigs (Figure 1), at the age of about one year and an average weight of around 100 kg, and rabbits (Figure 2), with an average weight of about 3.5 kg at the age of about 6 months. All preparations were fresh, as the dissection was performed immediately after the death of the animal. The spinal column was purified from the *musculus erector trunci* and laminectomy was performed.

Figure 1
Dissected spinal cord of the domestic pig

During the process of preparation, a regularly exposed portion of the tissue was constantly moisturized with a saline solution to prevent drying of the sample. The maximum time of dissection in specimens was about 150 minutes. After the extirpation surgery, in order to verify the continuity of the spinal cords, all of them were tested under the stereomicroscope. The damaged samples were ultimately discarded.

Figure 2
Dissected spinal cord of the domestic rabbit with a body weight of 4kg

The spinal cords were transferred to 0.9% sodium chloride environment at 37^{o}C, in order to achieve osmotic equilibrium and prevent the drying of samples. The distal ends of the spinal cord were dried with tissue paper. The samples were placed in a special attachment between two multi-grooved rubber pads in the jaws of the Zwick Z050 testing machine. To prevent slippage of the tissue during the measurement, two different methods of mounting the sample were employed: distal sections of the domestic pig spinal cord were wrapped around the polyethylene rollers (Figure 3), whereas the ends of the domestic rabbit's spinal cord tissue were glued with cyanoacrylate adhesive (Figure 4). We were very careful not to violate the samples during installation. During the measurements all samples were moistened with the saline solution.

For the present study quasi-static uniaxial tensile tests were performed (n_1=10 for porcine spinal cords and n_2=8 for domestic rabbit's spinal cords). The spinal cords were elongated, without preconditioning, at a quasi-static rate of 0.05 mm/s.

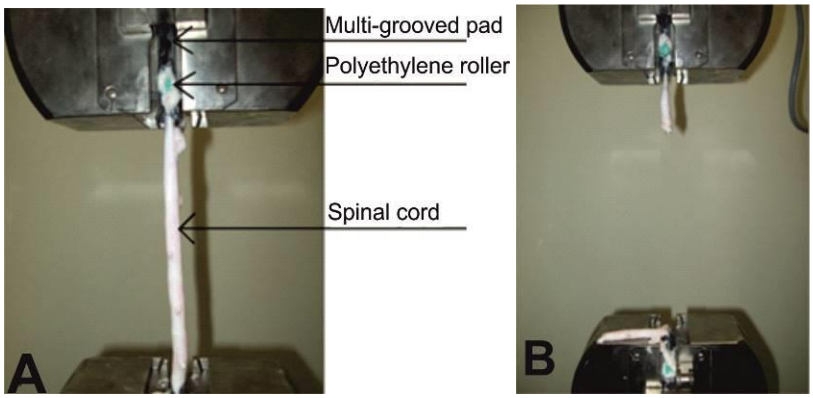


Figure 3
Preparation of the porcine spinal cord. A - at the time of the tensile test. B - rupture and the end of measurement

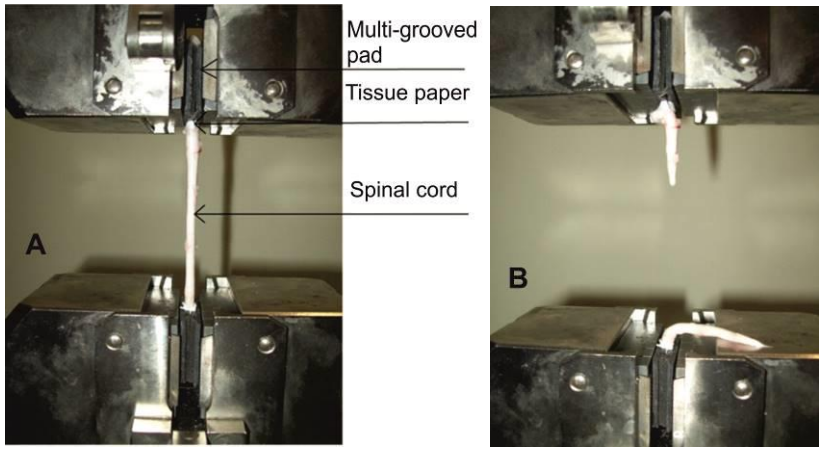
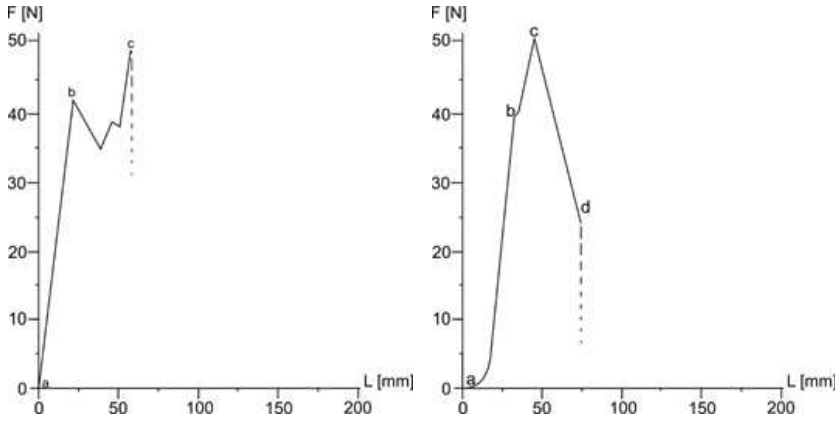


Figure 4
The preparation of the spinal cord of the domestic rabbit. A - at the time of the tensile test. B – rupture of the sample and the end of the measurement

3 Experimental Results

Ten samples of the domestic pig spinal cord were tested in the experiment. During the uniaxial tensile test the average pig spinal cords were deformed by 61.1%. Examples of graphs in the force-displacement scenario have been shown in Fig. 5. On the A-B-C stretch, there is a clear increase in forces with relatively small displacement. In point C, sample A5 and A7 sample in point D the tearing of the

dura mater was observed. In the case of point D, sample, A5 and A7 sample in point E the complete rupture of the tissue took place.



Sample A5 Sample A7

Figure 5

The characteristics of the force-displacement of the spinal cord [N-mm]

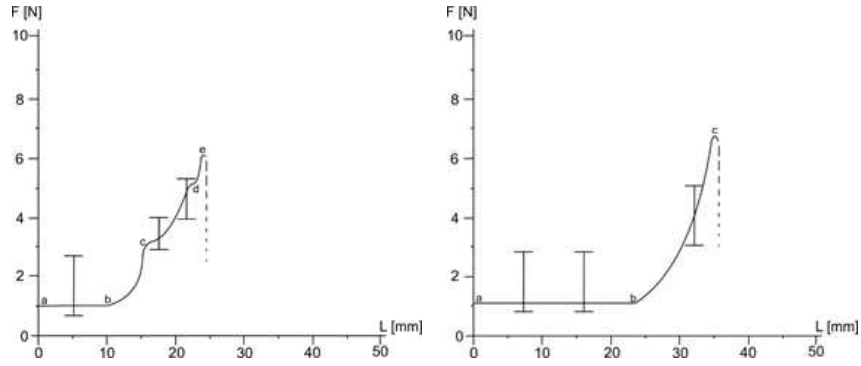
The results of all preparations tested are shown in Table 1. The maximum average force acting on the pig's spinal cord is 47 N. Based on the mean from 10 measurements a value of Young's modulus of 0.323 MPa was obtained.

Table 1

Summary of the results of the porcine spinal cord tests

No. of sample	F_{max} [N]	l [mm]	Δl_{max} [mm]	ε_{max} [%]	Young's modulus [MPa]
A1	35.5	60	32	53	0.273
A2	41.0	224	148	66	0.254
A3	49.2	60	40	67	0.311
A4	55.8	203	130	64	0.356
A5	51.8	97	62	64	0.330
A6	47.0	128	88	69	0.278
A7	50.4	126	72	57	0.361
A8	50.8	66	40	61	0.340
A9	32.4	66	38	58	0.288
A10	56.1	23	12	52	0.440

In a study of the mechanical properties of the rabbit spinal cord, we have used eight preparations. The uniaxial tension of the samples increased in length at 68% in average. The mean maximum force transmission through the spinal cord of the domestic rabbit was 6.75 N. The mean of eight measurements was used to determine Young's modulus of 0.106 MPa.



Sample B2 Sample B3

Figure 6
Characteristics of the force-displacement [N-mm] for the spinal cord of the domestic rabbit

Examples of graphs from the study are shown in Figure 6. Points A - B - pronounced displacement with a small change in force; point D, sample B2 - interruption of the dura mater, point E, sample B2 and point C, sample B3 - the complete rupture of the tissue.

Table 2
Summary of the experimental results from the scrutiny of the domestic rabbit's spinal cord

No. of sample	F_{max} [N]	l [mm]	Δl_{max} [mm]	ε_{max} [%]	Young's modulus [MPa]
B1	7.7	10	6	60	0.136
B2	6.7	38	24	63	0.113
B3	7.1	55	38	69	0.109
B4	5.2	40	27	68	0.081
B5	5.9	25	18	72	0.087
B6	6.2	75	50	67	0.098
B7	9.6	50	38	76	0.134
B8	5.6	58	40	69	0.086

4 The Boundary Conditions of Numerical Investigations and the Results of FEM Analysis

The spinal cord is an inhomogeneous, composite material with a complex construction. Therefore, it is understandable that the biomechanics of the spinal cord injury can prove to be complex [41]. The use of numerical modeling of biological systems provides a more accurate analysis of the biomechanics of tissues [4, 9, 10, 11, 42, 43, 44, 45, 46, 47, 48]. Soft tissues are susceptible to

deformation. The behavior of the materials at large scale deformation can be described using Ogden's hyperelastic model [31, 49, 50, 51]. The Ogden strain energy potential W is defined as:

$$W = \sum_{i=1}^{N} \frac{\mu_i}{\alpha_i^2}\left(\bar{\lambda}_1^{\alpha_i} + \bar{\lambda}_2^{\alpha_i} + \bar{\lambda}_3^{\alpha_i} - 3\right) + \sum_{k=1}^{N} \frac{1}{d_k}(J-1)^{2k} \tag{1}$$

where: $\bar{\lambda}_p (p = 1,2,3)$, are the deviatoric principal stretches, defined as $\bar{\lambda}_p = J^{-1/3}\lambda_p$, λ_p - principal stretches of the left Cauchy-Green tensor $\lambda_p = l_p / l_{p0}$ the ratio of deformed length l_p to the original length l_{p0} in principal directions, respectively $p = 1,2,3$, J - determinant of the elastic deformation gradient, N, μ_p, α_p, d_p - material constants. The initial shear modulus μ is defined by $\mu = 0.5\sum_{i=1}^{N} \alpha_i \mu_i$ and initial bulk modulus κ is defined by formula $\kappa = 2/d_1$. For a simplified, uniaxial tension, taking incompressibility into account, the relationship between the stress and stretch, Ogden's strain energy density function can be defined as follows:

$$\sigma = \sum_{i=1}^{N} \frac{2G_i}{\alpha_i}\left(\lambda_1^{\alpha_{i-1}} - \lambda_1^{-0.5\alpha_{i-1}}\right) \tag{2}$$

For $N = 1$ and $\alpha_1 = 2$, the Ogden material model is equivalent to the Neo-Hookan model. For $N = 2$ and $\alpha_1 = 2$ and $\alpha_2 = -2$, the Ogden model is equivalent to the two parameter Mooney-Rivlin material model.

Computer simulations of the porcine spinal cord tension using a finite element method (FEM) were conducted in the ANSYS 5.7 software. Two numerical models were built of different spinal cord cross-sectional topologies. The first model included only the gray and white matters, while in the second model, the dura mater together with the pia mater were added. While taking into account the rheological properties of the materials, the numerical investigations incorporated the mechanical properties of the individual areas of the spinal cord [Table 3]. Parameters were determined on the basis of the experiments conducted by Ichihara *et al.* [34], Ozawa *et al.* [32], Wilcox *et al.* [35], as well as the work of Czyż *et al.* [4].

Table 3
The mechanical properties of the anatomical structures of the spinal cord [4], revised by the author

Material	Young's modulus [MPa]	Poisson's ratio	References
Grey matter	0.656	0.499	Ichihara et al. 2003 [34]
White matter	0.277	0.499	Ichihara et al. 2003 [34]
Pia mater	142	0.45	Wilcox et al. 2003 [35]
Dura mater	2.3	0.3	Ozawa et al. 2004 [32]

The analysis makes use of a simplified geometry in cross section (an ellipse) of the spinal cord. The geometric model in a 2D spinal cord was made in SolidWorks 2013, taking into account the thickness of the pia mater is constant and equals 0.1mm and the dura mater equals 0.4 mm [52]. On the basis of the average sample size used in experimental testing, a 2D geometric model was created in ANSYS to the length of 100mm to form the 3D geometry. The discrete model takes into account the longitudinal plane of symmetry. The tested experimental anatomical deformation of the preparations was above 60%; thus, the values were increased and the numerical models were adjusted by stretching them up to 70 mm. Finite elements: SOLID186 (20-nodes) and SHELL93 (8-nodes) were used. Numerical models contained from 13,450 to 21,220 finite elements.

The hyperelastic constitutive relation used to describe the tissue response was derived from the two parameters Ogden material model (Table 4). The Ogden strain energy potentials are nonlinear in terms of the constants. The Marquardt-Levenberg method for nonlinear least squares fitting procedure was used.

Table 4
Fitted coefficients of mechanical properties Ogden hyperelastic model of the anatomical structures of the spinal cord

Material	**Constant μ_1 [kPa]**	**Constant α_1 [kPa]**	**Constant d_1 [1/kPa]**
Grey matter	32	4.7	6.4656
White matter	32	4.7	6.4656
Pia mater	0.13	0.0	299.85
Dura mater	1200	16.2	0.1724

In the model without the dural sac (Figure 7) the highest equivalent stress was observed in the gray matter at 0.488 MPa (Figure 7A). Because of the boundary conditions adopted at the ends of the spinal cord model (no transverse displacements), relatively large equivalent strain, approximately 135% (Fig. 7B), appeared locally on the white matter. Within a significant distance from the boundary disturbance zone, the maximum equivalent strain occurred equal to approximately 85%.

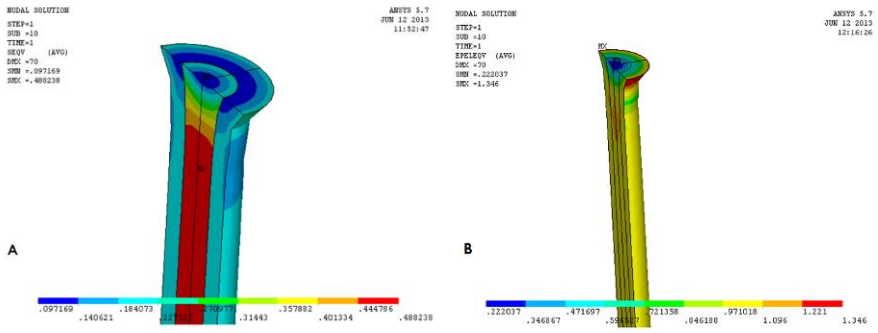


Figure 7

The spinal cord with only the gray and white matters: A – The equivalent stress [MPa] in a state of uniaxial tension, B - The equivalent strain [x100 =%]

In the dural sac's model (Figure 8), the maximum local equivalent stress in the clamping area, in the spinal cord equals 128.6 MPa. In this case, the boundary effect induced by the adopted boundary conditions (Figure 8A) is clearly visible. From a practical point of view, the most important data about the stress and the strain can be measured in the middle part of the model. In this area, the maximum equivalent stress is about 100MPa in the outer dura mater, with the strain of about 107%. The gray matter and white matter stress did not exceed 0.03 MPa. The dural sac, with much higher values of Young's modulus and Poisson's ratios lower, when compared to the both kinds of matter, protects them from the high stress and, by doing so, also protects from the spinal cords rupture.

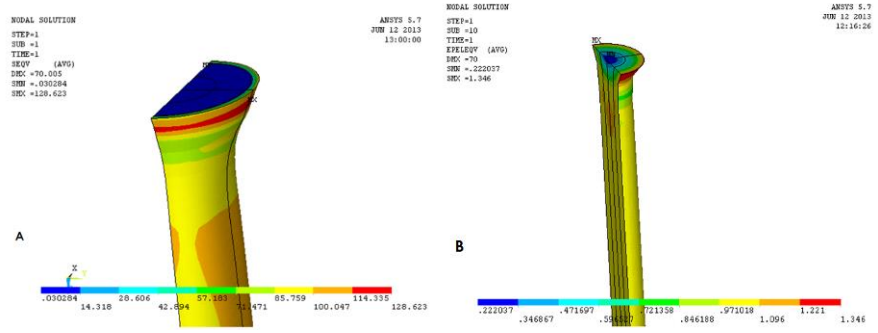


Figure 8

The spinal cord with the gray and white matters, the pia, mater and the dura mater. A - the equivalent stress [MPa] B - the equivalent strain [x100 = %]

A force-displacement diagram has been obtained for the simulation of the spinal cord's tensile testing in the ANSYS 5.7 program (Figure 9).

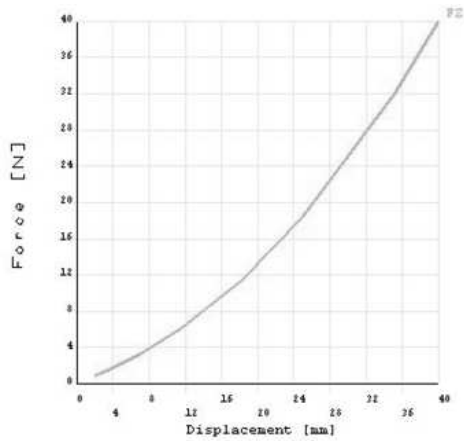


Figure 9

Characteristics of the axial force-displacement of the spinal cord obtained with the ANSYS software

In order to analyze the influence of the cerebrospinal fluid, when modeling the mechanisms of the spinal cord's injury, the next model includes such structures, as: the dura mater, the CSF, as well as, the white and gray matters, similar to what has been done by Maikos [53]. The CSF has been modeled using bulk modulus 6.67 kPa, Poisson's ratio 0.49 [53]. The properties of the spinal cords' remaining elements have been modeled in a way that is shown in Table 3. In this model (Figure 10), the largest reduced deformations caused by the tension testing were localized in the area of the CSF's occurrence. The equivalent and principal strains are presented in Figure 11. The measurement was performed in half of the length of the spinal cord in the cross section.

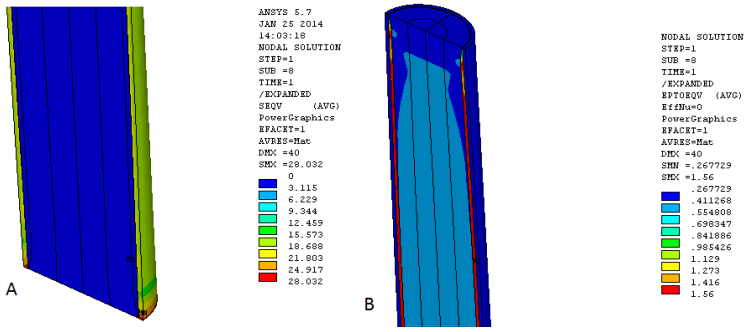


Figure 10

The spinal cord: the dura mater, the CSF, the white and gray matters: A – equivalent stress [MPa] in the state of uniaxial tension. B –equivalent strain [x100 = %]

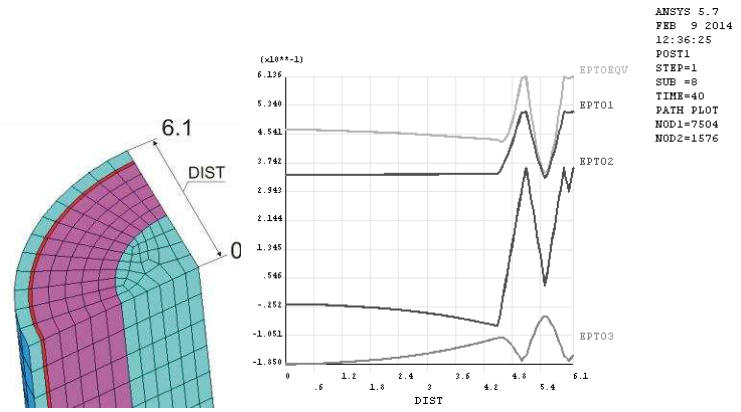


Figure 11

Strains into cross section area of the spinal cord measured in the middle of the length of the cord, (equivalent von Mises strain EPTEQV, principal strain EPT01, 2, 3 [x100 = %].

The analysis of the results has been carried out in the STATISTICA software, ver. 10.0 (StatSoft, Poland).

The *in vitro* tensile testing attempts have been compared with the data coming from the ANSYS program (Figure 12). In order to clearly present the obtained results, the nonlinear regression method has been employed.

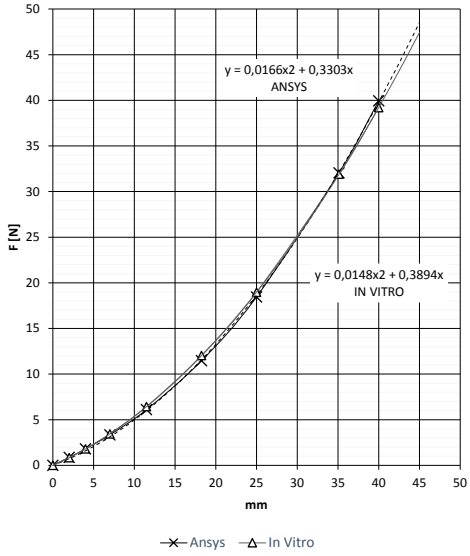


Figure 12

Comparison of FEM results and experimental data: Axial force Fz [N] versus displacement [mm]

The conducted comparative analyses show that the obtained experimental results are highly congruent with the data coming from the computer simulations.

Discussion and Conclusions

On the basis of the mechanical properties of the spinal cord in a state of uniaxial tension, nonlinear characteristics of the force-displacement system have been obtained. The nonlinear behavior of the spinal cord may arise from the extension of the individual fibers in the tissue during stretching.

The spinal cord experiences changes in its mechanical properties, after the death of the animal [54]. After 6 hours stiffness increases in all biological tissues [29]; therefore, all the samples were tested within three hours after death. During the test it was observed that, even though the sample was torn in the middle of the length measurement, the first dura mater injuries occurred near the area of the mechanical connection. Similarly, it has been demonstrated in a numerical model that the local stress concentrations are highest in the area close to the attachment of the tissue. Within the numerical model of the spinal cord with the dural sac, in the area of the primary damage to the dural sac the highest scalar strain took place. The spinal cord of the pig and the rabbit increased its length by more than half of its original length before the complete rupture. The average maximum force acting on the spinal cord of the pig was almost seven times higher than the maximum force acting on the spinal cord of the rabbit. In the computer simulation, after the addition of the dural sac to the model, the observed stress levels were much lower. This fact confirms the claims about the protective function of these structures.

Similar experiments using uniaxial tension on the spinal cord in an *in vitro* setting were carried out by Bilston and Thibault [31], Oakland *et al*. [29], and Clarke *et al*. [27] (Table 5). Young's modulus values that were obtained in this test amounted to 0.323 MPa for the porcine spinal cord and 0.106 MPa for the rabbit samples and were similar to the results obtained *in vivo* by Hung *et al*. [24] (Table 5). Strain values, which were obtained in the study – 61.1% for the porcine spinal cord and 68% for the spinal cord of the rabbit – were very different from the results listed in Table 5. A similar Young's modulus and tissue strain values were obtained by Ichihara *et al*. in 2001 [55], in the studies of the mechanical properties of the gray and white matters in an *in vivo* tension test (Table 6).

The results may differ, due to the differences in the scrutinized sections of the spinal cord, the use of different strain rates, the assembly of the samples in the testing machine, the testing apparatus, in general, the time elapsed since death, species differences, the physical condition and the age of animals and the quantity of the tested preparations [56].

Table 5
Summary of the results from the studies over the mechanical properties of the spinal cord – uniaxial tensile test [54], revised by the author

Reference	Specimen	Region	Environment	Size (mm)	Number of samples	Max. strain	Strain rate	Young's modulus (MPa)
Clarke *et al.* (2009)	Rat (14 days)	Not specified	In vitro	Not specified	8	5%	$0.002s^{-1}$, $0.02s^{-1}$, $0.2s^{-1}$	0.010, 0.013, 0.015
Oakland *et al.* (2006)	Cow	Not specified	In vitro	130 – 80	1	~ 8.5%	$0.24\ s^{-1}$	1.19
Hung *et al.* (1981c)	Cat	T8 – L1	In vivo	25	4	8 – 12%	$0.0008s^{-1}$	0.4
Hung and Chang *et al.* (1981a)	Puppy (3 – 5kg)	L1 – L2	In vivo	8	3	1.7%	$0.003s^{-1}$	0.265
Bilston and Thibault (1996)	Human (30 – 84 years)	Cervical and thoracic	In vitro	30 – 45	3	~10%	$0.048s^{-1}$, $0.120s^{-1}$, $0.225s^{-1}$	1.02, 1.17, 1.37
Polak *et al.* (2014)	Domestic pig	Cervical (C1 – C7) – denticulate ligaments	In vitro	Shown as a triangle - the calculated cross-sectional area was 0.45 mm^2	98 (mean)	2.35 %		1.95

Table 6
Mechanical properties of the gray and white matters [54]

Reference	Specimen	Region	Environment	Length	No. of samples	Max. deformation	Speed of deformation	Young's module
Ichihara *et al.* (2001)	Cow (2 y. o.)	C3 white matter	In vitro	17 mm	6	40% (until failure)	$0.05\ s^{-1}$	0.166 MPa
Ichihara *et al.* (2001)	Cow (2 y. o.)	C3 gray matter	In vitro	17 cm	6	55% (until failure)	$0.05\ s^{-1}$	0.025 MPa

It is likely that there are significant differences in the gray to white matter ratios and the degree of vascularization between humans and different animal species. Estes and McElhaney, upon comparing the rhesus macaque (*Macaca mulatta*) tissues to human ones, established that human tissues are more deformable [27]. It should also be emphasized that in the case of an *in vitro* setting, the tissue response to the external load is a bit different than in the living body. Some of the many reasons for this may be, the degradation of the post-mortem tissue, or the pressure of blood perfusion in the spinal cord in a living body. Perfusion pressure can cause the hydraulic stiffness effect, which can possess the means of influencing the behavior of the tissue [28]. The results obtained in the work show the *in vitro* tissue response to the tensile forces. One should be careful in interpreting these results, when relating to human tissues. During a traumatic spinal cord injury, different forces may have an effect on the tissue, such as: compression, shearing, or twisting. In order to obtain valid results, when it comes to numeric models of the spinal cord, which would prove helpful in individual diagnostics and recuperation of the patients with the spinal cord injuries, it is necessary to include all of the parts constituting the spinal cord [33].

Further studies leading to a better understanding of the central nervous system should provide *in situ* measurements of constitutive compounds in the tissues [57], which take into account the effects of blood, with its pressure, as well as, the cerebrospinal fluid and the flow of both liquids, while also taking into account, the mechanotransduction process. The biomechanical configurations of living subjects are not yet well described, which makes numerical modeling all the more demanding. The development of automated equipment and computer-aided surgical treatments [58] emphasize the need for further studies of the mechanical properties of the spine – spinal cord interactions, with the inclusion of the intraoperative and external loads, *e.g.* surgical instruments. This knowledge is necessary to determine the boundary conditions for the mathematical descriptions of tissues for diagnosis, treatment and prognosis of injuries and diseases of the spinal cord. Determining these values will also allow for the prediction of the cord tissue's translocation during surgery, immediately after the injury, for example, and the removal of splinters of bone.

Acknowledgement

The authors wish to thank Dr. Zbigniew Zawada for providing animal samples.

References

[1] Kiwerski J, Kowalski M, Krakuski M, Schorzenia i urazy kręgosłupa, Wyd. PZWL, Warszawa, 1997

[2] Sharma HS, Pathophysiology of Blood-Spinal Cord Barrier in Traumatic Injury and Repair, in Banks WA (ed.), Current Pharmaceutical Design, Bentham Science Publishers, 11(11): 1353-1389, 2005

[3] Maikos JT, Shreiber DI, Immediate Damage to the Blood Spinal Cord Barrier due to Mechanical Trauma, in Povlishock JT (ed.), Journal of Neurotrauma, National Neurotrauma Society, Mary Ann Liebert 24(3): 492-507, 2007

[4] Czyż M, Ścigała K, Jarmundowicz W, Będziński R, The Biomechanical Analysis of the Traumatic Cervical Spinal Cord Injury using Finite Element Approach, in Acta of Bioengineering and Biomechanics, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 10(1): 43-54, 2008

[5] Galle B, Ouyang H, Shi R, Nauman E, Correlations between Tissue-Level Stresses and Strains and Cellular Damage within the Guinea Pig Spinal Cord White Matter, in Guilak F, Journal of Biomechanics 40: 3029-3033, 2007

[6] Lu J J, Benzel E C Biomechanics of the Spinal Cord, in Seminars in Spine Surgery 17(1):13-18, 2005

[7] Miller K, How to Test Very Soft Biological Tissues in Extension?, in Journal of Biomechanics, 34: 651-657, 2001

[8] Miller K, Wittek A, Joldes G, Biomechanics of the Brain for Computer-Integrated Surgery, in Acta of Bioengineering and Biomechanics 12(2): 25-37, 2010

[9] Joldes GR, Wittek A, Miller K, Suite of Finite Element Algorithms for Accurate Computation of Soft Tissue Deformation for Surgical Simulation, in Medical Image Analysis 13: 912-919, 2009

[10] Czyż M, Ścigała K, Jarmundowicz W, Będziński R, Numerical Model of the Human Cervical Spinal Cord – the Development and Validation, Acta of Bioengineering and Biomechanics, 13(4): 51-58, 2011

[11] Czyż, M, Ścigała K, Będziński R, Jarmundowicz W, Finite Element Modelling of the Cervical Spinal Cord Injury – Clinical Assessment. Acta of Bioengineering and Biomechanics 14(4): 23-29, 2012

[12] Melińska A, Czamara A, Szuba L, Będziński R, Klempous R. Balance Assessment during the Landing Phase of Jump-Down in Healthy Men and Male Patients after Anterior Cruciate Ligament Reconstruction. Acta Polytechnica Hungarica 12(6): 77-91, 2015

[13] Bharathi S, Sudhakar R, Balas VE. Hand Vein-based Multimodal Biometric Recognition. Acta Polytechnica Hungarica 12 (6): 213-229, 2015

[14] Dobránszky J, Ring G, Bognár E, Kovács R, Bitay E. New Method for Evaluating the Visibility of Coronary Stents. Acta Polytechnica Hungarica 11(5):81-94, 2014

[15] Luna C, Detrick L, Shah SS, Cohen AH, Aranda-Espinoza H, Mechanical Properties of the Lamprey Spinal Cord: Uniaxial Loading and Physiological Strain, J Biomech 46(13): 2194-2200, 2013

[16] Brett PN, Fraser CA, Henningan M, Griffiths MV, Kamel Y, Automatic Surgical Tools for Penetrating Flexible Tissues, IEEE Engineering Medicine and Biology 14(3): 264-270, 1995

[17] Villotte N, Glauser D, Flury P, Burckhardt CW, Conception of Stereotactic Instruments for the Neurosurgical Robot Minerva, Engineering in Medicine and Biology Society, 1992 14th Annual International Conference of the IEEE 3: 1089-1090, 1992

[18] Taylor RH, Mitterlstadt BD, Paul HA, Hanson W, Kazanzides P, Zuhars JF, Williamson B, Musits BL, Glassman E, Bargar WL, An Image-Directed Robotic System for Precise Orthopaedic Surgery, in Taylor R. H., et al. (eds), Computer-Integrated Surgery: Technology and Clinical Applications, MIT Press, pp. 379-395, 1995

[19] Sackier JM, Wang Y, Robotically Assisted Laparoscopic Surgery: from Concept to Development, in Taylor RH, et al. (eds), Computer-Integrated Surgery: Technology and Clinical Applications. MIT Press, pp. 577-580, 1995

[20] Schenker PS, Das H, Ohm TR, A New Robot for High Dexterity Microsurgery, Proc CVRMed95. Lecture Notes Computer Science, 905, Springer-Verlag, pp. 115-122, 1995

[21] Burdea, G, Force and Touch feedback for Virtual Reality, Wiley, New York, 1996

[22] Joldes GR, Wittek A, Miller K, Real-Time Nonlinear Finite Element Computations on GPU – Application to Neurosurgical Simulation, Comput. Methods Appl. Mech. Engrg 199: 3305-3314, 2010

[23] Tunturi AR (1978), Elasticity of the Spinal Cord, Pia, and Denticulate Ligament in the Dog, Journal Neurosurgery, 48(6): 975-979, 1978

[24] Hung TK, Chang GL, Biomechanical and Neurological Response of the Spinal Cord of a Puppy to Uniaxial Tension, Journal of Biomechanical Engineering 103(1): 43-47, 1981a

[25] Hung TK, Chang GL, Chang JL, Albin MS, Stress–Strain Relationship and Neurological Sequelae of Uniaxial Elongation of the Spinal Cord of Cats, Surgical Neurology, 15(6): 471-476, 1981b

[26] Hung TK, Chang GL, Lin HS, Walter FR, Bunegin L, Stress–Strain Relationship of the Spinal Cord of Anesthetized Cats, Journal of Biomechanical Engineering, 14(4): 269-276, 1981c

[27] Clarke EC, Cheng S, Bilston LE, The Mechanical Properties of Neonatal Rat Spinal Cord in Vitro, and Comparisons with Adult, Elsevier, Journal of Biomechanics, Australia 42:1397-1402, 2009

[28] Fiford RJ, Bilston LE, The Mechanical Properties of Rat Spinal Cord in Vitro, Journal of Biomechanics, Australia, 38:1509-1515, 2005

[29] Oakland RJ, Hall RM, Wilcox RK, Barton DC, The Biomechanical Response of Spinal Cord Tissue to Uniaxial Loading, Proceedings of the Institution of Mechanical Engineers [H], 220(4):489-492, 2006

[30] Shetye SS, Troyer KL, Streijger F, Lee JHT, Kwond BK, Cripton PA, Puttlitz CM, Nonlinear Viscoelastic Characterization of the Porcine Spinal Cord, Acta Biomaterialia 10(2): 792-797, 2014

[31] Bilston LE, Thibault LE, The Mechanical Properties of the Human Cervical Spinal Cord in Vitro, Annals of Biomedical Engineering, 24(1): 67-74, 1996

[32] Ozawa H, Matsumoto T, Ohashi T, Sato M, Kokubun S, Mechanical Properties and Function of the Spinal Pia Mater, Journal of Neurosurgery. Spine, 1: 122-127, 2004

[33] Polak K, Czyż M, Ścigała K, Jarmundowicz W, Będziński R, Biomechanical Characteristics of the Porcine Denticulate Ligament in Different Vertebral Levels of the Cervical Spine—Preliminary Results of an Experimental Study, Journal of the Mechanical Behavior of Biomedical Materials 34 (2014): 165-170, 2014

[34] Ichihara L, Taguchi T, Sakuramoto I, Kawano S, Kawai S, Mechanism of the Spinal Cord Injury and the Cervical Spondylotic Myelopathy: New Approach based on the Mechanical Features of the Spinal Cord White and Grey Matter, J. Neurosurg., 99, Suppl. 3: 278-285, 2003

[35] Wilcox RK, Bilston LE, Barton DC, Hall RM, Mathematical Model for the Viscoelastic Properties of Dura Matter, J Orthop. Sci., 8: 432-434, 2003

[36] Kwan M, Wall E, Massie J, Garfin S, Strain, Stress and Stretch of Peripheral Nerve: Rabbit Experiments in Vitro and in Vivo, Acta Orthopaedica Scandinavica, 63: 267-72, 1992

[37] Ozawa H, Matsumoto T, Ohashi T, Sato M, Kokubun S, Comparison of Spinal Cord Gray Matter and White Matter Softness: Measurement by Pipette Aspiration Method, Journal of Neurosurgery: Spine 95: 221-224, 2001

[38] Sheng, SR, Wang XY, Xu HZ, Zhu GQ, Zhou YF, Anatomy of Large Animal Spines and its Comparison to the Human Spine: a Systematic Review, Eur. Spine J. 19: 46-56, 2010

[39] Sparrey CJ, Keaveny TM, The Effect of Flash Freezing on Variability in Spinal Cord Compression Behavior, J. Biomech. Eng, 131: 111010, 2009

[40] Sparrey CJ, Keaveny TM, Compression Behavior of Porcine Spinal Cord White Matter, J. Biomech. 44: 1078-1082, 2011

[41] Maikos J, In Vivo Tissue Level Thresholds for Spinal Cord Injury, New Brunswick, New Jersey, 2007

[42] Winkelstein BA, Myers BS, The Biomechanics of Cervical Spine Injury and Implications for Injury Prevention, Med Sci Sports Exerc 29(7 Suppl): S246-55, 1997

[43] Yoganandan N, Kumaresan S, Voo L, Pintar FA, Finite Element Model of the Human Lower Cervical Spine: Parametric Analysis of the C4-C6 unit, J. Biomech. Eng. 119(1): 87-92, 1997

[44] Willinger R, Kang HS, Diaw B, Three-Dimensional Human Head Finite-Element Model Validation against Two Experimental Impacts, Ann Biomed Eng 27(3): 403-410, 1999

[45] Wheeldon J, Khouphongsy P, Kumaresan S, Yoganandan, N, Pintar FA, Finite Element Model of Human Cervical Spinal Column, Biomed. Sci. Instrum. 36: 337-342, 2000

[46] Tillier Y, Paccinia A, Durand-Revilleb M, Baya F, Chenota JL, Three-Dimensional Finite Element Modelling for Soft Tissues Surgery 2003, International Congress Series 1256: 349-355, 2003

[47] Miller K, Chinzei K, Orgssengo G, Bednarz P, Mechanical Properties of Brain Tissue In-Vivo: Experiment and Computer Simulation, J. Biomech. 33(11): 1369-1376, 2000

[48] Miller K, Jia L, On the Prospect of Patient-Specific Biomechanics without Patient-Specific Properties of Tissues, 2013 Journal of the mechanical behavior of biomedical materials 27: 154-166, 2013

[49] Ogden, RW, Large Deformation Isotropic Elasticity - on the Correlation of Theory and Experiment for Incompressible Rubberlike Solids, Proceedings of the Royal Society London Series A 326: 565-584, 1972

[50] Fung YC, Biomechanics - Mechanical Properties of Living Tissues second ed., Springer-Verlag, New York, 1993

[51] Miller K, Chinzei K, Mechanical Properties of Brain Tissue in Tension, Journal of Biomechanics 35: 483-490, 2002

[52] Nicholas DS, Weller RO, The Fine Anatomy of the Human Spinal Meninges. A Light and Scanning Electron Microscopy Study, J. Neurosurg. 69: 276-282, 1988

[53] Maikos JT, Qian Zh, Metaxas D, Shreiber DI, Finite Element Analysis of Spinal Cord Injury in the Rat, Journal of Neurotrauma 25: 795-816, 2008

[54] Cheng S, Clarke EC, Bilston LE, Rheological Properties of the Tissues of the Central Nervous System, Medical Engineering & Physics, Australia, 30: 1318-1337, 2008

[55] Ichihara K, Taguchi T, Shimada Y, Sakuramoto I, Kawano S, Kawai S, Gray Matter of the Bovine Cervical Spinal Cord is Mechanically more

Rigid and Fragile than the White Matter, J Neurotrauma 18(3): 361-367, 2001

[56] Będziński R, Biomechanics, Komitet Mechaniki PAN, Wyd. IPPT PAN, Warsaw, 2001 (in Polish)

[57] Elliott NSJ, Bertram CD, Martin BA, Brodbelt AR, Syringomyelia: A Review of the Biomechanics, Journal of Fluids and Structures, 40(2013): 1-24, 2013

[58] Takács Á, Kovács L, Rudas IJ, Precup R, Haidegger T. Models for Force Control in Telesurgical Robot Systems. Acta Polytechnica Hungarica 12(8): 95-114, 2015

Cogging Torque Reduction by Magnet Pole Pairing Technique

Szilárd Jagasics, István Vajda

Óbuda University, Bécsi út 96/b, H-1034 Budapest, Hungary
jagasics.szilard@kvk.uni-obuda.hu, vajda@uni-obuda.hu

Abstract: A high performance electrical drive needs a smooth torque waveform and a high torque to inertia ratio. The power density and performance needs can be, in most cases, fulfilled by using a permanent magnet synchronous machine (PMSM). This paper explores a new cogging torque reduction technique. This method can be used without reducing the power density of the machine and it can also be applied in a mass production process.

Keywords: cogging torque; finite element analysis; optimization; pulsating torque reduction

1 Introduction

Pulsating torque is usually harmful for electric drives. They can create disturbing vibration and noise which need to be eliminated. The pulsating torque of PMSM machines is the sum of torque ripple and cogging torque.

Torque ripple is produced if the induced voltage graph of the machine or the power inverter has harmonic content. Cogging torque is a magnetostatic effect: pulsating torque arises due to magnetic energy variation in the air gap as the rotors magnet pole passes over a slot opening. The pulsating torque components have a usual user accepted level: it is defined in the ratio of the rated torque, which is usually 0.5% for cogging torque and 3% for torque ripple.

There are many well-known cogging torque minimizing techniques that are able to maintain a defined pulsating torque range. These methods may use of dummy slots, magnet poles or slot skewing and can effectively decrease a pulsating torque component, but they can also decrease the torque to inertia ratio of the machine.

There are many other cogging torque and torque ripple reducing methods presented, but usually these methods are effective for only one pulsating torque component. A method which is effective for cogging torque reduction is usually not effective for torque ripple reduction. Moreover, sometimes a cogging torque reducing method produces high torque ripple. The designer has to find a solution that is effective, for both pulsating torque components, this paper presents such a solution.

2 Physical Background

A PMSM machine has several magnet poles and several slots. Pulsating torque wave is generated if a magnet pole edge passes a slot opening. The shape of the wave is the function of the slot opening and magnet pole design. Different magnet pole shape may produce a significantly different cogging wave for the same slot opening (Figure 1). The designers try to create such a magnet pole shape that produces the most optimal cogging wave. The wider slot opening produces the higher amplitude for cogging wave. The slope of the cogging wave is usually the function of the magnet pole shape. If the magnet width is constant, the cogging wave is like the red one on Figure 1. If the magnet width is maximal at the middle of the magnet and becomes narrower towards the pole edges the slope of the cogging torque wave gets lower, like the blue (dashed) wave in Figure 1.

The individual magnet pole-slot opening related cogging torque graphs can be simulated by finite element analysis for each slot. The identical waves are summarized mechanically by the stator and rotor lamination for all slots and poles. The key question is the physical distribution of the magnet poles and stator slots which is defined by the magnet pole-stator slot number combination. In another aspect of view, the resultant cogging torque wave of the machine is the summing of different graphs which are in the same phase or have phase offset between each other, this phase offset is defined by the mechanical position of the slots and magnet poles.

In some cases the slot number-pole number combination gives the opportunity to use such a pole pitch ratio, that the cogging torque graph of the two magnet pole edge of an individual magnet pole is in the same phase, but with opposite sign. That is, the two cogging waves may cancel each other.

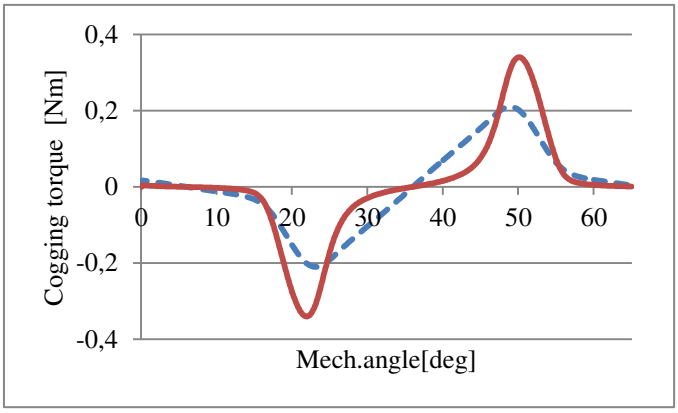


Figure 1
Cogging torque graph of an individual magnet pole for different magnet pole shape design.

The other opportunity is to use special magnet positioning technique. The positioning of the magnets may be done so to have magnet pole edge pairs having the individual cogging waves in the same phase but with opposite sign. By this way some of the individual cogging torque graphs can be cancelled and the cogging torque level of the complete machine can be effectively reduced. This self-cancelling cogging torque reducing technique is useful because it doesn't reduce the torque density of the machine. Special manufacturing technology is not needed either.

The individual cogging torque of one magnet pole and slot opening can be calculated by finite element method (FEM). The summing of these graphs can be done by analytical equations.

Let's call the identical cogging torque wave for one magnet pole f_{sp}. This wave can be accurately calculated by FEM. The slot number (Z) and pole number $(2p)$ is known for an analyzed machine. The mechanical angle between slots (β) and poles (α):

$$\alpha = \frac{360°}{2p}, \qquad \beta = \frac{360°}{Z}, \quad \gamma = \frac{360°}{LCM(2p, Z)} \tag{1}$$

The period of cogging torque of the complete machine is γ, $\mathrm{LCM}(2p, Z)$ is the least common multiple of the slot and pole number. The cogging torque wave of the machine is the summary of the unique cogging waves of each slot and magnet pole.

The cogging torque wave for one slot and the whole rotor for one complete mechanical rotor revolution can be generated: f_{sp} has to be summed for each magnet pole by adding α mechanical phase offset for each magnet pole. Also magnet positioning error (φ_i) can be taken into account. The cogging torque wave for one slot for one rotor revolution is the following:

$$f_{s\_360} = \sum_{n=1}^{2p} f_{sp}(x + n \cdot \alpha + \varphi_i) \tag{2}$$

The cogging torque graph of the machine can be created by the sum of the cogging torque graphs of the slots:

$$f_{cogg} = \sum_{m=1}^{Z} f_{s\_360\_m}(x + m \cdot \gamma) \tag{3}$$

The phase offset marked by φ_i may be magnet positioning error or a defined magnet positioning phase offset. If f_{sp} is known, the optimal value for other parameters, like pole pitch ratio, etc., can be found.

These equations can be used to create the cogging torque waveform of a machine from the identical cogging torque wave of one magnet pole-slot opening

interaction. This calculating technique is called hybrid method. Special magnet positioning or magnet pole width manipulation can be also taken into account and the resultant cogging torque wave of the modified machine can be calculated in a short time.

Cogging torque calculation by FEM for different pole positioning cases or for different pole width needs a different model. These models need to be created and calculated which usually takes quite a long time. The hybrid method can be used to find the optimal geometry in a short time and the final calculation and optimization can be done by FEM.

3 Cogging Torque Compensation by Pole Width Modification

Let's check some pole number-slot number combinations in Table 1. For best cogging torque compensation effect slot pitch, or multiplied slot pitch angle should be equal with the pole pitch. That is, f_{sp} waves should be summed in such way to compensate each other.

If the magnet pole pitch is too narrow, the harmonic content of the induced voltage graph will be high and in this case the ripple torque of the machine will be also high, usually higher than the application acceptable level. Table 1 contains the maximal magnet pole pitch values for some pole number cases.

The 27 slot 6 pole machine seems to be a good combination for self-cancelling cogging waves for each magnet pole. The maximal pole pitch for $2p = 6$ machine is 60º. The slot pitch for a $Z = 27$ slot machine is 13,33º. The nearest value around the maximal pole pitch value is 53,33º, which means four slot pitch.

Table 1

Mechanical angle for slot pitch and magnet pole pitch for some pole number and slot number cases

Number of poles	**2**	**4**	**6**	**8**	**10**	**12**
Pole pitch [mech. angle]	180º	90º	60º	45º	36º	30º

Slot number	**9**	**12**	**15**	**18**	**21**	**24**	**27**
Slot pitch*1	40º	30º	24º	20º	17.14º	15º	13.33º
Slot pitch*2	80º	60º	48º	40º	34.29º	30º	26.67º
Slot pitch*3	120º	90º	72º	60º	51.43º	45º	40º
Slot pitch*4	160º	120º	96º	80º	68.57º	60º	53.33º

The f_{cogg} wave for the complete machine was calculated by FEM. The outer diameter of the analyzed machine was 150 mm, length of stator lamination was 100 mm. The air gap length is 1 mm. Rated torque of the machine was 20 Nm.

The cogging torque graphs for the machine for different magnet pole width value magnet can be seen on Figure 2.

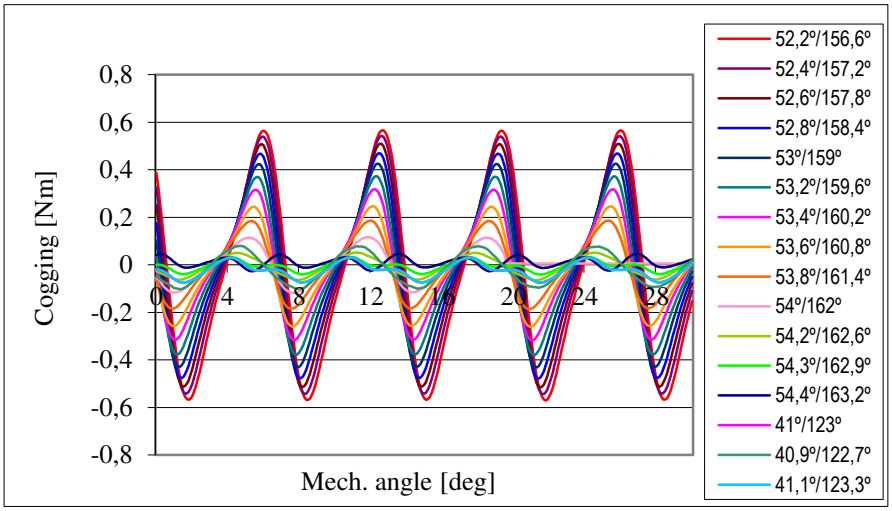


Figure 2

Magnet pole pitch dependence of cogging torque for a 6 pole 27 slot machine

Simulation was made by finite element analysis. The calculation of each graph took about 28 hours, so the complete simulation took more than 18 days. The drawing of the geometry for the rotor versions and also the model building for the simulation took about 2 hour per variant. This time is only calculating time without any rest for the computer. If the timing of the different models is not automatized, and the time gap between the different calculations is not minimal the time consumption may be much higher.

The lowest peak value cogging wave was calculated for the case of 54.3° magnet pole pitch. The 54.3° mechanical pole pitch means 162.9° in electrical angle which is a good value for low torque ripple level. The same wave can be produced if the magnet pole width is not 4 times but 3 times the value of the slot pitch. Please check the cogging torque wave for 54.3° and 41.1° magnet pole pitch. The analysis was made by 0.2° steps.

The self-compensating method has been validated:

$$f_{c_i} = (-1) \cdot f_{c_{i+N}} \tag{4}$$

where f_{c_i} is the identical cogging torque wave for the i^{th} slot, $N = 1,2,3$ ... integer, so $f_{c_{i+N}}$ is an identical cogging torque for a slot in the neighborhood of the i^{th} one.

This comprehensive analysis was made to validate the modeling method and also to check the effect of the diagonal magnetized magnet poles. The optimal pole width was 54.3° and 41.1° in mechanical angle. The slot pitch is 13.33°, 3 slot

pitch is 40º, 4 pole pitch is 53.33º. The pole width is more than 1º wider than the slot pitch.

If the pole pitch is different from the optimal value the amplitude of the sum cogging torque wave increases rapidly. This cogging torque reducing method act sensitively for magnet pole width related manufacturing tolerances. Normally cogging torque graph acts sensitively for magnet positioning error but by the case of the magnet pole self-cancelling method the sensitivity for magnet positioning error is low.

The peak to peak value of cogging torque is 0.08 Nm, which is 0.4% of the rated torque of the machine so the usage of other cogging torque reducing method is not needed.

The finite element optimization of the machine was done on the way of step-by step analysis but time consumption was quite high, a faster calculation method would be preferred.

The shape of the cogging graph can be scaled for different magnet pole width situations and the scaled graphs can be summed. The result can be achieved in a short time by running for example a Matlab script which contains an identical cogging torque graph, the pole number-slot number combination and the different pole width cases. The cogging torque for the rotor variants can be easily plotted and the optimal pole width can be found in a short time. This optimal pole width can be used for the finite element analysis as a starting point and the optimizing process could be much faster.

Let's check the error possibilities for the analytical method. If the magnetization of the magnet is radial, the identical cogging graph can be transformed narrower with same amplitude. If the magnetization is diagonal and the magnet pole width is modified, the shape of the identical cogging torque graph usually changes. If the pole width is thinner, the amplitude of the cogging wave gets higher. At the middle of the magnet pole the magnetization direction is perpendicular with the stator lamination but towards the magnet edges this angle changes and also the orthogonal projection changes. In this case the effective air gap length is higher towards the magnet pole edges and the amplitude of cogging torque decreases.

The variation of identical cogging torque graph was analyzed by finite element analysis for the case of several different magnet pole width magnet version (Figure 3.). If the identical magnet cogging wave would be only scaled to different width, amplitude error would arise. Equation (3) is still valid for the machine having the modified rotor configuration, which means only the exchanging of the f_{sp} wave. The scaling can be made for example by the *interp1* function in Matlab.

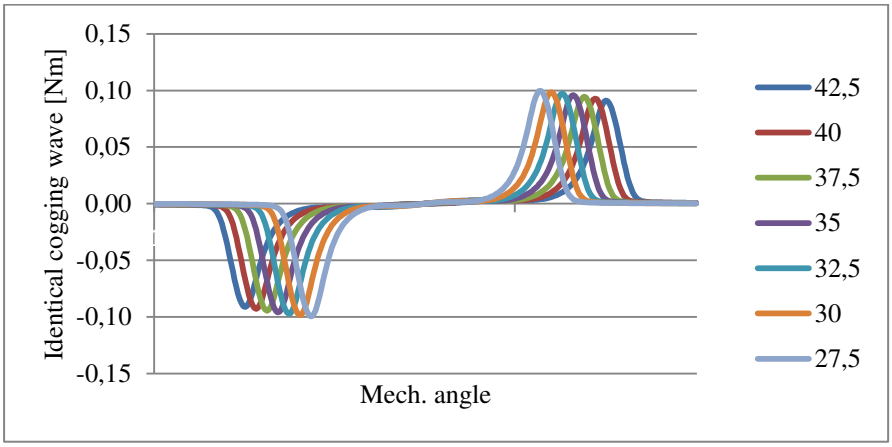


Figure 3
Identical cogging wave for different magnet pole width cases

If the geometry contains some geometry locations with high saturation level some unexpected cogging torque harmonics may appear which cannot be handled by the hybrid analytical method. For example, if the yoke is too narrow of the machine, the saturation level in the yoke region will be different for the identical cogging torque calculation. The identical cogging torque wave is calculated when only two magnet is present in the model. If all magnets are inserted in the model, the saturation level may be much higher.

Let's check the case of a 9 slot 6 pole machine. The magnetization of the magnets is also diagonal. The slot pitch of this machine is 40°. The magnet pole width analysis for cogging torque reduction was also made. The analysis was made by finite element method, results can be found on Figure 4.

The optimal pole width value for the actual slot opening, magnet pole shape and air gap length is 41.9° in mechanical angle. The maximal magnet pole pitch for a 6 pole machine is 60°. The 41.9° mechanical angle pole pitch means 125.7° in electrical angle which is very low and would create high torque ripple level. In this case the magnet pole self-compensating method is not applicable. The effect of diagonal magnetization acts also in this case: the magnet pole width is wider than the slot pitch.

In this case other pulsating torque reducing method should be used like the magnet pole-pairing method. This technique means such a magnet positioning that some magnets are placed by a defined phase offset of their original symmetrical position.

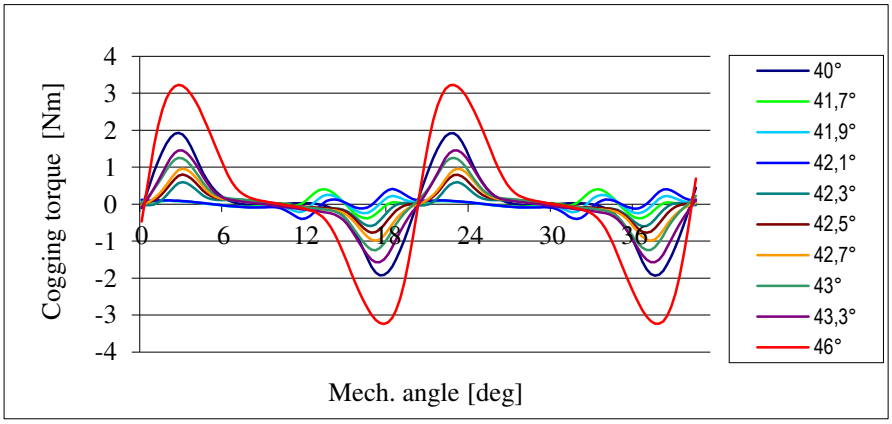


Figure 4

Magnet pole pitch dependence of cogging torque for a 6 pole 9 slot machine

If diagonal magnet is used and magnet pole width is modified, the analytic pole width scaling method may contain higher error level. The optimal pole width value can be tightened to a range of 1°. Another opportunity is the modification of the position of some magnets and/or to take advantage of the cogging compensation effect, for magnet pairs, not for identical magnets as in this point.

4 Pole Pairing Technique

In many slot number-pole number cases the magnet poles can be arranged such a way that the magnet pole self-cancelling method can arise for pole edges of different magnet poles.

This eliminating method means that the positive and negative peaks of the individual cogging waves are in phase for different pole pairs. For better understanding see Figure 5. The magnet pairs A-A and B-B are the self-eliminating pairs for cogging torque. By the actual rotor position one B magnet leaves a slot opening, the other B magnet arrives to a slot opening. The magnets marked by C and D are creating their own cogging torque graphs and their position was not changed. The sum cogging torque graph is only the sum of the graph of magnet C and D.

The identical cogging torque wave does not need to be modified, only the phase angle should be modified, so the analytic result will have very good accuracy. The optimal pole positions were calculated and such a rotor geometry was imported to the FEA software to validate the result. The difference between the optimal pole width value from analytic and FEA result was 0.2°. The final optimization was fast.

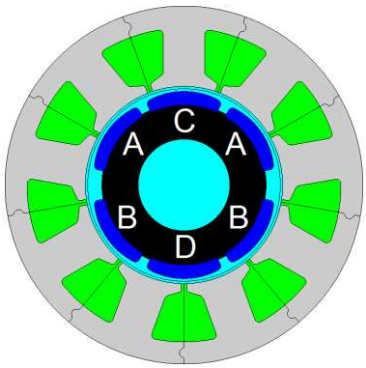


Figure 5
Magnet pole pairing technique applied for a 6 pole 9 slot machine

If the magnet arrangement on the rotor is symmetric, there is 120° mechanical symmetry for the rotor and stator. In this case the cogging torque of the 3 magnet groups is summarized in phase. This effect can be eliminated by the modified magnet arrangement.

The usual applied cogging torque reducing method is magnet pole skewing. The effects of pole skewing and the magnet pole pairing technique can be compared. The outer diameter of the analyzed machine is 80 mm, the stator stack length is 30 mm, rated torque is 3.5 Nm. The comparison of different rotor configurations as regular magnet pole arrangement, pole paired rotor configuration and skewed rotor was analyzed by FEM. The magnet pole shape (pole pitch, magnet material, geometry) was not changed.

Table 2
Comparison of cogging torque peak values for regular, skewed rotor and a rotor with modified magnet positioning

	cogging torque [Nm]	ripple torque [Nm]	rated torque [Nm]
limit	0.0175	0.105	3.50
regular rotor	0.0440	0.350	3.71
pole pairing	0.0064	0.097	3.68
skewed rotor	0.0040	0.320	3.42

The cogging torque of the regular rotor arrangement is higher than the limit. The skewed rotor configuration can pass the cogging torque limit but the ripple torque is higher than the limit. Unfortunately, skewing is only effective for one pulsating torque harmonic. Skewing also reduces the torque density of the machine.

The pole paired rotor configuration can pass both cogging and torque ripple limits and the torque density also remains at a high level. The partial compensation of the identical cogging waves reduces effectively the cogging torque.

Unfortunately, the pole pairing method is not usable for all pole number-slot number combinations. In some cases, the usage of the cancelling technique creates a high torque ripple ratio. The goal of the machine optimization is to achieve low level of cogging torque and torque ripple. In such cases, another pulsating torque reducing method should be used, which may also reduce the torque density of the machine. When the pole pairing method is applicable, it can be effectively used and the torque density of the machine remains higher than using one of the other cogging torque reducing methods.

5 Reduction of Sensitivity for Air Gap Eccentricity

Permanent magnet synchronous machines act are usually sensitive for manufacturing misalignments. The most frequent mechanical misalignment cases for mass-production are air gap eccentricity and magnet positioning error. Magnet positioning error creates a slot number order cogging torque harmonic, air gap eccentricity creates pole number order cogging torque harmonic.

The magnet positioning error can be eliminated by special rotor design: If internal permanent magnet (IPM) rotor design is chosen, the magnets are fixed in the slots of the rotor lamination. The air gap eccentricity problem still has to be investigated.

Air gap eccentricity may arise in various cases: The electric machine is built up of many parts, which are, for rotor side, the shaft, rotor lamination stack and magnets and for stator side, stator housing, laminated stator stack, end-shields (for both sides or only one side) and bearings.

Each of the many parts, have manufacturing tolerances. If the tolerances are too wide-ranging, the parts may need to be grouped or the air gap eccentricity of the machine may vary too much. If the manufacturing tolerances are too strict the scrap ratio may become too high. It is a common scenario to check the allowable air gap eccentricity level of the machine and also to apply the result for defining the tolerances for each part. This is a must, because it is cheaper to select the scrapped main parts then to scrap the complete machine if its cogging torque is higher than the prescribed limit.

A comprehensive finite element analysis was created for checking the sensitivity of the cogging torque for air gap eccentricity on a common PMSM machine pole-slot number combinations. The analysis was made for same size machines: The outside diameter was 150 mm, the stator stack length was 100 mm and the mechanical air gap was 1mm. Three cases were analyzed: The regular case with 0mm, then a 0.2 mm and a 0.4 mm air gap eccentricity. The results can be seen in Figure 6 below.

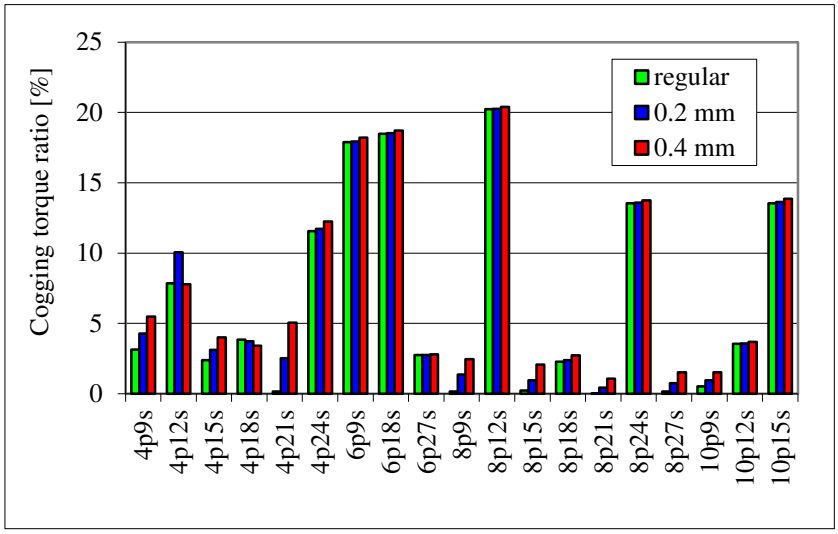


Figure 6
Comprehensive analysis for sensitivity of cogging torque for air gap eccentricity

As it can be seen, the machines can be divided into two groups: one group has low regular cogging torque values which increases rapidly for air gap eccentricity and the other group where the rated cogging torque graph has a high amplitude ,cogging torque graph but the sensitivity for air gap eccentricity is low.

The results can be explained by the hybrid method. The resultant cogging torque graph of the machine can be calculated by the summing of the identical cogging torque graphs.

The shape of the identical cogging torque graph is a function of air gap length. The shape and width of the graph remains the same, but the amplitude changes: higher air gap lengths results in lower cogging torques. The behavior of the identical cogging torque graph for different air gap length cases can be calculated by FEA. This function ($c(g_m)$) is usually defined by the saturation level of the magnetic circuit of the machine. If the saturation level is high, the graph is hyperbolic, if it is low, the $c(g_m)$ graph is almost linear. Also, the shape of the $c(g_m)$ graph is the function of the regular air gap length: if the regular air gap is small, the magnetic circuit become more sensitive.

For a defined air gap eccentricity case, the actual air gap length value for each slot can be calculated. The scale factor can be defined for the identical cogging torque graphs for each slot opening by using the $c(g_m)$ graph. The resultant cogging torque graph can be defined in (4):

$$f_{cogg} = \sum_{m=1}^{Z} c(g_m) \cdot f_{s\_360\_m} \tag{4}$$

The mechanical offset angle between slot dependent $f_{s\_360\_m}$ graphs is γ. If the pole number-slot number combination is relative prime, the offset angle between each $f_{s\_360\_m}$ graph is different. If the pole number-slot number combination has a common divider, there will be some $f_{s\_360\_m}$ graphs in the same γ phase, their number is the function of the common divider.

Let's check (4) and Figure 7. Normally, when air gap eccentricity is not present, the scale factor for $c(g_m)$ would be 1.

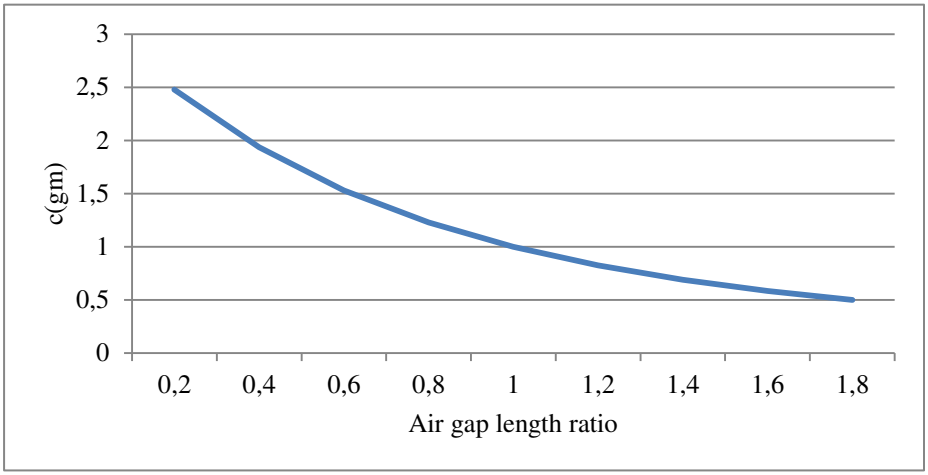


Figure 7

The scale factor graph for identical cogging graph for different air gap length values

If air gap eccentricity is present, the $f_{s\_360\_m}$ graphs should be scaled and then summed. The $c(g_m)$ graph increases heavily for smaller air gap values, that's why the cogging torque amplitude greatly increases for the case of machines having relative prime pole number-slot number combination.

Table 3

γ offset angle for identical cogging graphs for different slots for 10p12s and 8p9s machine

slot.No	1	2	3	4	5	6	7	8	9	10	11	12
γ -10p12s	0°	6°	12°	18°	24°	30°	36°	42°	48°	54°	60°	66°
	0°	6°	12°	18°	24°	30°	0°	6°	12°	18°	24°	30°
γ -8p9s	0°	5°	10°	15°	20°	25°	30°	35°	40°			

For example, a 10 pole - 12 slot machine, has a common divider, it is 2. In this case the $f_{s\_360\_m}$ graphs can be classified in two groups: slot 1-6 and slot 7-12. The γ angle for each slot is 6°, the α angle is 36°. The offset between slot 1 and 7 is 36°, which means one magnet pole shift, so these cogging waves are in the same phase.

In regular case without air gap eccentricity, equation (3) can be written as:

$$f_{cogg} = s \cdot \sum_{m=1}^{Z/s} f_{s\_360\_m}(x + m \cdot \gamma) \quad (5)$$

where s is the common divider between slot number and pole number, and also the number of the mechanical symmetry axes for the machine. For the 10p12s machine $s = 2$. If air gap eccentricity is present, (5) can be expressed as:

$$f_{cogg} = \sum_{m=1}^{Z/2} c(g_m) \cdot f_{s\_360\_m}(x + m \cdot \gamma) + \sum_{m=\frac{Z}{2}+1}^{Z} c(g_m) \cdot f_{s\_360\_m}(x + m \cdot \gamma) \quad (6)$$

The air gap length for slot 1-6 is higher than for the regular value, for slot 7-12 the air gap length is lower than the regular value.

Due to mechanical symmetry, slot 1-7, 2-8 etc. are placed in opposition and the difference from regular air gap for the slot pairs is the same but with opposite sign. The air gap dependent $c(g_m)$ scale factor for (4) can be defined for each slot. For the case of the 10p12s machine, if the $c(g_m)$ graph would be linear, the sum of the scale factor for the higher and lower side $f_{s\_360\_m}$ graphs (slot 1-6, 7-2, etc) would remain 2, the positive and negative deviation from 1 for each slot pair would be the same.

For the case of 8p9s machine the phase offset between the $f_{s\_360\_m}$ is different, the previous compensating effect is not present.

If the value of s is greater than 1, eccentricity compensation is present, but the regular cogging torque peak value is higher because the number of $f_{s\_360\_m}$ graphs that are added in the same phase, is more than 1. If $s = 1$, the sensitivity for air gap eccentricity in the aspect of cogging torque is higher but the regular cogging torque level is lower.

Conclusions

This paper presents a new cogging torque analysis method which is called a "hybrid method". It can be used to easily understand the nature of cogging torque. The cogging torque of the machine can be thought of as the summary of several graphs. The hybrid method can be used in many cases to reduce the cogging torque of the machine or reduce the sensitivity for air gap eccentricity in such a way to be applied in a mass production scenario, without the reduction of the rated torque for the machine.

References

[1] Nicola Bianchi, Silverio Bolognani: Design Techniques for Reducing the Cogging Torque in Surface-Mounted PM Motors, IEEE Transactions on Industry applications, 2002, pp. 1259-1265

[2] Min Dai, Ali Keyhani, Tomy Sebastian: Torque Ripple Analysis of a PM Brushless DC Motor Using Finite Element Method, IEEE Teansactions on Energy Conversation, 2004, pp. 40-45

[3] Szilard Jagasics: Comprehensive Analysis on the Effect of Static Air Gap Eccentricity on Cogging Torque, IEEE RAAD 2010, pp. 447-449

Effects of Increasing the Power of Retail Chains on Competitive Position of Producers

Stipe Lovreta[1], Jelena Končar[2], Ljiljana Đ. Stanković[2]

[1] Faculty of Economics, University of Belgrade, Kamenicka 6, 11000 Belgrade, Serbia; slovreta@ekof.bg.ac.rs

[2] Faculty of Economics Subotica, University of Novi Sad, Segedinski put 9, 24000 Subotica, Serbia; koncarj@ef.uns.ac.rs; ljstankovic@ef.uns.ac.rs

Abstract: There is a retailing revolution in progress, whose basic characteristic is strengthening the power of retailers. This strengthening causes shifting of the powers between members in the marketing channels and thus it leads to the change of positions and relationships within the channel. The trade revolution affects all the segments of goods and services, and it reflects, with the greatest extent and with special content, on the sphere of food and other products of everyday purchase and consumption. Precisely in this domain, the position of the large retail chains extremely strengthens in comparison to other players in the marketing channels. The effects which cause contemporary changes in the sphere of retail on relationships in the marketing channels are of substantial importance for functioning and survival of the channels. The question of a new position of producers in them is of special importance. All those, and especially smaller producers, are highly affected by the situation in which they are not able to resist the accumulated purchasing power of retailers. However, despite the fact that producers both face the ever-growing and more complex demands the large retail chains impose on them and at the same time become more dependent to the decreasing number of more and more powerful retailers, they also recognize the advantages of doing business with such partners. Simultaneously, the role of modern state is of high relevance. A state is expected to precisely and comprehensively define adequate rules for market players and to work intensively on their enforcement and sanctioning their infringement.

Keywords: marketing channels; retail chains; producers, power

1 Introduction

Strengthening the retailer's power is a basic characteristic of the modern retailing revolution. It characterizes the commercial life in the European Union and other developed trade economies in which the retail structure has been transformed from the presence of numerous minor and independent retailers into the existence of dominant, strong national and international retail chains. Retailers are rapidly

growing, developing and their power in marketing channels constantly increases. They strengthen their market share and market power, and at the same time increase retail market concentration. In accordance with the stated, an ever-growing part of the market is handled by a limited number of traders that become gatekeepers for approaching consumers [8].

Retailers become individually stronger, and retail as a sector, at national, as well as in the global market, becomes ever stronger. Parallel to increases of retail market concentrations and the increases of market shares of the individual retail chains, the power of retailers is constantly increasing compared to the power of the other members of the marketing channels [14]. Large retailers which undergo the process of restructuring, also apply modern technological innovations and conduct internationalization of their activities, become the mainspring and main force for development of trade and market within the national and international framework.

Measured in volume of turnover, property or share capital, retailers are today, in most trade groups, significantly bigger than producers. They are starting to dictate conditions even to those huge global producers whose prevailing part of production is put on the market exactly through global trade giants. As producers become more dependent on retailers, their negotiating position weakens and as a result invest more efforts in building relationships with retailers, and less in building relationships with the consumers. In this kind of situation, retail takes over numerous marketing functions which were, traditionally, performed by wholesalers and producers. Retail also takes over the leading role in developing relations with consumers, and thus aquires the dominant position within the marketing channels. At the same time, as a consequence, numerous members of marketing channels become irrelevant.

Shifting the power from producers to retailers relates primarily to increasing concentration of retail industry, successful introduction of private label products, development of a concept of category management by retailers, ever growing application of informational technology in retail, and insufficient, limited shelf space for numerous new products introduced on market every day and the like. Consolidation and raising the level of retail sector concentration enables a better negotiating position of the retailers in comparison to producers, lower purchase prices, more efficient and more effective retail business activities, and strengthening of the price competition along with a higher level of services for the consumers.

It is positively sure that in the mentioned processes of strengthening the power of retailers, the significant potential effects of adding new value for consumers are becoming more and more evident. However, changes within the retail sector affect substantial number of stakeholders and they intensively influence the activities of all other market players: producers, wholesalers, retailers, in addition to and state institutions. In this sense, increasing the power of the retailers in marketing

channels intensively opens numerous theoretical and practical questions. Basically, they all lead to the same question, is the retail concentration really in the interest of other members within the marketing channels, and among others – the producers?

That is the question which requires, both in this study and generally, thorough research in order to find concrete answers. The focus is, certainly, on the power of retailers and its usage on other players in marketing channels. In that sense, fundamental questions are if and how producers handle the asymmetric relationships within the modern marketing channels. Can the producers confront the accumulated power of retailers, the pressure retailers put in relation to prices and numerous other elements of their mutual business relationships, as well as to use the advantages of doing business with large retail giants?

Producers' problems mostly manifest in trade of fast moving consumer goods, which means food and other products of everyday purchase and consumption, because the position of retailers is the most dominant in this field. This is the basic reason why the focus of research in this study is the position of producers in this field. Furthermore, in this domain of market, the need for state intervention and harmonization of positions in regard to current problems and open questions of trade policy becomes more evident, which is a very difficult task for countries in transition, such as Serbia. Transition to the modern market, which includes protection of market players' freedom and their free competition within the market up to the level when the market abuses are detected, is not an easy task.

In any case, modern changes in the retail sphere are of high complexity. In order to comprehend those, it is necessary to comprehend the functioning of marketing channels, balance and (inter) dependence among the members of the marketing channel, influencing capability of members who have the power, not only the positive and constructive, but also the negative and destructive sides of power, and finally, the structural changes in the overall economy and market competition, which are the result of increasing the power of retailers in the marketing channels. In accordance with this, thorough research of changes in regard to the power of members and impact to the management and general functioning of marketing channels on every market become highly important. Also, due to the importance for the national market and overall economy development, the position and satisfaction of producers become an important new focus. Bearing in mind this, the special enigma to be solved is the future role of the state in the process of regulating and creating a desirable retail structure.

Shifting the market power from producers to retailers, thus eliciting changes in the marketing channels has been the focus of research since the beginning, of the last decade, of the twentieth century. Simultaneously, with the increase of market concentration and strengthening of the powerful retailers, an intensive academic dispute on how the changes in regard to the power of members in marketing channels reflect on the management of channels, positions and survival of certain

members itself, i.e. general functioning of the marketing channels. Despite the fact that phenomenon of existence of power and demonstration of the power belongs to, by now, the most popular subjects in the field of marketing and trade, these questions continue to draw attention of scientists and experts on a daily basis and with greater care.

Many researchers considered that power was a negative aspect and that those with power would strive to change the behavior of the partner using the strategy of compulsory impact. However, the attitude that power could be useful in case that it created natural distribution of activities and coordination among the members of marketing channels became the prevailing one. The latest research suggest that power in marketing channels could be strategically used, and strategic uses of power could make differences in inter-firm relationships and distribution efficiency [19]. The market players are expected to use power as an effective tool for supply chain management [4], creating the atmosphere of fair relationships and enhancing the efficiency of the overall market. In accordance with it, today the power is broadly accepted, not only in academic circles, but in practice itself by market players. The significance of powerful retailers becomes more evident, and the total theory and practice of marketing changes along with the stated trends.

However, at the same time, academic circles, also, more often raise question whether the increase of gigantic retailers is unambiguously useful to other channel players, and, finally and most important, to consumers. Questioning of the increase of a retailers power and use of that power raises many disagreements and controversies which could be worrisome. Especially in regard to potential limitations and constraints in the vertical dimensions of the channel and jeopardizing the survival of other players in marketing channels, primarily of producers and small and medium-sized retailers, higher prices for consumers, and making barriers for entering the retail market and alike. But, regardless of this, it is doubtless that the dominant attitudes in the works of modern authors are that although there is concern, whatever the problems concerning the concentration of buying power, retailers need to be monitored as long as they compete fairly for market share and pass resulting benefits on to consumers [10]. Also, when this is not the case, then with adequate anti-monopolistic policy and credible and effective anti-monopolistic institutions, one could eliminate those imperfections and deviations on the market which threatens to highly jeopardize its efficiency [17].

Greater power means the possibility of greater influence on its own position and position of its partners, but also means the possibility of showing selfish disregard for others. Unfortunately, there is always a danger of it abusing its own power and the position of weaker partners. The bigger the asymmetry in balance will mean lesser advantages for the weaker side, but there is also the possibility of imposing punishments and enforcement by the more powerful one. However, for many companies, the lure of partnering with a mega-distributor is irresistible [18], and

acceptance of power-imbalance is a key first-step to successful relationship building [11].

In accordance with this, in modern market conditions producers, and especially those with less power, try to, primarily, modify and present their role as a more significant one to those more powerful retailers. Their goal is to become their necessary suppliers, being aware that a sustainable position in modern marketing channels can be established only on the basis of a long-term cooperation and partnership in relationships, characterized by mutual dependency of its members. Be aware that both the level and form of mutual dependency define power relations between the channel members; and, it is also the seed of potential channel conflict [13]. That is the reason why they are vigorously and purposely dedicated to cultivating and managing the development of their long-term mutual relationships. Such long-term relationships develop mutually beneficial outcomes and are characterized by mutual trust, open communication, common goals, commitment to mutual gain, and organizational support [6]. Success in this endeavor, in current conditions of high power concentration and polarization of relationships, will become to producers, especially the weaker ones, preconditions for survival. And, as the economic trends become less predictable the long-term relationships between distribution channels partners becomes an increasingly more important part of the company's long-term strategy [5].

The above mentioned clearly suggests that the traditional theory of exchange now is being analyzed in the light of new relationships established in the modern market conditions between a retail subject and other players in marketing channels, that, positively, the theories of exchange that served marketing well for 40 years are giving way to relational concepts [1]. Besides, analyzed trends make more complex, both in theoretical and methodological sense, the estimation of effects which are the outcomes of increasing the power of retailers in marketing channels.

However, in practice also, more attention is paid to the analysis of the existing relations on the market. Be aware that today a firm's only sustainable advantage is its ability to learn and anticipate market trends faster than competition [12], producers try to understand better numerous concrete situations in which they fail to adequately confront the demands of powerful retail giants. Primarily, during the negotiation on starting and developing partnership between them as suppliers and representatives of procurement department of modern retailers, producers face four key and mandatory aspects [15]:

- sufficient quantities to the whole chain;
- supplier interest in the long-term relationship with the retailer;
- possibility of traceability back to the primary producer'; and
- supplier presence, in one way or another, on the chain's market.

After that, they face the pressure made by the highly powerful buyers who try to minimize the seller's differences in price and decrease transactional value to the maximum in order to, primarily, have space for calculating their own higher margin in price while selling goods in the next phase of distribution. Producers relent under their capability to mercilessly and cruelly lower purchase prices to a low uneconomical level which often is a threat to producer's survival. It is assumed that in this situation suppliers will participate in the supply chain channel as long as his profit is non-negative [9].

In addition, producers, mostly the smaller ones, often complain that retail management favors only certain suppliers and mostly the biggest ones. They are concerned because they believe that category management in this way leads to narrowing categories of products to several brands and that due to this other suppliers are in constant danger of having their products deleted from the assortment. They often see category management as a barrier for entering in the assortment, especially of secondary and tertiary brands, and they are not far from the truth. In accordance with this, today, it is more evident that "category management (CM) has become one of the core areas of interest to both producers and retailers" [7].

Besides, existing of own private label products in retail assortment, based on decisions of retailers` category managers, means that retailers are no longer only agents selling producers' brands: they are now also their competitors [2]. That phenomenon within the analysis of producers' position in conditions of strengthening the retailers cannot be overlooked. Retailers` private label products today highly compete for the shelf space and in these conditions, from producers' point of view, this kind of competition looks unfair. Competing for the shelf space has become so exhausting that many producers perceive it as the most expensive real estate in the world.

And not only for that. Realization of higher profits and greater sales volume at the same time makes it possible for powerful retail chains to become more and more powerful on a daily basis. Their purchasing agents become more skilled negotiators and with a larger horizontal market share will have larger orders to place with the firm's supplies, thereby reducing its invoice costs and boosting its power as a buyer" [16]. By increasing their "upstream" market share simultaneously transferring additional costs to producers and taking additional benefits from them, retailers will increase their horizontal participation and/or profits in this way. This, further, enables them to become more successful in "upstream" competition, which leads retailers, again, to achieve larger horizontal market share and/or profits, and in circles.

2 Objective and Hypothesis of the Research

The subject of this research relates to the effects caused by modern changes in the sphere of retail to the competitive position of producers. The aim is to evaluate the consequences of increasing the power of retailers in comparison to producers, which includes thorough and comprehensive analysis of positive and negative influences on the position and realistic interests of producers. In order to enlighten the mentioned influences, we have conducted research which, also, should confirm or disprove the defined hypothesis:

H1: *Increasing the power of retailers brings a higher level of satisfaction to large producers than to the small and medium-sized producers.*

H2: *Producers, in the conditions of existence of powerful retail chains, express more satisfaction in operating with large retailers than with the small and medium-sized ones.*

H3: *Increasing the power of retailers pushes, primarily, small and medium-sized producers to associate in order to meet the requirements of large retailers for developing long-term interrelations between them and, in this way, to ensure their own survival in marketing channels.*

A detailed survey on the satisfaction of manufacturing firms in Serbia should, in accordance with the defined elements of business relationships, show the level of producers' satisfaction in doing business with other members of marketing channels in modern conditions characterized by increasing the power of retailers. This is especially, focused on cooperation with large and small and medium-sized retailers, separately. The whole picture of that cooperation will be significantly completed by researching the satisfaction indices per individual elements of business relationship of producers and large and small and medium-sized retailers. Besides, producers' concrete responses will contribute to the evaluation of the existing level of their satisfaction in doing business with large and small and medium-sized retailers, as well as understanding of their needs for basic changes.

Results of this research should make clear, primarily, to the creators of trade policy what is the direction and influence of ongoing changes in relationships between members of marketing channels, in conditions characterized by increasing retail power. They should clarify the existence of positive influence that those processes have on trade sector of economy.

The results should enable redirecting the trade policy away from uncritical accusations of modern, large, retailers of creating an unfavorable competitive position of producers in marketing channels, in the situation in which producers mostly lose their positions and influence due to their insufficient pro-activity in terms of adjusting to new competitive conditions on market. The mentioned uncritical approach to acting and influencing of modern powerful retail chains is typical, especially, for the countries in the initial phase of retail market

consolidation. Unfortunately, those countries, due to the lack of knowledge and understanding of modern changes, often resort to state intervention in the market that is directed towards the limitation of development and power growth of retail chains. Also, countries in transition slowly develop modern concepts and contents of trade policy that especially leans toward the domain of anti-monopolistic and, generally, a policy of regulating competition on the market, which can be useful and should be used in cases where powerful retailers do abuse their power.

However, as the countries in transition, such as Serbia, are slowly adapting to creation of a modern market environment, producers are equally slowly undergoing changes. This primarily refers to small and medium-sized producers. It is expected the research will confirm that their position, in the conditions characterized by increasing the power of retailers, is significantly worse than those of large producers. It is assumed that the analysis of satisfaction of small and medium-sized producers with small and medium-sized retailers, as well as the analysis of the certain elements of business relationship between producers and retailers and their concrete responds will lead to a conclusion that the priority strategy for strengthening positions of small and medium-sized producers is their mutual connecting and associating in order to be able to cooperate with, to them highly attractive, large retailers and ensure their survival in marketing channels.

3 Research Methodology

With a view of researching the effects of demonstrating the power of large retailers on producers in the marketing channels within the market of the Republic of Serbia, in accordance with the defined methodology, the concrete research has been conducted, as well as analysis of the results and estimation of the effects of increasing the power of retailers in comparison to the producer's position. The obtained results were basis for giving answers to research hypothesis.

Instead of the (hardly feasible) possibility of measuring precisely the large retailers power and of giving a precise evaluation of the positive and negative effects of its impact on producers, researchers took into consideration the attitudes, i.e. opinions, of a significant number of producers. For the current conditions on the market of Serbia, extensive surveys were conducted which offered a relatively good picture on the significance and effects of increasing the power of retailers on the competitive position of producers in marketing channels.

In order to understand a new position of producers and their relationship with other members in marketing channels, satisfaction of producers in operating with other members in marketing channels was determined by calculating satisfaction indices for each individual relationship. To test defined hypothesis, it is highly relevant to conduct thorough analysis by dividing producers and retailers into large and small and medium-sized, where the criteria of size was a number of

employees (up to 200 employees for small and medium-sized and 200 employees for large ones).

Satisfaction indices of producers were calculated for the following individual business relationships (ij): producers (all) with wholesalers (1a); producers (all) with retailers (1b); producers (all) with small and medium-sized retailers (1c); large producers with wholesalers (2a); large producers with large retailers (2b); large producers with small and medium-sized retailers (2c); small and medium-sized producers with wholesalers (3a); small and medium-sized producers with large retailers (3b); small and medium-sized producers with small and medium-sized retailers (3c).

To calculate the satisfaction indices for each individual relationship, nine key criteria were defined (k) (elements of business relationship), and they were the basis for evaluating cooperation between members, and for each of them the survey respondents specified:

- relevance for cooperation – points from 1 to 10; and
- degree of satisfaction with the existing cooperation (satisfaction rate) – points from 1 to 5.

The following table (Table 1) shows the elements of a business relationship which were used in the questionnaire:

Table 1

Elements of a business relationship – criteria (k)

(k)	**Elements of a business relationship – criteria (k)**
1.	Prices, rebates and additional payments
2.	Terms of payment (due to date) and regularity of payments
3.	Sales potential
4.	Range of assortment
5.	Activities and costs to be borne by a concrete member
6.	Cooperation in terms of promotional and other marketing activities
7.	Data exchange and electronic communication
8.	Level of trust
9.	Potential for development and possibility of a long-term partnership

Additionally, the significance of single elements of business relationship were analyzed, as well. Satisfaction indices per all the elements of a business relationship, for relationships of producers towards large retailers and towards small and medium-sized retailers, were calculated.

Within the obtained data processing, the first step was normalization of satisfaction rate ($V_{ij,k}$) obtained for each element of a business relationship (k) individually, as follows:

$$X_{ij,\,k} = 100/4\ (V_{ij,k} - 1\ ), \tag{1}$$

(i=1, 2 ili 3; j=a, b ili c; k=1, 2,..., 9)

(in case that a producer rated a sales potential of a retailer with 5, then a value obtained by normalization would be 100, with 4 – 75, with 3 – 50, with 2 – 25 and with 1 – 0).

Satisfaction index $X_{ij,}$ which measures the relationship (*ij)*, was calculated as a weighted average of the obtained normalized rates $X_{ij,k}$ by each criteria. Ponder was the obtained rate of relevance for cooperation of certain criteria i.e. rated elements of business relationship (from 1 to 10). Calculating satisfaction index for each separate relationship was performed according to the formulae for weighted average:

$$X_{ij} = \frac{\sum w_{ij,k} X_{ij,k}}{\sum w_{ij,k}}, \qquad (2)$$

where $W_{ij,k}$ stands

for average ponders of a k-criteria of an individual relationship (*ij*).

The following table (Table 2) illustrates procedure for calculating satisfaction index:

Table 2

Illustration of procedure for calculating satisfaction index

Elements of business relationship – criteria (k)	Satisfaction rate by criteria	Normalization of satisfaction rate	Relevance for cooperation by criteria	Result
1	3	50	8	400
2	3	50	8	400
3	3	50	10	500
4	3	50	10	500
5	3	50	8	400
6	4	75	9	675
7	2	25	7	175
8	3	50	9	450
9	3	50	10	500
Total			79	4000

Satisfaction index in the presented case values was 4000/79=50.63. The whole procedure provides value of satisfaction index in the interval from 0 to 100, and the same methodology is also applied for calculating satisfaction index by the elements of business relationship individually (by criteria individually) for relationships of producers with wholesalers, with large retailers and with small and medium-sized retailers.

Methodology, explained in the previous section, was applied for processing the data collected in the field research, on the sample of 30 producers operating within the market of the Republic of Serbia. The research was conducted on a

stratified and purposely (non-random) chosen sample. In the chosen sample, 17 producers with more than 200 employees (big producers) and 13 producers with less than 200 employees (small and medium-sized producers) were interviewed. On average, they co-operate with 210 retailers and 23 wholesalers. All the producers in the field have fast moving consumer goods. In defining the sample, adequate geographic spread, size of undertakings, measured by number of employees and turnover, as well as group of products with the largest share of turnover of a selected producer was taken into consideration. The sampling included around 45% of a turnover of food and other fast moving consumer goods, which is generated in the producer-retailer relation within the Serbian market. The calculated sample error was 17,8%. The basic method used for obtaining data from producers was a direct personal interview (face-to-face) which was guided by a pre-defined questionnaire, during the first quarter of 2013, which was conducted on the basis of previously developed questionnaire. The data collected within the conducted survey research were processed and analyzed using statistical program SPSS 15.0.

4 Research Results

In the questionnaire they filled out, the producers determined, primarily the significance of the certain offered elements of a business relationship with wholesalers and with retailers. After that, they rated the level of current satisfaction they had within the existing business relationships with wholesalers, and, now separately, with large and small and medium-sized retailers, based on the same elements. According to the defined methodology, based on the obtained rates, the average satisfaction indices of producers in operating with different partners in marketing channels were obtained. The results are presented in the following table (Table 3):

Table 3

Average satisfaction indices of producers in operating with other members in marketing channels

Satisfaction of → With ↓	**Producers**	**Large producers**	**Small and medium-sized producers**
Producers			
Wholesalers	62.24	63.36	54.49
Large retailers	63.89	64.62	61.68
Small and medium-sized retailers	55.17	56.40	51.47

In current conditions of reached, relatively significant, level of concentration within the market of the Republic of Serbia, all of the producers, express the highest level of satisfaction in terms of their relationship to large retailers.

The satisfaction is even higher when we observe the business cooperation of large producers with large retailers. This is the highest level of satisfaction of producers in relation to the different business relationship they have with their partners within the marketing channels. Average satisfaction index of small and medium-sized producers with their business relationship with large retailers is markedly lower than the average satisfaction index of large producers.

The obtained data indicate that operating with wholesalers within the market of Serbia brings to the producers, whether we analyze all the producers or individually large and small and medium-sized ones, a somewhat lower level of satisfaction in comparison to operating with retailers. However, this satisfaction is in all cases more significant than the business relationship which producers have with small and medium-sized traders.

Also, if we compare the satisfaction of large producers to the satisfaction of small and medium-sized producers, we can instantly notice the significant differences. Large producers in comparison to small and medium-sized producers are noticeably more satisfied with their business relationship with wholesalers, with large retailers and with small and medium-sized retailers.

One can clearly notice the significant difference of satisfaction of producers in their business relationship with large and small and medium-sized retailers. The obtained average satisfaction indices of producers are presented in the following table (Table 4):

Table 4

Satisfaction of producers with the business relationship with large and small and medium-sized retailers ranked by an average satisfaction index

	Satisfaction		**Index**
1	large producers	large retailers	64.62
2	producers	large retailers	63.89
3	small and medium-sized producers	large retailers	61.68
4	large producers	small and medium-sized retailers	56.40
5	producers	small and medium-sized retailers	55.17
6	small and medium-sized producers	small and medium-sized retailers	51.47

The obtained data indicate that when operating with large retailers, producers show a much higher level of satisfaction in comparison to operating with small and medium-sized retailers. This can be applied not only to the large producers, but to producers as a whole or small and medium-sized ones.

One can easily notice that the biggest difference in expressed satisfaction is related to the satisfaction of small and medium-sized producers with their business relationship with large retailers, on one side, and with small and medium-sized retailers on the other side. Also, it is interesting that satisfaction of small and medium-sized producers with large retailers is significantly higher than the satisfaction of large producers with small and medium-sized retailers. Surely, the lowest satisfaction is expressed by small and medium-sized producers with small and medium-sized retailers.

The collected data enabled conduction of a deeper analysis. In that sense, the results of the research depicted the business relationship of producers with retailers by each element of business relationship, as well. The following table shows the relevance for cooperation of certain elements for the producers in relation to business relationship with retailers and satisfaction indices of producers with large and small and medium-sized retailers on the basis of those elements, with ranking (Table 5).

Table 5
Satisfaction indices of producers by elements of business relationship of producers to large and small and medium-sized retailers

Rank	Elements of business relationship of producers with retailers	Relevance for cooperation	Large retailers		Small and medium-sized retailers	
			Satisfaction index	Rank	Satisfaction index	Rank
1	**Prices, rebates and additional payments**	9.07	54.50	8	68.93	1
2	**Terms of payment (due to date) and regularity of payments**	8.73	58.11	7	56.20	5
3	**Sales potential of retailers**	7.67	81.41	1	57.83	2
4	**Level of trust in retailer**	7.37	60.97	5	52.71	7
5	**Range of assortment of retailer**	7.23	73.04	2	56.91	4
6	**Cooperation in terms of promotional and other marketing activities**	7.20	68.98	3	55.79	6
7	**Activities and costs to be borne by a concrete member**	7.17	50.58	9	52.33	8
8	**Potential for development and possibility of a long-term partnership**	7.00	67.74	4	57.26	3
9	**Data exchange and electronic communication**	5.97	59.64	6	38.55	9
AVERAGE SATISFACTION INDEX			**63.89**		**55.17**	

For producers, the greatest relevance in their relationship with retailers reflects in prices, rebates and additional payments, which includes sale prices and rebates given to retailers, as well as additional payments to retailers, such as payment for "enlisting the products" and alike. It is followed by terms of payment (due to date) and regularity of payments. Elements of business cooperation with the smallest relevance for producers are potential for development and possibility of a long-term partnership and data exchange and electronic communication.

When it comes to prices, rebates and additional payments, the most significant elements for producers, they express a very low level of satisfaction in relationship with large retailers, in comparison to the relationship with small and medium-sized retailers where this element has the highest level of satisfaction and the highest satisfaction index. However, judging by the obtained satisfaction indices, except for prices, rebates and additional payments and the element related to activities and costs to be borne by producer, in all other elements producers express higher level of satisfaction in operating with large retailers compared to operating with small and medium-sized retailers.

Higher level of satisfaction of producers in operating with large retail chains is confirmed by their concrete responses to questions related to their attitudes on operating with large and small and medium-sized retailers. The obtained results are presented in the following table (Table 6).

Table 6

Concrete producers' responds to the questions related to operating with large and small an medium-sized retailers

Questions	**Responses**	
Do you prefer operating with large, modern, retailers or with small and medium-sized, traditional, retailers (who do you rather choose for a partner)?	With large, modern, retailers	Small and medium-sized, traditional, retailers
	86.70%	13.30%
Do you think that development of retail chains is in your and interest of other producers?	I think	I do not think
	70.00%	30.00%

The responses definitely confirm that producers prefer operating with large, modern retailers to small and medium-sized traditional retailers. To be specific, 86.70% of the respondents rather choose for a partner modern, powerful retail chains. 13.30% of the respondents rather choose small and medium-sized traditional retailers. Their responses to the question if they think that development of retail chains is in their interest and interest of other producers, 70% of them responded positively, while 30% of them did not think that this development was in their interest.

Within the scope of questionnaire, the producers were expected to estimate, from 1 to 5, the role of the state in terms of creating modern, incentive, market conditions, for predefined fields. The obtained average rates are presented in the following table (Table 7).

Table 7
Rates for the role of the state in terms of creating modern market conditions given by producers

Rank	Field	Rate
1	**consumer protection policy**	3.30
2	**initiating new (international) trade chains to enter the Serbian market**	2.90
3	**competition policy**	2.90
4	**involving market players in the creation of legislation framework process**	2.80
5	**protection against unfair market competition**	2.43
6	**preventing abuse of market power of large retailers**	2.40
7	**small and medium-sized retailers protection policy**	2.17
8	**initiating domestic retail chains to enter foreign markets**	1.83
	AVERAGE	2.59

The first-ranked field by producers is consumer protection policy. The producers gave noticeably lower rates to the policies which are highly important for protection of their competitive position: competition policy, protection against unfair market competition and preventing abuse of market power of large retailers. On the basis of the obtained rates, the average rate of the role of the state is calculated as (2.59).

5 Testing the Hypotheses

Research has confirmed the first hypothesis that "*increasing the power of retailers brings a higher level of satisfaction to large producers than to small and medium-sized producers".* Average satisfaction indices of producers, in operating with other members in marketing channels: with wholesalers and large and small and medium-sized retailers, indicate that, in current conditions of reached, relatively significant, level of market concentration in Serbia and significant role of large retail chains, producers, generally, but also large and small and medium-sized producers separately, express the highest level of satisfaction in terms of operating with large retailers. This confirms that strengthening large retailers brought a higher level of satisfaction to producers.

Large retailers, whose role on the market becomes stronger every day, ensure the greatest satisfaction to large producers. That satisfaction, in accordance with the results obtained in this research, as far as small and medium-sized producers are concerned is at a remarkably lower level. And, in any case, large producers express more satisfaction, in current conditions, with all the partners in marketing channels: wholesalers, large retailers and small and medium-sized retailers in comparison to small and medium-sized producers. This, without doubt, confirms that the development of large retail chains is in the interest of large producers which are able to cooperate with them and have the opportunity to use the

advantages that this cooperation provides. Thus, their competitive position in marketing channels strengthens and, surely, places small and medium-sized producers in a more inferior position, confirming that, generally, increasing the power of large retailers is not in the interest of small and medium-sized retailers.

Researches have also confirmed the second thesis that "*producers, in the conditions of existence of powerful retail chains, express more satisfaction in operating with large retailers than with small and medium-sized ones.*" Average satisfaction indices clearly indicate that the producers, all of them and, individually, large and small and medium-sized ones, express higher level of satisfaction, when operating with large retailers than with small and medium-sized retailers. The fact that, in conditions characterized by powerful retail chains, producers express more satisfaction in operating with large retailers than with small and medium-sized ones is confirmed, also, by the analysis of their satisfaction indices per individual elements of business relationship with large and small and medium-sized retailers. Out of 9 elements of business relationship of producers with retailers, the satisfaction index is by 7 elements higher when observing relationship of producer with large retailers then with small and medium–sized ones.

The highest level of satisfaction with its business relationship with large retailers is expressed by large producers. This difference in satisfaction is especially noticeable in the case of small and medium-sized producers which express substantial differences in satisfaction with their business relationship with large retailers, on one hand, and on the other hand, with small and medium-sized ones.

Finally, concrete responses the producers gave to the questions related to large and small and medium-sized retailers have definitely and clearly confirmed that producers prefer operating with large, modern, retailers to operating with small and medium-sized, traditional, retailers. A significant majority of them believe that development of large retail chains is in their interest, and in other producers' interest as well. Considering this, the advantage the large retailers have over small and medium-sized retailers is founded on higher level of satisfaction which producers express in operating with large instead of operating with small and medium-sized retailers.

The research confirmed the third thesis that "*increasing the power of retailers pushes, primarily, small and medium-sized producers to associate in order to meet the requirements of large retailers for developing long-term interrelations between them and, in this way, to ensure their own survival in marketing channels*". The research has indicated that large and small and medium-sized producers express a higher level of satisfaction concerning cooperation with large retailers. At the same time, small and medium-sized producers show significant differences in satisfaction related to operating with large retailers in comparison with small and medium-sized retailers. However, in accordance with the structural changes occurring in the retail sector and intensive growth of retail chains, the issue of availability of sufficient quantities for the whole chain of retail shops becomes one of priority criteria for starting cooperation and development of

partnership between producers and modern retailers. Along with the intensive growth of retailers this problem becomes more evident because there is decreasing number of producers which, with their own developing capacities, can cope with highly dynamic development of retailers' demand for products.

The fastest way for a producer to meet the demands for increased quantities is to establish horizontal cooperation with other producers, of same or similar products. Producers are additionally pressured to develop such long term cooperation by growing wishes of modern retailers to establish long term partnership with their suppliers on various segments. Horizontal cooperation and development of the long-term relationships, primarily, with producers, and later with large retailers is exactly the way for producers to achieve a higher level of satisfaction per elements of business relationship with retailers.

The research has proved that producers find that the most important element of a business relationship with retailers is related to prices, rebates and additional payments. Satisfaction index of producers for that element, in a case of relationship with large retailers, which present highly desirable partners, is at a low level. The same situation is with the second important element which relates to terms of payment and regularity of payments or the element related to activities and costs which are borne by producer. Those three elements are related with the lowest satisfaction indices of producers. Horizontal cooperation of producers which leads to strengthening their power and negotiating position is the way to change the current, unsatisfying situation. This in particularly refers to small and medium-sized producers which have limited ability of access and the weakest negotiating positions in relation with large retail chains.

Constantly strong pressure made on producers by large retailers and their purchasing managers is directed at lowering the transactional prices and increasing retail margins. In that sense, producers will be able to meet demands of large retailers if they expand their production capacities and thus secure the effects of economies of scale and further increase the efficiency of their work.

Horizontal cooperation and development of long-term interrelations between producers, with the aim of expanding their capacities so as to satisfy demands of large retailers in terms of prices and quantities, will be of great importance also in the case of manufacturing private label products for retailers. Anyhow, the smaller producers, the greater need for their joining and long-term strategic operating so as to meet the demands of large retailers and ensure long-term cooperation with them as the most preferred partners. That, also, means survival in marketing channels on the modern, every day more concentrated, and global market.

Conclusions

By analyzing the effects of increasing the power of retail chains on competitive position of producers in marketing channels and relationships between large retailers and producers, it becomes evident that demonstrating their significantly increased market and purchasing power causes visible effects on the competition at the level of producer. All, and especially small producers, are strongly affected

by the situation in which they are not able to endure the purchasing power of retailers and are pressured to lower sale prices up to the level which can hardly be survived by any of them. The effects of such a situation are a threat to sustainable growth and development even for those most efficient producers. However, the research results presented in this study indicate that producers, those in position to cooperate with large retail giants, express a high level of satisfaction with this cooperation. Also, those who did not manage to establish and/or sustain such cooperation, are intensively trying to find the way to do so.

In presented framework, numerous questions, both of theoretical and of practical side, are opened. The phenomenon of large retail chains domination intensively reflects on overall market structure and creation of the effective competition between all the members in the marketing channels. This is the reason that almost all the parts of the competitive process and relationships in the marketing channels impose the need for special analysis.

As for the effects of increasing the power of retail chains on competitive position of producers, there are numerous questions to be answered. Space for new research certainly lies in conducting those per products groups, which was not performed in this one due to the insufficient sample volume. Furthermore, it would be interesting to compare satisfaction of producers on similar foreign markets which undergo consolidation processes, as well as on highly developed markets. Also, there is highly emphasized need for researches which refer to tendencies of producers for cooperation at the horizontal and vertical level, as well as for exclusive business arrangements between retailers and producers.

The large enigma that should be solved, is the future role of the state in the process of regulating market and creating retail structures so as to enable effective competition on market. The research results about the role of the Republic of Serbia in terms of creating modern market conditions, which are based on the producers' rates, indicate that we are still far from a satisfying situation. Issues related to competition policy, protection against unfair market competition and preventing abuse of market power of large retailers are certainly open questions in Serbia. The mentioned questions justifiably draw attention of adequate state organizations in almost all the countries of the European Union.

It becomes clear that the fundamental task of the state is to create and manage activities which result in building a modern structure of market and trade. The state should provide conditions for as intensive competition as possible between market players in general. To be more precise, the role and activities of a state should be in interest, and certainly not at the expense of development and market freedom. Furthermore, its role in terms of strong market control is necessary so that competition and market freedoms should not be abused and should not be neglected.

Bearing in mind the stated trends, it is evident that the need for conducting exhaustively thorough research on intensive modern changes in marketing

channels is becoming more emphasized and popular, especially in the sphere of retail and new interrelations established among all market players. Thereto, the open question which requires thorough research in the European Union, as well as in Serbia, relates to the acceptable "tolerance level" of merging large trade chains, taking care of the interests of all members in marketing channels, state and other stakeholders.

References

[1] Achrol, S. R., Kotler, P. (2012) „Frontiers of the Marketing Paradigm in the Third Millennium", *Journal of the Academy of Marketing Science*, No. 40: 35-52

[2] Amrouche, I, Zaccour, G. (2009) „A Shelf-Space-Dependent Wholesale Price when Producer and Retailer Brands Compete", *OR Spectrum*, No. 31: 361-383

[4] Belaya, V., Hanf J. H. (2009) „The Two Sides of Power in Business-to-Business Relationships: Implications for Supply Chain Management", *The Marketing Review*, Vol. 9, No. 4: 361-381

[5] Black, S. G. (2010) „Relationalism: A Vintage But Sound Concept in Distribution Channel Relationships", *Atlantic Economic Journal,* No. 38: 245-246

[6] Bradford, D. K., Weitz, A. B. (2009) „Salespersons' Mangement of Conflict in Buyer – Seller Relationships", *Journal of Personal Selling and Sales Management,* No. XXIX (1, winter): 25-42

[7] Campo, K., Gijsbrechts, E. (2005) „Retail Assortment, Shelf and Stockout Management: Issues, Interplay and Future Challenges", *Applied Stochastic Models in Business and Industry*, No. 21: 383-392

[8] Dobson, P., Waterson, M., Davies, S. (2003) „The Patterns and Implications of Increasing Concentration in European Food Retailing", *Journal of Agricultural Economics*, Vol. 54, No. 1: 111-125

[9] Ertek, G., Griffin, M. P. (2002) „Supplier- and Buyer-driven Channels in a Two-Stage Supply Chain", *II E Transactions*, Vol. 34, No. 2: 691-700

[10] Fels, A. (2009) „The Regulation of Retailing – Lessons for Developing Countries", *Asia Pacific Business Review,* Vol. 15, No. 1: 13-27

[11] Hingley, K. M. (2005) „Power Imbalance in UK Agri-Food Supply Channels: Learning to Live with the Supermarkets", *Journal of Marketing Management,* No: 21: 63-88

[12] Kumar, V., Jones, E., Venkatesan, R., Leone, P. R. (2011) „Is Market Orientation a Source of Sustainable Competitive Advantage or Simply the Cost of Competing?", *Journal of Marketing,* Vol. 75 (January): 16-30

[13] Li, G. Z., Dant, P. R. (2001) „Channel Interdependence: Conceptual and Operational Considerations", *Journal of Marketing Channels*, Vol. 9, No. 1/2: 33-64

[14] Lovreta, S., Končar, J., Stanković, Lj. (2015) „Effects of Increasing the Power of Retail Chains on Competitive Position of Wholesalers" ", *Acta Polytechnica Hungarica*, Vol. 12, No. 3: 213-228

[15] Skytte, H., Blunch, J. N. (2005) „Buying Behavior of Western European Food Retailers", *Journal of Marketing Channels,* Vol. 13, No. 2: 99-129

[16] Steiner, L. R. (2008) „Vertical Competition, Horizontal Competition, and Market Power"*, The Antitrust Bulletin,* Vol. 53, No. 2: 251-270

[17] Stojanović, B., Stanišić, T., Veličković, M. (2010) „The Problem of Competitor Protection in Retail Trade in Serbia", *Škola biznisa,* No. 3: 57-66

[18] Thomas, R. A., Wilkinson, J. T. (2011) "The Devolution of Marketing: Is America's Marketing Model Fighting Hard Enough to Keep Up?", *Marketing Management*, spring: 19-25

[19] Zhuang, G., Herndon, N., Zhou, N. (2006) „Exercises of Power in Marketing Channel Dyads: Power Advantage versus Power Disadvantage", *Int. Rev. of Retail, Distribution and Consumer Research*, Vol. 16, No. 2: 1-22

Decision Making Process of Hexapods in a Model of Complex Terrains

Vladimir Socha[1,3], Patrik Kutilek[1], Alexandr Stefek[2], Lubos Socha[3], Jakub Schlenker[1], Karel Hana[1]

[1] Czech Technical University in Prague, the Faculty of Biomedical Engineering, Nam. Sitna 3105, 272 01, Kladno, Czech Republic, e-mail: vladimir.socha@fbmi.cvut.cz, kutilek@fbmi.cvut.cz, jakub.schlenker@fbmi.cvut.cz, hanakar1@fbmi.cvut.cz

[2] University of Defence, Faculty of Military Technology, Kounicova 65, 662 10, Brno, Czech Republic, e-mail: alexandr.stefek@unob.cz

[3] Technical University of Košice, Faculty of Aeronautics, Rampová 7, 041 21, Košice, Slovakia, e-mail: lubos.socha@tuke.sk

Abstract: The paper describes behavior of a cognitive control system model, which enables a hexapod to walk in an obstacle-free terrain as well as in a complex terrain including obstacles. This cognitive system model is based on reinforcement learning and assumes the concept of static-stable walking. The decision making process was tested using three different types of terrain models. The results of decision making process trigger actions in the form of changes in the state of six-legged body to maintain stable walking forward. New method have been developed to describe a group of obstacles of different sizes in a complex terrain. The results suggest a relationship between the predefined number of actions and the maximum total walked distance in terrain. In case of the terrain without obstacles, the optimized actions are the same. Thus, the way of moving the trunk and legs in the terrain is always the same and cyclic. The results also indicate that the maximum total walked distance is reduced due to a growing number of obstacles to overcome. The maximum total walked distance is reduced more significantly in the case of overcoming a greater number of small obstacles compared the case of smaller number of large obstacles. The way of moving the trunk and legs in the terrain with large obstacles is acyclic. The methods proposed for the study of the cognitive system and the sensory system of a hexapod, for the simulation of six-legged walking, as well as for the characterization of terrain with obstacles may find application in bioengineering, robotics, military system and other fields.

Keywords: complex terrain; obstacles; hexapod; reinforcement learning; static-stable walking

1 Introduction

Hexapod represents a six-legged arthropod with a set of three pairs of legs controlled by a nervous system, [1] while relying on the concept of six-legged undercarriage previously introduced in robotics. The model of the cognitive system used for modelling of a six-legged insect`s gait [2] employed a set of sensors (eyes) and actuators (leg muscles), [1, 2]. Generally, the cognitive system sends information to the motoric system and receives information from the sensory system, which is a concept known as the sensor-actuator loop [2]. The central cognitive system is responsible for the strategy of leg coordination and the transition from current-state to future-state [2], see Fig. 1.

The process of learning and decision-making in a cognitive system can be modelled using computational methods [2-4]. A number of methods based on artificial intelligence has been developed to control the trajectory of a walking hexapod robot [5-7], [8] however, these have not been designed for leg coordination [2]. There are also methods for controlling legs on a flat surface without any obstacles [9-13]. The methods do not offer solutions for crossing of small obstacles in the terrain [2] and only a few of the methods have been proposed to allow for this technique of walking [2, 14]. These methods are based i.e. on neural networks [15-16] or genetic algorithms [17]. Only a small number of methods is based on Reinforcement Learning (RL) [2, 3, 9, 10], which is accepted as a method that describes the decision-making process of living organisms [18-21]. The results of decision making process based on RL trigger in the form of changes in state of six-legged body to maintain stable forward walking [20]. An issue in evaluating the effectiveness of methods based on RL, however, is that the total walking distance in terrain covered by the number of actions of legs is affected not only by the algorithm parameter settings but the terrain complexity as well. This shortcoming thus calls for a method which would allow for the terrain complexity description.

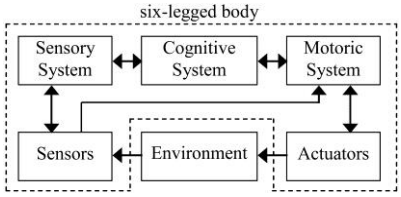


Figure 1
The sensor-actuator loop

A large number of methods focusing on the complexity of terrain has been designed in the field of robotic studies, and some of these are based on the description of rough terrain. Although procedures for planning foot trajectory and body trajectory of walking robots in 3D space have already been designed [22-25],

no method or variable to describe a group of obstacles of different sizes in a complex terrain corresponding to the walked distance or number of actions has been developed yet. In contrast with methods describing obstacles of different heights corresponding to the dimensions of the robot's trunk already introduced [26], a method of assessing lengths (horizontal sizes) of obstacles, however, has never been described comprehensively. A terrain analysis based on the envelope, slope and the curvature of the surface using certain scale has also been designed [27], however failed to provide characteristics of the geometry and size of obstacles in relation to the trunk and/or leg characteristics in terms of size and workspace.

Thus, the aim of this article is to introduce a method of describing a complex terrain and analyse the applicability of the method based on RL to enable stepping over obstacles in a terrain. The methods for the study of the six-legged walking, as well as for the characterization of terrain with obstacles could find application in robotics, military systems, rehabilitation and other fields.

2 Methods

2.1 Model of the Decision-Making Process of Hexapod

Walking through a complex terrain with obstacles is made possible through acyclic gait. The assumed model of cognitive system uses decision-making process based on RL for coordination of the legs of hexapod [2, 20, 21]. To ensure static stable locomotion [20, 21], the cognitive system uses the known states of the position of the legs to maintain static stability [13]. For the decision-making, knowledge of the condition for static-stable posture is assumed [20, 21]. The static-stable walking is represented by vector $t=(l_{R1}, l_{L1}, l_{R2}, l_{L2}, l_{R3}, l_{L3})$, $t \in T$, which describes the 15 states of six-legged body to maintain static stability [2], where binary (i.e. true/false) variables ($l_{R1}, \ldots, l_{L3}$) represent the states of left (L) and right (R) legs: 1 - leg is in the swing phase; 0 - stance phase of the leg, [2, 20, 21].

To transport the trunk, the legs change their position in relation to the trunk. If the movement of each leg is autonomous, we can describe the position of each leg in leg workspace (LW) by the value n_i, [2]. This value represents the requirement for leg movement in LW, [23]. Maximum front position in LW is represented by the value of 0. Value 1 represents maximum requirement for leg movement of the leg in back position, [2]. Vector $r = (n_{R1}, n_{L1}, n_{R2}, n_{L2}, n_{R3}, n_{L3})$, $r \in R$, represents requirements for all legs movements.

Normally, the feet must not touch the obstacles, [22]. If we know where the obstacles are located in LW, it is possible to identify possible step lengths within the LW. Vector $p = (k_{R1}, k_{L1}, k_{R2}, k_{L2}, k_{R3}, k_{L3})$, $p \in P$, represents information about maximum possible step lengths k_i of all legs, [2].

The designed cognitive system uses Q-learning (QL) as a RL technique, [2, 20, 21], see Fig. 2. In the QL, the state-action pairs are represented by a Q-table, [2]. The Q-table stores the information about the relations between the states p, r and proper actions t. The actions represent changes in the state of the six-legged body to maintain static stable walking forward. States for each leg are represented by variables $n_i \in N$ and $k_i \in K$, and the actions of each leg are represented by $l_i \in L$ variables, [2].

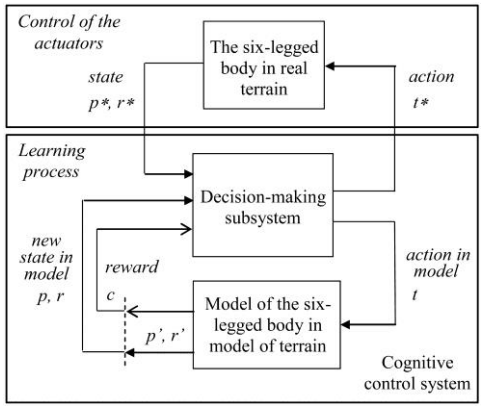


Figure 2

Cognitive control system based on reinforcement learning, six-legged body and terrain model [2]

Based on the information stored in the Q-table, the current situation is evaluated in order to select the best (most reward-promising) action to accomplish the task, [2, 21]. The new state, entered as a consequence of the execution of action, is evaluated by the reinforcement function. Its qualitative criterion (reinforcement) is used by the algorithm that adjusts the Q values, [2, 18, 19]. The RL algorithm applied is described in Table 1, [2]. The learning algorithm searches for possible options of walking and evaluates each action $t \in T$. The p^* and r^* are the initial states and the initial positions of the six-legged body in a terrain model. Let $MaxQ(p',r',t)$ be the maximum Q value of the next state (p',r') related to all the possible actions $t \in T$, [2], Table 1. For each new state of the six-legged body, different possible actions can be tested n times. The range of distance $l_{required}$ for the prediction of t actions in the terrain model is determined by the capabilities of the sensors. The variable for determining the (reached) total walking distance l_{walked}, i.e. maximum total walking distance after the learning process, is a predefined m number of t actions to cover this total distance. The value of c is the immediate (or expected) reward for the state change from the old state (p,r) to the

new state (p',r'). We assume that the success of the action t is defined by the covered distance represented by l, thus the reward is $c \approx l$, $0 \leq l \leq l_{max}$ [2]. The learning rate $\alpha(p,r,t)$ (see Table 1) is used to ensure the convergence of the iterative procedure, [2, 19]:

$$\alpha = \frac{\alpha_0 n_0}{n_0 + n_\alpha(p,r,t)} \tag{1}$$

$n_\alpha(p,r,t)$ is the number of times the $Q(p,r,t)$ value, i.e. state-action pair (p,r,t), has been called during learning, [20]. The α_0 and n_0 are parameters used to control the convergence of the iterative procedure. Based on previous experiments, we set $\alpha_0=1$ and $n_0=1$. The γ is the discount factor in the range $0 \leq \gamma < 1$, we set $\gamma=0.9$ [20].

Table 1

Algorithm for RL of the hexapod cognitive control system [2, 21]

1.	Set the initial parameters of the learning process
2.	Identify the state $p = p^* \in P$ and state $r = r^* \in R$ of the six-legged body in real terrain
3.	Select and execute an $t \in T$ action of the model of six-legged body in the model of terrain
4.	Identify the new state $p' \in P$ and $r' \in R$ of the model of six-legged body, and obtain the immediate reward c and update the learning rate α
5.	Update the Q-table, i.e. Q value: $Q(p,r,t) = Q(p,r,t) + \alpha (c + \gamma Max\, Q(p',r',t) - Q(p,r,t))$
6.	Assign $p = p'$ and $r = r'$
7.	While $n \neq n_{repetitions}$ return to 3
8.	While $l_{walked} \neq l_{required}$ return to 2
9.	Execute the most appropriate action $t^* = t \in T$ of the six-legged body in real terrain

The strategy used to select the action during learning is based on Boltzmann´s exploration [2, 20, 28]:

$$P(t) = \frac{e^{\frac{Q(p,r,t)}{E}}}{\sum_{t \in T} e^{\frac{Q(p,r,t)}{E}}} \tag{2}$$

where E is a parameter known as the computational temperature. High temperatures cause all actions to be nearly equiprobable, whereas low temperatures cause greedy action selections. The parameter value E decreases gradually by:

$$E_{n+1} = E_{\min} + \beta \cdot (E_n - E_{\min}) \tag{3}$$

where n is the number of the cycle repetition, i.e. the number of iteration cycles. Based on previous experiments and recommendations [2, 20, 28], the values of the parameter are $E_0=E_{max}=0.9$, $E_{min}=0$ a $\beta=0.9$. RL parameters and their descriptions can be found in [2, 20].

Table 1 presents the algorithm used to explore the possible states and actions of a six-legged body in a modelled terrain with obstacles. The cognitive system tries to cover a required maximum distance $l_{required}$, and thus the number of iterations is determined by the size of the terrain model. The number of $n_{repetitions}>0$ if (repetitions) selecting an action t is given by the quality of the cognitive abilities of the system (we use n=1). The learning process is stopped when the required maximum distance $l_{required}$ or the number of cycles (i.e. locomotion) to walk through the same terrain model by an m number of t actions is achieved. Of course, the distance $l_{required}$ can be reached by a certain number of cycles to walk through the entire length of the terrain during learning. After learning, the final, optimized (t) actions at the (p, r) states are performed [2]. It is assumed that the total walked distance l_{walked} as well as the m number of t actions will decrease in a more complex terrain and increase in a less complex one. In the following sections the focus will be put on the study of the relationship between the number of cycles required to walk through the entire length of the terrain to achieve the maximum distance, number m of t actions to cover the maximum distance and the terrain complexity.

2.2 Method for Terrain Description and Gait Evaluation

The second aim of article is to introduce a method for describing a complex terrain and analyse the applicability of the method based on RL to enable stepping over obstacles in a terrain. However, a method for assessing the obstacles of different lengths has never been described comprehensively before. We assume that the cognitive system should select the best walking strategy to overcome the longest (i.e. maximum predefined) distance. Also, we assume that obstacles in a terrain are already identified, e.g. using methods based on recursive density estimation and evolving Takagi–Sugeno fuzzy systems [29] or 2D laser range finder and obstacle/gap detection based on edge detection [30]. An example of a terrain model with complicated distribution of obstacles is shown in Fig. 3 exported from the simulation software (MatLab, MathWorks Inc.) including descriptions.

To verify the designed RL based control system, it is necessary to describe lengths (horizontal sizes, [30]) of the obstacles in the direction of movement to consider only those obstacles which are on the path of the hexapod and corresponding with the section of the complex terrain, i.e. the covered total walking distance l_{walked}. Two types of obstacles need to be distinguished: small obstacles which can be stepped over by one leg without translating the trunk and large obstacles which have to be overcome by translating the trunk and can't be stepped over by a single leg movement.

The maximum possible step length d_{kSmax} in a terrain without obstacles is defined by the LW geometry [2, 17, 30]. The lengths of small obstacles are smaller or equal to the length of the maximum possible step h d_{kSmax}, Fig. 3, [2]. The length of large obstacles is greater than the maximum possible step length d_{kSmax}, but still

has to be smaller than the maximum walking distance by one action t to ensure movement of the trunk in a terrain, i.e. l_{max}, see Fig. 3. Theoretically, l_{max} can reach up to $2 \cdot d_{kSmax}$, assuming the static-stable locomotion and geometry of the trunk and LW as described in [2, 21]. It also has to be ensured that two obstacles larger than maximum possible step length d_{kSmax} are not situated in both workspaces of the left and right leg at the same time, i.e. in the workspaces of the two opposite legs. If the condition above is met, then the path corresponding to the section of the complex terrain with obstacles, i.e. reached total walking distance l_{walked}, can be described by a numeral denoting the number of small obstacles and a numeral informing about the number of large obstacles, see Fig. 3. A more objective description would be the following: The path corresponding to the section of the complex terrain with obstacles, i.e. covered total walking distance l_{walked}, is described by numerals that represent the number of small obstacles to be overcome by the left side (LS) of the body (left legs), the number of the large obstacles to be overcome by the LS of the body (left legs), as well as the number of the small obstacles to be overcome by the right side (RS) of the body (right legs), and finally the number of large obstacles to be overcome by the RS of the body (right legs).

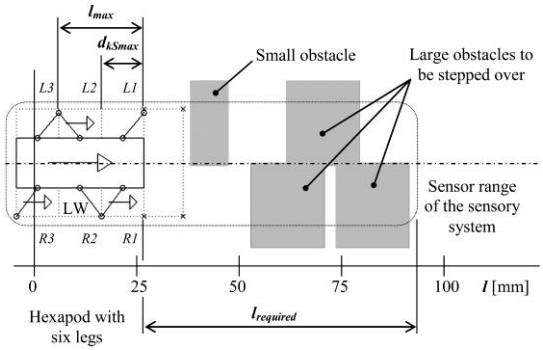


Figure 3
Simplified model of a six-legged body (with legs in swing/stance phase) and sensory information in a terrain model with obstacles [2]

The remaining question is what kind of relationship, if any, is there between the covered total walked distance l_{walked}, the predefined m number of t actions and the numbers of cycles (to reach the maximum total walked distance during learning) to walk over the terrain with specific types and numbers of obstacles. The information about possible relationship could be used to develop algorithms for effective RL, thus reducing the time and complexity of computing the learning process.

2.3 Verification Method

The verification is focused on testing the adopted approach to evaluate the hexapod's decisions in complex terrain. For this purpose, the second aim of article is to introduce a method of describing a complex terrain and analyse the applicability of the method based on RL to enable stepping over obstacles in a terrain. However, a method for assessing obstacles of different lengths has never been described comprehensively before. We assume that the cognitive system of hexapod should select the best walking strategy (i.e. the most appropriate actions t^*) to overcome the longest (i.e. maximum predefined) distance.

Table 2

Relationship between the predefined number of actions and the used types of complex terrains with small and large obstacles (LS – left side, RS – right side of the hexapod's body)

Terrain	Type of obstacles	Predefined number m of actions				
		3	6	9	12	15
		Number of obstacles to be overcome				
1.	small LS	0	0	0	0	0
	large LS	0	0	0	0	0
	small RS	0	0	0	0	0
	large RS	0	0	0	0	0
2.	small LS	2	3	4	5	6
	large LS	0	0	0	0	0
	small RS	0	2	3	4	4
	large RS	0	0	0	0	0
3.	small LS	2	3	3	3	4
	large LS	0	0	1	1	1
	small RS	0	1	1	1	2
	large RS	0	0	1	2	2

The method for the decision making based on RL described above was tested on models of terrain based on method for describing complex terrains in a MatLab software environment. The terrain model with obstacles is (on purpose) designed to be complex and ill-structured. It is assumed that the terrain model is to be explored and identified by the sensory system. The initial parameters used for the learning algorithm define the initial position of the six-legged body in the terrain model and the characteristics of terrain obstacles (dimensions and positions), see Fig. 3. Then the maximum distance $l_{required}$=140 mm from the six-legged body (i.e. trunk) to the farthest detected obstacle or target in the terrain is chosen. During the prediction of walking, the maximum number m of predicted t^* actions of the six-legged body in the model of terrain represents another limiting factor, see Table 2.

The issue with the evaluation of the effectiveness of setting algorithm parameters, however, is that the covered total walked distance l_{walked} achieved by the predefined m number of t actions is affected not only by the parameters set for the algorithm, but also by terrain complexity. Consequently, a more advanced testing session was performed using three types of terrains, Table 2.

The first type of terrain was a flat terrain without any obstacles; the second type of terrain involved only small obstacles; and the third type of terrain was covered with both small and large obstacles. To walk through the three types of terrains, the predefined m numbers of t actions were 3, 6, 9, 12 and 15. The predefined number of cycles to walk through the same terrain during learning was 2500.

2.4 Statistical Analysis

Each type of terrain was tested ten times by each predefined m number of actions. After calculating the covered total walked distance l_{walked}, the predefined m number of actions and the number of cycles to walk through the same terrain, the statistical analysis of these characteristics was performed using MatLab software. Maximum and minimum values were identified, the median, first quartile (Q1) and third quartile (Q3) was calculated for the number of cycles (i.e. locomotions) to reach the maximum total walking distance (i.e. select the optimized actions). The Jarque–Bera test (in MatLab software) was used to test the normal distribution of all parameters. The test returns the value of h=1 if it rejects the null hypothesis at the level of significance of 5%, and h=0 if does not.

3 Results

In this section, results for approach adopted to evaluate the hexapod's decisions by RL in three different complex terrains (see Table 2) are demonstrated. Each type of terrain was tested ten times by each predefined m number of actions and the Jarque–Bera test returned h=1 in all trials. The data were compared to identify the relationship between the predefined m number of t^* actions and the maximum total walking distance (l_{walked}), see Table 3 and Fig. 4. Using the RL method, the hexapod (i.e. model of the hexapod in MatLab software environment) is able to plan several actions in advance, see Table 3. The planned actions represent the planned changes in the state of the six-legged body to maintain static stable walking forward. The maximum total walking distance is achieved by the predefined number (m) of actions (t^*) after the learning process.

Table 3

Maximum total walked distances achieved by the predefined number of actions after the learning

Terrain	Predefined *m* number of actions *t**				
	3	6	9	12	15
	Maximum total walked distance [mm]				
1.	24	60	96	132	168
2.	24	49	64	79	94
3.	24	49	66	88	114

To determine the relationship between the number of cycles (i.e. locomotions) to reach the maximum total walking distance and the predefined *m* number of *t** actions, compare Table 4 and Fig. 5.

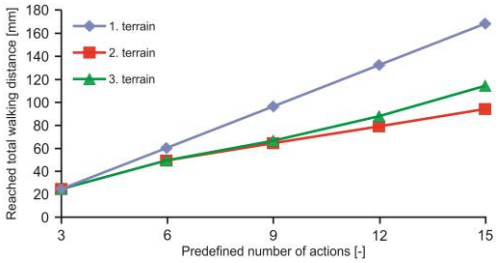


Figure 4

Diagram illustrating relationships between the maximum total walking distances and predefined numbers of actions in different terrains

Table 4

Numbers of cycles to reach the maximum total walking distance and to select the optimized actions in specific terrains

		Numbers of cycles to reach the maximum total walking distance														
		1st terrain					2nd terrain					3rd terrain				
		Min	Max	Median	Q1	Q3	Min	Max	Median	Q1	Q3	Min	Max	Median	Q1	Q3
Predefined number *m* of actions	3	3	2037	45	12	59	26	1199	155	63	259	7	617	167	90	456
	6	76	2214	205	157	313	147	1686	331	239	696	108	1822	310	177	1294
	9	74	1268	211	118	730	169	2227	564	348	1109	223	1715	519	343	778
	12	59	1342	139	113	390	156	2373	485	267	1882	313	1633	1016	886	1539
	15	79	741	150	112	231	435	2025	938	816	1347	404	2058	1381	798	1715

It is clear that a higher value of the defined *m* number of actions increases the number of obstacles to be overcome, Table 2. Furthermore, the total walking distance is also increased, Table 3. When traversing the three different types of terrains, all obstacles (both small and large) were overcome.

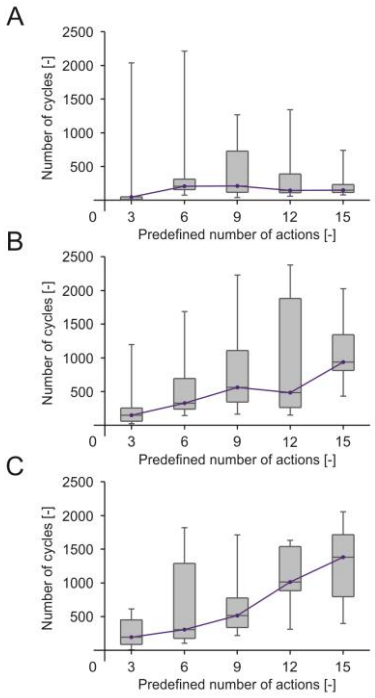


Figure 5

Diagram illustrating relationship between the number of cycles to reach the maximum total walking distance and the predefined numbers of actions in 1st terrain (A), 2nd terrain (B) and 3rd terrain (C)

4 Discussion

The data presented in Fig. 5B and Fig. 5C point out to the relationship between the predefined number of actions and the number of cycles (i.e. locomotions) to reach the maximum total walking distance and indicate the increase of the median of the number of cycles to reach the maximum total walking distance, helping to select the optimized actions, in the case of the 2nd and 3rd terrain. The data presented in Fig. 5A indicate a gradual increase and a subsequent decrease of the median of the number of cycles to reach the maximum total walking distance, and help select optimized actions, in the case of the 1st terrain (without obstacles). This is due to

the fact that the identified states (p, r) of the six-legged body and terrain are the same or similar during walking, and thus the optimized actions are the same and computation time of the learning process to find the optimized actions can be reduced. In all cases, the Jarque–Bera test confirmed that we can reject the hypothesis that the number of cycles to reach the maximum total walking distance has a normal distribution. In the case of the 1st and the 2nd terrain (including only small obstacles), strong asymmetrical (skewed) distributions of the number of cycles to reach the maximum total walked distance is found. The median of the number of cycles to reach the maximum total walking distance during learning is much lower than the maximum value, see Table 4. Thus, the maximum total walking distance is achieved by a lower number of the cycles. In the case of the 3rd terrain with large obstacles, the maximum total walking distance is achieved by a higher number of the cycles.

The data also indicate that the increase in the number of cycles to reach the maximum total walked distance is hampered (i.e. reduced) as the identified states (p, r) of the six-legged body and terrain are similar during walking in the case of the 1st and the 2nd terrain (including only small obstacles) and predefined greater number of actions. In the case of the 3rd terrain (with small and large obstacles), the rate of increase in the number of cycles is not constant or reduced since the terrain is complex and the identified states (p, r) of the six-legged body and terrain are not identical during walking. In general, moving over a more complex terrain reduces the number of actions and results in higher values of maximum number of cycles (i.e. locomotions). On the other hand, the maximum total walking distance is lower as terrain complexity reduces the total number of possible actions denoted by t.

In view of the adjustment of learning process described above, the approximate number of cycles (i.e. locomotions) sufficient for the subsequent selection of the optimal action t can reach up to 1500. The approximate number of cycles can be lower than 1000 in the case of a less complex terrain and the predefined m=9 number of t actions. With a more complex terrain, the approximate number of cycles can be lower than 1000 in the case of a lower predefined (m=3) number of t actions. In general, the use of a lower predefined (m=3) number of actions is appropriate for a complex terrain.

The data suggesting the relationship between the predefined number of actions and the maximum total walking distance, Fig. 4, indicate an increase in the maximum total walking distance in all cases, i.e. the terrains are passable. In case of the 1st terrain (without obstacles), the curve of the increase in the maximum total walking distance is constant because the optimized t actions are the same. Thus, the type of moving the trunk and legs in the terrain is always the same. For static stable walking forward, the model of hexapod can use acyclic and/or cyclic gait. The type of gait depends on the results of the search for the most appropriate sequence of actions to achieve the maximum total walked distance by the predefined number of actions. In case of the 1st terrain, model of the hexapod

opted for cyclic gait, as is apparent from Fig. 4, where the walked distance after the individual actions remains the same. The decision making process of the hexapod shows that acyclic gait suits a very complex terrain and cyclic gait on the other hand a less complex terrain. The data in Fig. 4 also indicate that the maximum total walking distance is reduced due to a growing number of obstacles to be overcome (Tab. 2), especially in the 2nd terrain. The maximum total walking distance is reduced more significantly in the case of overcoming a greater number of small obstacles (2nd terrain) than in the case of smaller number of large obstacles (3rd terrain).

Proceeding from the above analysis, it was found that the results show the suitability of the adopted approach. The methods proposed for the study of the six-legged walking, as well as for describing a terrain with obstacles, allow us to quantitatively assess and evaluate the hexapod's decisions in complex terrain. It of course follows that, alternatively, other nonlinear methods adapted for hexapod's decisions in complex terrain could be used, e.g. the tensor product models [31], methods for delivery vehicle routing problem [32], nonlinear multivariable systems using recurrent cerebellar models [33], neural control mechanisms [16], etc. Application of these methods to the problem of hexapod's decisions may find its place in the context of follow-up studies. However, RL method was used as RL is widely accepted as a tool to understand the goal-directed behavior of real organisms (including six-legged insects) that learn and interact with their environment in real time. Thus the objective of RL is to select actions so as to maximize their long-term rewards [34]. Regarding the methods to describe the complexity of the terrain, no similar method to describe the complexity of terrain with respect to the hexapod geometry has been introduced before. In the past, methods for describing obstacles in a terrain were only mentioned [30]. Not only does our approach allow for the description of the complexity of a terrain, it also provides description of the size of the obstacle with respect to LW [2, 23]. Thus, the proposed method may complement existing methods designed for trajectory planning [5-8].

Conclusions

The proposed techniques for describing complexity of a terrain and hexapod's decisions in the terrain were described tested, and verified in the article. In order to meet the demands for quantitative description of the terrain, the method describing lengths (horizontal sizes) of the obstacles in the direction of the hexapod´s movement was designed. The technique for decision making process of hexapod based on RL selects the most suitable action for each state to overcome the longest (i.e. maximum) distance in the terrain which includes obstacles. The methods for the study of the decision making process of hexapods, as well as for the simulation of six-legged walking as well as describing terrains with obstacles may find application in bioengineering, robotics, military systems and other related fields.

In future works, the problem of vertical size of obstacles in terrain and the problem of accelerated motion in terms of its dynamics will be studied. The proposed methods rely on the movement of most insects since they move slowly, and the problem of vertical size of obstacles has been partially solved in [22]. However, vertical movement of the legs is determined exclusively by the height of the obstacles in direction of walking [30] and does not affect the actions (i.e. the strategy of leg coordination of hexapod and the transition from current-state to future-state) determined by the methods described in this article.

Acknowledgement

This work was done in the framework of CTU project SGS15/107/OHK4/1T/17. The authors would like to thank Andrej Madoran, BA, for the translation of this work.

References

[1] Barfoot T., Earon E., D'Eleuterio G.: Experiments in Learning Distributed Control for a Hexapod Robot, Robotics and Autonomous Systems, Vol. 54, 2006, 864-872

[2] Barfoot T., Earon E., D'Eleuterio G.: A Step in the Right Direction – Learning Hexapod Gaits through Reinforcement, Proceedings of the International Symposium on Robotics (ISR), Quebec, Montreal, 2000; pp. 487-492

[3] Parker G.B., Mills J.W.: Metachronal Wave Gait Generation for Hexapod Robots, Proceedings of the World Automation Congress (WAC), USA, Anchorage, 1998; pp. 365-370

[4] Youcef Z., Pierre C.: Control of the Trajectory of a Hexapod Robot based on Distributed Q-learning, Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE), France, Ajaccio, 2004; pp. 277-282

[5] Touzet C.: Q-learning for Robots, The Handbook of Brain Theory and Neural Networks, Cambridge: MIT Press, 2003; pp. 934-937

[6] Khriji L., Touati F., Benhmed K., Al-Yahmedi A.: Mobile Robot Navigation based on Q-Learning Technique, International Journal of Advanced Robotic Systems, 2011; Vol. 8, pp. 45-51

[7] Porta J., Celaya E.: Efficient Gait Generation using Reinforcement learning, Proceedings of the 4th International Conference on Climbing and Walking Robots (CLAWAR), Germany, Karlsruhe, 2001; pp. 411-418

[8] Zeidan B.; Sakyasingha Dasgupta S., Wörgötter F., Manoonpong P.: Adaptive Landmark-based Navigation System Using Learning Techniques, Lecture Notes in Computer Science, 8575: 121-131

[9] Porta J.: Rho-LEARNING: A Robotics-oriented Reinforcement Learning Algorithm, Technical Report IRI-DT-00-03, Institut de Robotica i Informatica Industrial, Barcelona, 2000

[10] Espenschied K.S., Quinn R.D., Chiel H.J., Beer R.D.: Leg Coordination Mechanisms in Stick Insect Applied to Hexapod Robot Locomotion, Adaptive Behaviour, Vol. 1, 1993, pp. 455-468

[11] Porta J., Celaya E.: Walking in Unstructured Natural Environments, Proceedings of the European Workshop on Hazardous Robotics (HEROS), Spain, Barcelona, 1996; pp. 99-107

[12] Aparna K., Geeta S.: Insect Inspired Hexapod Robot for Terrain Navigation, Journal of Research in Engineering and Technology, Vol. 2, 2013, pp. 63-69

[13] Tedeschi F., Carbone G.: Design Issues for Hexapod Walking Robots, Robotics, Vol. 3, No. 2, 2014, pp. 181-206

[14] Juang C., Chang Y., Hsiao C.: Evolving Gaits of a Hexapod Robot by Recurrent Neural Networks with Symbiotic Species-based Particle Swarm Optimization, IEEE Transactions on Industrial Electronics, 2011; 58: 3110-3119

[15] Belter D., Skrzypczynski P.: A Biologically Inspired Approach to Feasible Gait Learning for a Hexapod Robot, International Journal of Applied Mathematics and Computer Science, 2010; Vol. 20, pp. 69-84

[16] Goldschmidt D., Wörgötter F., Manoonpong P.: Biologically-inspired Adaptive Obstacle Negotiation Behavior of Hexapod Robots, Frontiers in Neurorobotics, Vol. 8, 2014, pp 1-16

[17] Irodova M., Sloan R.: Reinforcement Learning and Function Approximation, Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 05), USA, Clearwater Beach, 2005; pp. 455-460

[18] Sutton R.S., Barto A.G.: Reinforcement Learning: An Introduction, Cambridge: MIT Press, 1998

[19] Kutilek P., Kacer J.: The Locomotion Control of the Concyclically Walking Carriage, Cybernetics Letters, Vol. 3, No. 1, 2005, p. 8

[20] Hrdlicka I., Kutilek P.: Reinforcement Learning in Control Systems for Walking Hexapod Robots, Cybernetics Letters, Vol. 3, No. 1, 2005, p. 13

[21] Belter D., Skrzypczynski P.: Integrated Motion Planning for a Hexapod Robot Walking on Rough Terrain, Proceedings of the 18th World Congress of the International Federation of Automatic Control (IFAC 2011), Italy, Milano, 2011; pp. 6918-6923

[22] Hauser K., Bretl T., Latombe J. C., Harada K., Wilcox B.: Motion Planning for Legged Robots on Varied Terrain, International Journal of Robotics Research, Vol. 27, 2008, pp. 1325-1349

[23] Görner M., Chilian A., Hirschmüller H.: Towards an Autonomous Walking Robot for Planetary Surfaces, Proceedings of the 10th International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS 2010), Japan, Sapporo, 2010; pp. 170-177

[24] Rebula J. R., Neuhaus P. D., Bonnlander B. V., Johnson M. J., Pratt J. E.: A Controller for the LittleDog Quadruped Walking on Rough Terrain, Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA 2007), Italy, Rome, 2007; pp. 1467-1473

[25] Palis F., Rusin V., Schmucker U., Schneider A., Zavgorodniy Y.: Walking Robot with Articulated Body and Force Controlled Legs, Proceedings of the Research on Adaptive Motion in Animals and Machines (AMAM 2005), Germany, Ilmenau, 2005; pp. 1-6

[26] Pettersson L.: Terrain Analysis as a Design Tool for Autonomous Vehicles in Difficult Terrain, Proceedings of the Second NordDesign, Sweden, Stockholm, 1998; pp. 1-10

[27] Celaya E., Porta J.: Force-based Control of a Six-legged Robot on Abrupt Terrain using the Subsumption Architecture, Proceedings of the International Conference on Advanced Robotics (ICAR '95), Spain, Sant Feliu de Guixols, 1995; pp. 413-419

[28] Kianercy A., Galstyan A.: Dynamics of Boltzmann Q-learning in Two-Player Two-Action Games, Physical Review E, Vol. 85, No. 4, 2012, pp. 1-10

[29] Angelov P., Sadeghi-Tehran P., Ramezani R.: An Approach to Automatic Real-Time Novelty Detection, Object Identification, and Tracking in Video Streams based on Recursive Density Estimation and Evolving Takagi-Sugeno Fuzzy Systems, International Journal of Intelligent Systems, Vol. 26, No. 3, 2011, pp. 189-205

[30] Kesper P., Grinke E., Hesse F., Wörgötter F., Manoonpong P.: Obstacle/Gap Detection and Terrain Classification of Walking Robots based on a 2D Laser Range Finder, Proceedings of the 16th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines (CLAWAR 2013), Australia, Sydney, 2013, pp. 419-426

[31] Precup R., Dragos C., Preitl S., Radac M., Petriu E.: Novel Tensor Product Models for Automatic Transmission System Control, IEEE Systems Journal, Vol. 6, No. 3, 2012, pp. 488-498

[32] Khmelev A., Kochetov Y.: A Hybrid Local Search for the Split Delivery Vehicle Routing Problem, International Journal of Artificial Intelligence, Vol. 13, No. 1, 2015, pp. 147-164

[33] Chen C. H., Chung C. C., Chao F., Rudas I. J., Intelligent Robust Control for Uncertain Nonlinear Multivariable Systems using Recurrent Cerebellar Model Neural Networks, Acta Polytechnica Hungarica, Vol. 12, No. 5, 2015, pp. 7-33

[34] Ludvig E. A.: Reinforcement Learning in Animals, The Encyclopedia of the Sciences of Learning, New York: Springer, 2012, pp. 2799-2802

Modeling and Prediction of the end of Life Vehicles Number Distribution in Serbia

Miroslav D. Sokić[1], Ilija B. Ilić[2], Vaso D. Manojlović[1], Branislav R. Marković[1], Zvonko P. Gulišija[1], Milan D. Pavlović[3], Nada D. Štrbac[4]

[1] Institute for Technology of Nuclear and Other Mineral Raw Materials, 86 Franchet d'Esperey Street, 11000 Belgrade, Serbia; m.sokic@itnms.ac.rs, v.manojlovic@itnms.ac.rs, b.markovic@itnms.ac.rs, z.gulisija@itnms.ac.rs

[2] Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia; miljan.ilic@mbpvs.gov.rs

[3] Technical Faculty "Mihajlo Pupin" in Zrenjanin, University of Novi Sad, Đure Đakovića bb, 23000 Zrenjanin, Serbia, pmilan@sbb.rs

[4] University of Belgrade, Technical Faculty, VJ 12, 19210 Bor, Serbia, nstrbac@tf.bor.ac.rs

Abstract: The impact of various time-defendant factors on the recycling rate of end-of-life vehicles (ELV) in Republic of Serbia was investigated. Statistical distribution of the frequency of the number of ELV in the year of dismantling depending on the year of production of ELV is designed using the two-parameter Weibull distribution function and MATLAB software, based on a real time data. Obtaining the time-dependence of Weibull parameters, a statistical distribution of frequency of the number of ELV in the coming period in Serbia was simulated. These results in combination with amount of materials in the most abundant cars in Serbia were used to simulate the overall amount of materials, which are available for recycling, in the coming period. These results are essential for automotive recycling industry management, particularly for shredders, dismantlers and metal pre-processors.

Keywords: dynamic model; ELV; recycling; secondary raw materials; simulation

1 Introduction

Metallic secondary raw materials are generated from the industries that use metals. Different areas of application and the various methods of processing and use as structural materials leads to the formation of a very diverse scrap and/or waste material supply [1]. Industrial products include a variety of materials related

to each other in various ways. After the end of the product life cycle, due to deterioration or technological obsolescence, they need to be recycled to the valuation of the material from which the product is made. EU legislation has defined the high standards and stringent environmental regulations [2], therefore, the recycling of various metallic secondary raw materials becomes inevitable [3]. However, after recycling, metals may lose their initial characteristics, both in terms of purity and in terms of mechanical properties. Hence, the recycled metallic materials are commonly used in less demanding applications. In order to minimize the losses of these materials due to contamination during the recycling process, it is essential to understand the relations between the technological processes of recycling and materials physical and chemical characteristics [4]. In the primary natural resources, metals in the mineral forms are bounded in the different ores. Exploitation and production of base metals is a complex series of a large number of technological processes adapted to the relevant type of ore. Obtained metals are incorporated into products which, after the expiration of the life cycle, become the amortization waste. The waste is collected, prepared for processing and sent to recycling streams, which are introduced in the production of metals as industrial recycling resources. Thus, metals find their way back to the source of the cycle and provide the basis of industrial cycles.

Modern society is regulated by extensive use of complex multi-component products, of which, vehicles are a typical example. They are composed of different materials, provided that the individual components are increasingly minimized, whether they are expensive, or have a detrimental effect on the function of the product. From an economic point of view, in practice, the complete disassembly of the product is unattractive and unprofitable. Therefore, partial disassembly, shredding, chopping and crushing are performed and thus prepared material returns to the basic resources of the system. However, with such a treatment and processing, the pure metal cannot be obtained, because there is contamination due to strong joints of different materials in the product and the imperfections of the processes of separation. As a result, the metal material contains impurities that are difficult to distinguish using the existing metallurgical processes [5]. The results are significant non-refundable losses and a less quality of metal obtained from secondary raw materials.

Automotive recycling infrastructure is related to:

- Dismantlers
- Shredders
- Operators of non-ferrous metals

The flow of materials within the automotive recycling infrastructure is shown in Fig. 1.

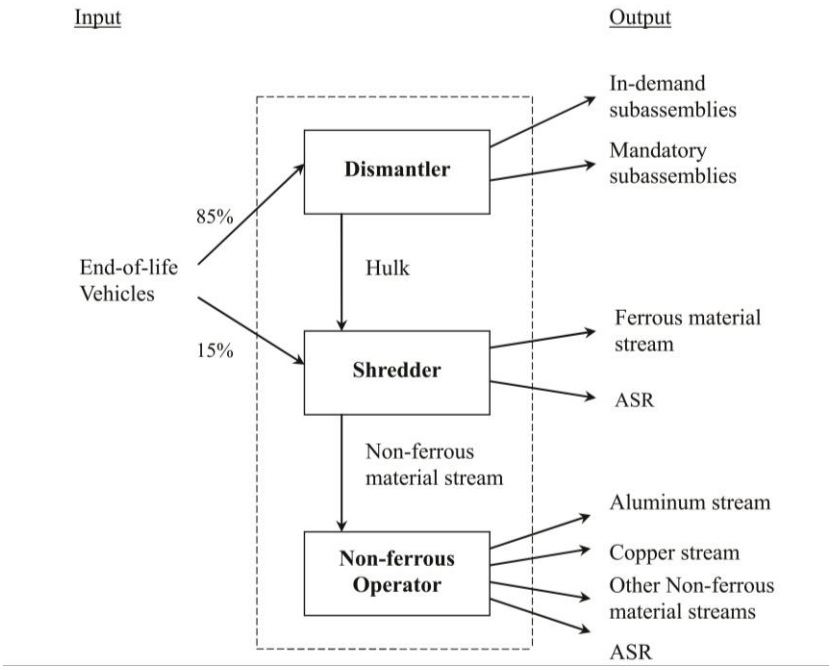


Figure 1
The materials flow of the ELV in the automotive recycling infrastructure [19]

Dismantlers remove parts from the car such as the engine, transmission, radiator, catalytic converter, petrol, fuel tank, fluids, tires, batteries and air bags because of their value or because of the Shredder requirements for their removal [6]. It is estimated that the content of non-metallic materials after disassembling is almost the same as before dismantling, because the parts removed by dismantlers are mainly based on metals. After removing the automotive parts, the car shell is sent to the shredder.

The shredding process enables the separation of the multicomponent materials by magnetic separation and by the methods based on differences in specific weight of materials. As shown in the diagram (Fig. 1), there are three main output streams of the shredding process: the stream of ferrous metals, the stream nonferrous metals, and the stream of automotive shredding residue (ASR). Ferrous metals, recycled by shredding process, are sold on the steelmaker market, non-ferrous metals are sold in the operators and non-ferrous metal market, and the shredder residue is deposited in the landfields [7, 8].

Operators of non-ferrous metals separate the stream of non-ferrous metals into different flows such as aluminum, copper and zinc.

A shredder plant with a capacity of 34000t/year for one shift, in the recycling center "Scholc"-Železnik in Belgrade, has successfully operated for many years.

2 Models for Monitoring the Impact of Various Parameters

The influence of parameters such as the life cycle design, metal content and material combination within the vehicle on recyclability of the car is significant. Examination of these parameters and monitoring of their trends and changes is essential for the successful management of the major economic issues within the automotive industry.

In one study [9], assessment of material/components flow and economic exchange within the life cycle of the material in the automotive industry has been developed. Issues which were processed were:

- What could be the effect to increase the recycling of plastic?
- What could be the impact in a decline in car sales?
- What could happen if most parts of a vehicle are built of aluminum, plastic and ultra-light steel?

Isaacs and Gupta [10] developed optimization model to describe the impact of the vehicle with a high content of polymeric materials on the automotive recycling infrastructure. The aim of this study was to determine the optimal level of plastics which are needed to be removed, after the shredding process, to maintain the profitability of all business areas within the infrastructure. Using the same approach, two additional studies were conducted, one to determine the increasing effect of the materials content based on aluminum in vehicle on the infrastructure [11], and other to compare the impact of electronic vehicles, hybrid vehicles and vehicles with a high content of polymeric materials on the automotive recycling infrastructure [12,13]. Optimization is the ideal approach in decision making for a given scenario, but this approach requires a precise mathematical determination of constraints and objectives.

Zamudio [14] developed a dynamic model to describe the recycling of cars. This model does not take into account materials which are connected with the replacement of parts, as well as, dismantling process or economic concerns for that process. Also, in the flow of the materials the interdependence is not assumed, so that the flow of various materials (e.g. steel and aluminum) will never be analyzed separately. It was concluded that the existing automotive recycling infrastructure should be modified in order to resist changes in the market.

Several other models use a dynamic system, in order to characterize the material flow and economic exchange within the recycling infrastructure. To explain the impact of changes in the material composition of the vehicle on the sustainability of the economy and the environment within the automotive recycling infrastructure, Bandivadekar et al. [15] developed a simulation model for recovery and recycling of cars. This model describes the flow of materials and economic

exchange of any business issue within the automotive recycling infrastructure. This model is then used to determine the profitability of each business issue due to a certain changes in the market and changes in the material composition of vehicles.

In another study based on a dynamic system, a strategy to meet the provisions of the EU ELV regarding recovery and recycling of materials was analyzed [16, 17, 18]. In the study, conducted by Kumar and Sutherland [19, 20], different technological strategies that can justify the effectiveness of the recovery of materials in the automotive recycling infrastructure were analyzed. The aim of their analysis was to determine the technological costs which could be related to the shredders and dismantlers used to achieve recovery of materials compared to those established in Europe.

Dynamic annual flow models, incorporating consumer discards and usage losses and featuring deterministic and stochastic end-of-cycle returns by the consumer, were developed for reused or remanufactured products, including fast/slow cycling and short/long-lived products [21].

In order to define the link between end-of-life products, their composition, and design a dynamic model was developed for cars [22, 23, 24]. This model permits the visualization of the influence of various parameters and distributions on the recycling rate over time. In addition, this model also permits the formulation of an improved definition of recycling rate. The architecture of the dynamic model, as shown in Fig. 2, will be used to capture the rapidly changing weight, composition, and lifetime of the car, which determine directly the recycling rate of vehicles. The final metal recovery depends on the prevalent thermodynamics and kinetics of the metallurgical processes, which are related to the material cycle [5]. Van Schaik and Reuter [24] showed that the recycling rate cannot be represented by an average or single value as required by EU legislation, but is largely dependent on the distributed nature and therefore the standard deviations of the time-varying lifetime, weight and composition of the car. The various simulations, using the dynamic system model, make clear that the weight and composition of the car at production and dismantling are determined by the different distributions and are highly dependent on their fluctuations.

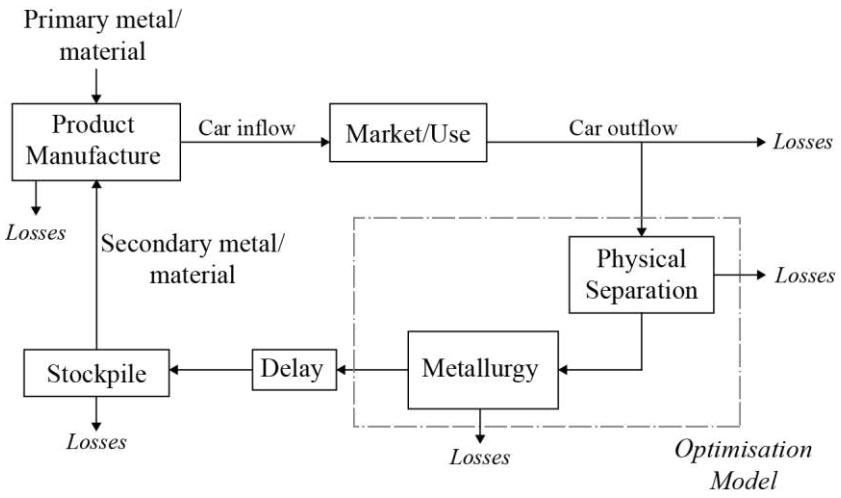


Figure 2
Scheme of the dynamic model [22]

It is estimated that in the Republic of Serbia, currently in use, are more than two million vehicles (2,274,770 vehicles in the year 2008, including trailers, tractors, commercial vehicles and motorcycles), whose average life cycle is between 18 and 19 years [25]. It is estimated that more than one million older vehicles, in various states of functioning, waiting to be recycled, and about 100,000 to 120,000 vehicles each year cannot be recycled, due to deterioration.

This state of the automotive recycling industry originated, primarily due to an unstable political and socio-economic situation in the country, in the period between 1990 and 2002. After 2002, the situation in the country significantly improved and was reflected in the automotive recycling industry. The dynamic model which describes the number of cars at the end of the life cycle, as well as, monitoring trends in the composition of the materials is of crucial importance for businesses within the automotive recycling industry. The model is crucial to properly plan processing capacity and to develop related business models that will ultimately lead to positive economic development and sustainability.

3 Recycling of Cars - Simulation

The statistical distribution of the number of cars on their end of life can be described by a normal distribution, depending on the year of production and year of dismantling of the cars (Fig. 3). These diagrams are constructed based on the number of cars that are available for recycling in a given year, along with the average age of the cars that are recycled in the same year. The diagrams were

constructed using the software package MATLAB. The mean value of the number of cars, which are recycled in one year, can be approximated by the number of unregistered cars in the same year. The average number of unregistered cars in the period between 1994 and 2001 year is approximately 58000 cars per year, and approximately 64000 cars in the period between 2002 and 2008 year [25].

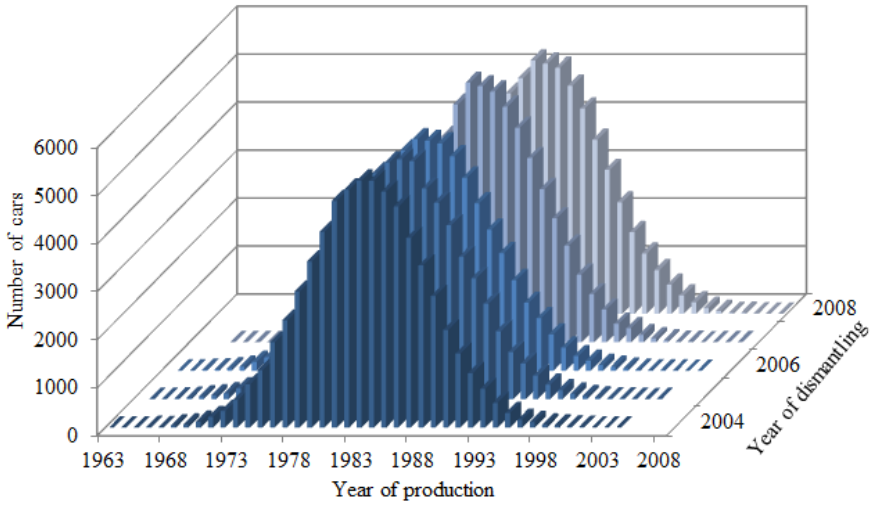


Figure 3
The normal distribution of the number of waste cars, by year of production and dismantling, in Serbia

The normal distribution curves in Fig. 3 can be described by the two-parameter Weibull distribution function:

$$W(t) = \frac{a}{b^a} t^{a-1} exp\left[-\left(\frac{t}{b}\right)^a\right] \quad (1)$$

where (a) is the shape parameter, and (b) is the scale parameter, (a, b>0). The shape parameter, also known as the Weibull slope, gives the Weibull distribution its flexibility. The scale parameter determines the range of the distribution.

Fitting of the curves in Figure 3 with the Weibull distribution function is done by using the MATLAB software package, wherein the correlation coefficients from R = 0.99 are obtained. Also distribution parameters (a) and (b) are determined, whereby the average value of the shape parameter is a = 4, and the scaling parameter, (b) is shown in functional dependence over the time period (Fig. 4). When (b) is increased, while (a) is kept the same, the distribution gets stretched out to the right and its height decreases, while maintaining its shape (Fig. 3).

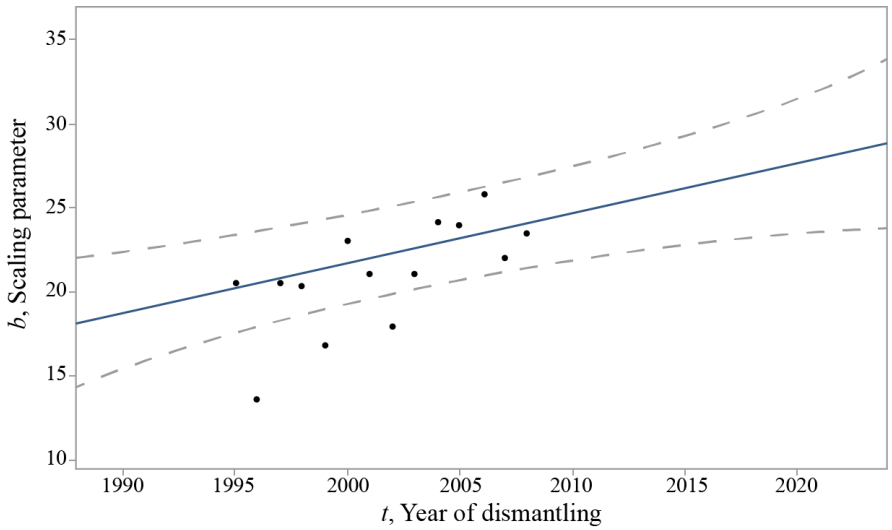


Figure 4

Dependence of the scaling parameter with the year of dismantling

The dependence of the scaling parameter (b) from the year disassembly (t) can be expressed by the following equation:

$$b(t) = 0{,}297 \cdot t - 573 \tag{2}$$

The coefficients in the equation (2) were calculated with confidence limits of 95%. Also, the prediction boundary of 95% was determined, which is shown with dashed lines in Figure 4. It should be noted that the values of the scaling parameter in 1996, 1999, 2002 years were rejected during determining of functional dependence. Major deviation of the parameter (b), in the aforementioned years, is attributed to the unstable socio-political situation in Serbia in those years.

The dependence of the parameter (b) from the time (Equation 2), together with the Weibull distribution function and the average dismantling year of ELV (Equation 1), are used in predicting the distribution of the number/frequency of the car that are recycled in the period between 2015 and 2025 year (Fig. 5).

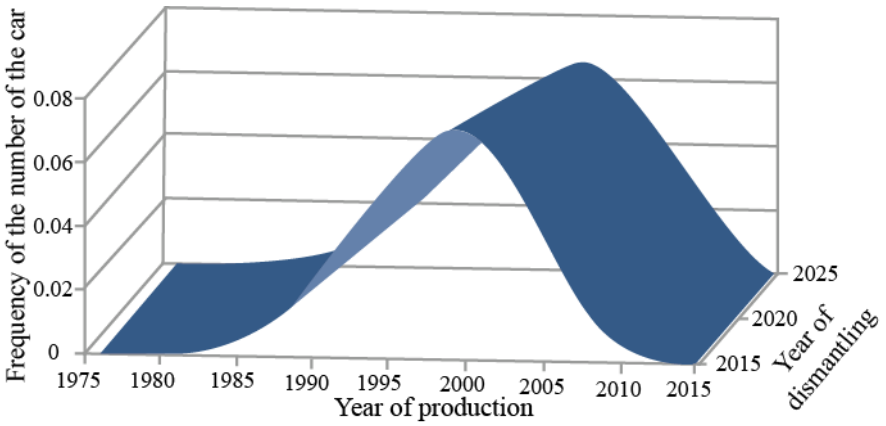


Figure 5
The distribution of the number of ELV, depending on the year of production and the year of dismantling in the forthcoming period

Based on the Fig. 5, we can see that the width of the number of ELV distribution will increase and the shift to the right to higher production years. Maximum number of ELV will decline slightly over time, as could be expected based on the trend of increase in the parameter (b).

An estimation of the quantity of materials obtained by recycling of the ELV, can be obtained by determining the trend of the most prevalent car models at the end of the life cycle and the amount of materials within those models. According to the Serbian Ministry of Interior Affairs, the most common brands of cars were Zastava, Opel and Volkswagen. The trend of the ELV share for domestic brand Zastava and other unregistered cars in the period between 2003 and 2008 year is shown in Table 1.

Table 1
Share (in %) of unregistered cars "Zastava" and others in the period since 2003 to 2008 year [24]

Brand/Year	2003	2004	2005	2006	2007	2008
Zastava	58.76	52.68	59.69	60.87	50.47	48.34
Others	41.24	47.32	40.31	39.13	49.53	51.66

Trend of the share of ELV which will be recycled in the forthcoming period can be determined according to Table 1 (Fig. 6). Domestic brand of cars "Zastava" has been stopped production since the end of the 2008 year, it is logical that their share in the forthcoming period of recycling will rapidly decrease.

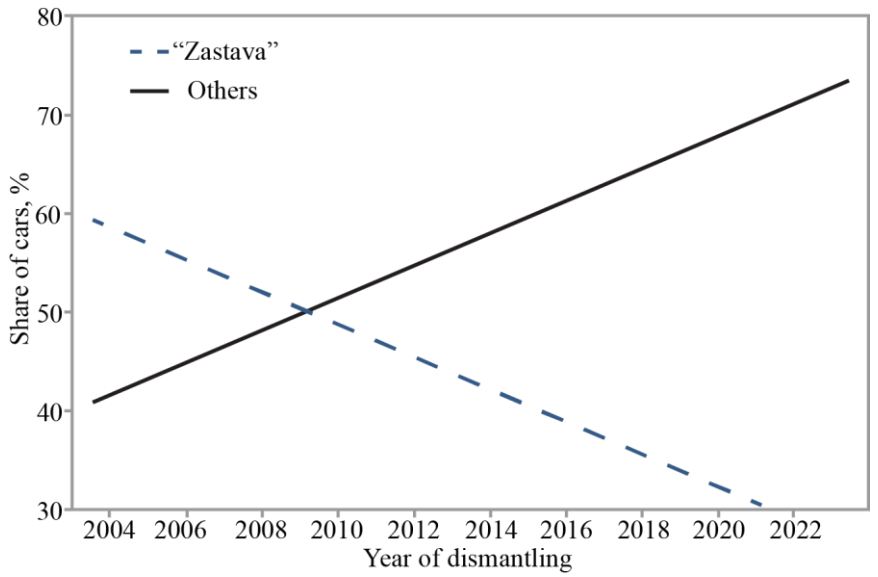


Figure 6
Trend of the share of unregistered cars "Zastava" and others in the forthcoming period in Serbia

The distribution of materials in cars Zastava and Opel Astra are given in Table 2.

Table 2
Share of materials in the car Zastava and Opel Astra

Material	Zastava		Opel Astra	
	Weight (kg)	Share (%)	Weight (kg)	Share (%)
Steel/iron	626	75.00	681	63.61
Other metals	48.9	5.86	89	8.33
Plastics	41.7	5.00	185	17.25
Pneumatics	25.9	3.10	30	2.82
Glass	30.9	3.70	31	2.85
Fluids	15.4	1.84	17	1.61
Rubber	8.3	1.00	17	1.61
Other	37.9	4.50	20	1.92
Totally	835	100	1070	100

According to these data, we can determine the total amount of waste material from cars that will be generated in the period between 2003 and 2023 year, with the approximation that the materials composition of other cars corresponds to the material composition of Opel Astra car as the dominate brand. Together with a simulation of the ELV number in the forthcoming period from Fig. 5, a diagram of the material amount prediction in dismantling year can be constructed.

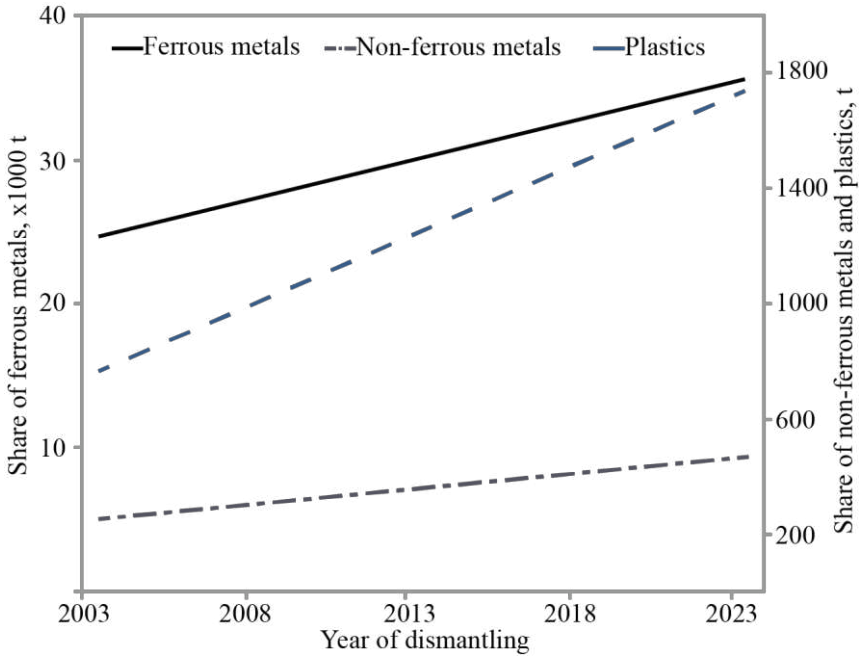


Figure 7

Predicting the amount of material per year of dismantling in Serbia

According to the diagram in Fig. 7, we see that the slope in plastic materials is the largest and the smallest is in non-ferrous metals. For ferrous metals, we also notice a growing trend, which is a result of an increase in the share of the foreign cars that have higher weight than domestic cars "Zastava".

Conclusions

The recycling industry in Serbia is in a growth stage, thus monitoring and forecasting of ELV is very important. Dismantlers, shredders, metal pre-processors and other related companies within the recycling industry must plan their business based on these data. Accuracy of the forecasting model depends on accuracy of real time data and simulation methods used for forecasting. The two-parameter, Weibull distribution function has been proven to be valid for fitting with the data of ELV distribution, in the past period, as well as, for forecasting of ELV which will generate within the 10 year period between 2015 and 2025. Results show that distribution of the ELV, in particular year of dismantling, will be spread and shift with regard to the higher year of their production.

The unstable socio-political situations in the country, in previous years, and the impact of domestic car production intensity, have been presented as the most influential parameters for the number of ELV and overall materials in ELV. The overall materials from ELV in year of dismantling obtained from the simulation of

the ELV in future period and from distribution of the most common types of cars in Serbia, indicate that plastic material have a higher growth rate, from 800 t (in 2003) to 1800 t (in 2023). Also, ferrous materials have a significant growth rate, from 25 kt (in 2003) to 35 kt (in 2023), due to the changes in the distribution of car types, in a particular dismantling year. Further work should be focused on simulating the various factors such as Weibull parameters and the distribution of the most common types of cars in Serbia in order to obtain different and more accurate forecasting models.

Acknowledgement

The authors wish to acknowledge the financial support from the Ministry of Education and Science of the Republic of Serbia through the projects TR34023 and TR35033.

References

[1] I. Ilić , Z. Gulišija, M. Sokić, *Recycling of Metallic Secondary Raw Materials*, ITNMS, Belgrade, 2010 (in Serbian)

[2] Directive 2000/53/EC. Commission of the European Communities, directive 200/53/EC of the European Parliament and the council on end of life vehicles. Off J Eur Union, Brussel (2000)

[3] V. Simić, B. Dimitrijević, Risk Explicit Interval Linear Programming Model for Long-Term Planning of Vehicle Recycling in the EU Legislative Context under Uncertainty, *Resour. Conserv. Recycl.*, 73, 197-210 (2013)

[4] M. B. G. Castro, J. A. M. Remmerswaal, M. A. Reuter, U. J. M. Boin, A Thermodynamic Approach to the Compatibility of Materials Combinations for Recycling, *Resour. Conserv. Recycl.*, 43, 1-19 (2004)

[5] M. A. Reuter, U. J. M. Boin, P. Rem, Y. Yang, N. Fraunholcz, A. Van Schaik, The Optimization of Recycling: Integrating the Resource, Technological, and Life Cycles, *JOM*, 56, 33-37 (2004)

[6] J. Staudinger, G. A. Keoleian, M. S. Flynn, Management of End-of-Life Vehicles (ELVs) in the US, Report for Japan External Trade Organization (JETRO). University of Michigan Center of Sustainable Systems, Report No. CSS01-01, Ann Arbor, MI (2001)

[7] D. Klempner, K. Frisch, B. Pokorski, V. Sendijarevic, Characterization of Various ASR Streams, *SAE Technical Paper 1999-01-0670*, doi:10.4271/1999-01-0670 (1999)

[8] Y. W. Cheng, J. H. Cheng, C. L. Wu, C. H. Lin, Operational Characteristics and Performance Evaluation of the ELV Recycling Industry in Taiwan, *Resour. Conserv. Recycl.*, 65, 29-35 (2012)

[9] A. Bustani, P. Mackay, E. Cobas-Flores, S. Yester, J. Sullivan, R.L. Williams, An Approach to Modelling the Vehicle End-of-life Process, *SAE Technical Paper 980099*, doi:10.4271/980099 (1998)

[10] J. A. Isaacs, S. M. Gupta, Economic Consequences of Increasing Polymer Content for the USA Automobile Recycling Infrastructure, *J. Ind. Ecol.*, 1, 19-33 (1997)

[11] J. E. Boon, J. A. Issacs, S. M. Gupta, Economic Impact of Aluminum - Intensive Vehicles on the USA Automotive Recycling Infrastructure, *J. Ind. Ecol.*, 4, 117-134 (2000)

[12] J. E. Boon, J. A. Issacs, S.M.Gupta, End-of-Life Infrastructure Economics for 'Clean Vehicles' in the United States, *J. Ind. Ecol.*, 7, 25-45 (2003)

[13] Siavash Sadeghi, Mojtaba Mirsalim, Arash Hassanpour Isfahani, Dynamic Modeling and Simulation of a Switched Reluctance Motor in a Series Hybrid Electric Vehicle, *Acta. Polytech. Hung.*, 7, 51-71 (2010)

[14] P. Zamudio, Economics of Automobile Recycling. MIT MS Thesis, Boston, MA (1996)

[15] A. P. Bandivadekar, V. Kumar, K. L. Gunter, J. W.Sutherland, A Model for Material Flows and Economic Exchanges within the USA Automotive Life Cycle Chain, *J. Manuf. Syst.*, 23, 22-29 (2004)

[16] P. Ferrao, P. Nazareth, J. Amaral, Strategies for Meeting EU End-of-Life Vehicle Reuse/Recovery Targets, *J. Ind. Ecol.*, 10, 77-93 (2006)

[17] P. Ferrao, J. Amaral, Assessing the Economics of Auto Recycling Activities in Relation to European Union Directive on End of Life Vehicles, *Technol. Forecasting Social Change*, 73, 277-289 (2006)

[18] J. Amaral, P. Ferrao, C. Rosas, Is Recycling Technology Innovation a Major Driver for Technology Shift in the Automobile Industry under an EU Context?, *Int. J. Technol. Policy Manage.*, 6, 385-398 (2006)

[19] V. Kumar, J. W. Sutherland, Achieving Higher Material Recovery Rates from End-of-Use Vehicles, *Trans of NAMRI/SME*, 35, 201-208 (2007)

[20] V. Kumar, J. W. Sutherland, Sustainability of the Automotive Recycling Infrastructure: Review of Current Research and Identification of Future Challenges, *Int. J. Sustainable Manuf.*, 1, 145-167 (2008)

[21] C. A. Tsiliyannis, Internal Cycle Modelling and Environmental Assessment of Multiple Cycle Consumer Products, *Waste Manage.*, 32, 177-193 (2012)

[22] A. Van Schaik, W. L. Dalmijn, M. A. Reuter, Impact of Economy on the Secondary Material Cycle, Proceedings COM Waste Processing and Recycling in Mineral and Metallurgical Industries IV, Toronto, Canada, 2001, pp. 407-423

[23] A. Van Schaik, M. A. Reuter, U. M. J. Boin, W. L. Dalmijn, Dynamic Modelling and Optimization of the Resource Cycle of Passenger Vehicles, *Miner. Eng.*, 15, 1001-1016 (2002)

[24] A. Van Schaik, M. A. Reuter, The Time-Varying Factors Influencing the Recycling Rate of Products, *Resour. Conserv. Recycl.*, 40, 301-328 (2004)

[25] Ministry of Internal Affairs of Serbia., Data on Registration of Motor Vehicles in Serbia for the Period of 1998 to 2008. In: Belgrade, Serbia, (2009)

Application of Potentials in the Description of Transport Processes

Katalin Gambár, Marianna Lendvay, Rita Lovassy, József Bugyjás

Institute of Microelectronics and Technology, Kálmán Kandó Faculty of Electrical Engineering, Óbuda University, Tavaszmező u. 17, H-1084 Budapest, Hungary
gambar.katalin@kvk.uni-obuda.hu, lendvay.marianna@kvk.uni-obuda.hu, lovassy.rita@kvk.uni-obuda.hu, bugyjas.jozsef@kvk.uni-obuda.hu

The most important equations of motion in physics are summarized in differential equations. Variational calculus is suitable to unify the different disciplines of physics; it is even classical mechanics, electrodynamics or modern field theories. The basic equations of the disciplines can be deduced from the least action principle, the Hamilton's principle. It is shown that the Lagrange function can be formulated for dissipative processes, in systems with infinite degree of freedom, thus the Hamilton's principle can be considered as a basis of the theory. The Lagrange function can be constructed by an introduced scalar (potential) field that defines the measurable physical quantities.
In the present paper we will construct a Lagrangian density function in such a way that the field equations (equations of motion) are known, but these equations contain non-selfadjoint operators. For this, it is necessary to introduce potentials. We suggest what possible directions are open in the study of the system, through the elaboration of the mathematical model.

Keywords: Hamilton's principle; dissipation; potential; adjoint operator; canonical formalism

1 Introduction

When Lagrange showed Euler his work, in which he deduced the principles of mechanics, based on the variational calculus, he did not think that it would spread to all disciplines of physics or the presented method might become a pillar of modern physics.

It is known that the equations for motion in physics can be formulated by differential equations. Most of them can be deduced from the least action principle, the Hamilton's principle. Experience shows that for those theories, where the Hamilton's principle can be applied and the related Hamilton-Lagrange

formulas can be elaborated not only the aesthetics" of the theory are impressive, but often, further improvements and discoveries develop.

The degrees of freedom of continuous media (physical fields) are infinite. Their descriptions change from place to place and moment to moment. We assign mathematical variables (field quantities) to individual points of geometric fields at each moment. The time evolution of these variables can be described by the field equation (equations of motion). In general, there are two basic steps to construct the actions for systems with infinite degrees of freedom: taking into account the properties of the system the Lagrange density function should be formulated by the field variables; the action can be obtained by the spatial and time integration of the Lagrange density function. If we know the action, we accept the Hamilton's principle, as an axiom (basic principle), then the complete Hamilton-Lagrange formalism can be built up and applied. It took a long time to recognize just how to establish this description for dissipative systems. It can be shown that the Lagrange function of a dissipative process, with an infinite degree of freedom, can be formulated by introducing a relevant physical-mathematical additional field, thus, Hamilton's principle gives the foundation of the theory [1, 2, 3]. The Lagrange function itself can be formulated with the help of this scalar field, which has a consistent mathematical relation, with measurable physical fields. Since the Hamilton principle is not sensitive as to how the Lagrange function is created, we can exploit the possibilities of variational calculus, the elaboration the complete canonical formalism and we can explore the various related research directions, based on the developed theory.

2 Construction of Lagrange Functions

2.1 Hamilton's Principle

In order to consider the Hamilton's principle, as a base of the theory, we need to formulate the Lagrange density function *L,* which the time integral is the action *S.*

$$S = \int L \, \mathrm{d}V\mathrm{d}t \tag{1}$$

The Lagrange density function for a Ψ scalar field is written by its first order derivatives as:

$$L = L\left(\Psi, \dot{\Psi}, \nabla\Psi\right) \tag{2}$$

In the sense of the Hamilton's principle, the variation of action is:

$$\delta S = \delta \int_{t_1}^{t_2} L \, \mathrm{d}V\mathrm{d}t = 0 \tag{3}$$

This means that for a realistic process of nature, the action is an extremum, i.e., the variation of action is zero. If the Lagrange function depends on the physical

field Ψ and its derivatives then after the variation of action the Euler-Lagrange differential equation can be obtained. The following describes the time and spatial evolution of the field:

$$\frac{\partial L}{\partial \Psi} - \frac{\partial}{\partial t}\frac{\partial L}{\partial \dot{\Psi}} - \nabla \cdot \frac{\partial L}{\partial \nabla \Psi} = 0 \tag{4}$$

Naturally, the Lagrange function may include more variables and higher order derivatives.

In general, if a linear operator O acts on the function Ψ in the Lagrange function then an expression

$$\tilde{O}\frac{\partial L}{\partial (O\Psi)} \tag{5}$$

appears in the Euler-Lagrange equation where $\tilde{O}$ is the adjoint operator of operator O. The most frequent operator – adjoint operator pairs in physics are:

$$O = \frac{\partial}{\partial t} \rightarrow \tilde{O} = -\frac{\partial}{\partial t} \tag{6}$$

$$O = \mathrm{grad}(= \nabla) \;\rightarrow\; \tilde{O} = -\mathrm{div}(= \nabla \cdot) \tag{7}$$

$$O = \mathrm{rot}(= \nabla \times) \rightarrow\; \tilde{O} = \mathrm{rot}(= \nabla \times) \tag{8}$$

$$O = \mathrm{div}(= \nabla \cdot) \rightarrow \tilde{O} = -\mathrm{grad}(= \nabla) \tag{9}$$

$$O = \frac{\partial^2}{\partial t^2} \rightarrow \tilde{O} = \frac{\partial^2}{\partial t^2} \tag{10}$$

$$O = \Delta \rightarrow \tilde{O} = \Delta \tag{11}$$

If the equality $O = \tilde{O}$ is completed then we speak about a self-adjoint operator. The Lagrange function can easily be constructed if equation of motion of a variable is known.

2.2 The Example of Electrodynamics

In the knowledge base for electric charges and currents, using the Maxwell equations, an electromagnetic field can be formulated in a simpler form, introducing the scalar and vector potentials. These define the measurable field variables as [4]:

$$\vec{E} = -\frac{\partial \vec{A}}{\partial t} - \mathrm{grad}\varphi \tag{12}$$

$$\vec{B} = \mathrm{rot}\vec{A} \tag{13}$$

In order to simplify the equations the potentials can be modified as:

$$\overrightarrow{A'} = \vec{A} + \mathrm{grad}\psi \tag{14}$$

$$\varphi' = \varphi - \frac{\partial \psi}{\partial t} \tag{15}$$

i.e., the definition of potentials are free from a gradient or time derivative, of an arbitrary scalar field. This means a free choice of potentials, suitably, e.g. it can be (Lorenz measure):

$$\mathrm{div}\vec{A} = -\varepsilon_o \mu_o \frac{\partial \varphi}{\partial t} \tag{16}$$

In that part of space where the charge density and the currents are zero, the Lagrange density function for the Maxwell equations can be expressed by the potentials:

$$\mathcal{L} = \frac{1}{2}\varepsilon_o \left(\frac{\partial \vec{A}}{\partial t} + \mathrm{grad}\varphi\right)^2 - \frac{1}{2\mu_o}\left(\mathrm{rot}\vec{A}\right)^2 \tag{17}$$

The Euler-Lagrange differential equations can be obtained by the variation of $\vec{A}$:

$$\mathrm{rot}\,\mathrm{rot}\frac{\vec{A}}{\mu_o} - \varepsilon_o \frac{\partial}{\partial t}\left(-\frac{\partial \vec{A}}{\partial t} - \mathrm{grad}\varphi\right) = 0 \tag{18}$$

which results the Euler-Lagrange differential equations. By the application of definition equation (13) the well-known Maxwell equations are obtained:

$$\mathrm{rot}\vec{B} = \varepsilon_o \mu_o \frac{\partial \vec{E}}{\partial t} \tag{19}$$

After the variation by φ the equation

$$\mathrm{div}\left(\frac{\partial \vec{A}}{\partial t} + \mathrm{grad}\varphi\right) = 0 \tag{20}$$

can be deduced. Applying equation (12), we get

$$\mathrm{div}\vec{E} = 0 \tag{21}$$

formulating the next Maxwell equation. The equations

$$\mathrm{div}\vec{B} = 0 \tag{22}$$

$$\mathrm{rot}\vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{23}$$

are completed automatically by the application of definition equations (12) and (13).

Considering the $\vec{A}$ dependence in the Lagrange density function, the operators $O_1 = \frac{\partial}{\partial t}$ and $O_2 = rot$ appear. According to the previous mathematical list in equations (6) – (11) the terms $-\frac{\partial}{\partial t}\frac{\partial L}{\partial(\frac{\partial \vec{A}}{\partial t})}$ and $\mathrm{rot}\frac{\partial L}{\partial(\mathrm{rot}\vec{A})}$ appear in the Euler-Lagrange equation. The validity can be easily checked, since

$$-\frac{\partial}{\partial t}\frac{\partial L}{\partial\left(\frac{\partial \vec{A}}{\partial t}\right)} = -\varepsilon_o \frac{\partial}{\partial t}\left(\frac{\partial \vec{A}}{\partial t} + \mathrm{grad}\varphi\right) \tag{24}$$

and

$$\mathrm{rot}\frac{\partial L}{\partial(\mathrm{rot}\vec{A})} = -\frac{1}{\mu_o}\mathrm{rot}\,\mathrm{rot}\vec{A} \tag{25}$$

from which

$$\frac{1}{\mu_o}\mathrm{rot}\,\mathrm{rot}\vec{A} = \frac{1}{\mu_o}\mathrm{rot}\vec{B} = \varepsilon_o\frac{\partial}{\partial t}\left(\frac{\partial\vec{A}}{\partial t} + \mathrm{grad}\varphi\right) = \varepsilon_o\frac{\partial\vec{E}}{\partial t} \tag{26}$$

Taking the φ dependence of the Lagrange density function, the operator $O_3 = \mathrm{grad}$ stands, thus the term $-\mathrm{div}\frac{\partial L}{\partial(\mathrm{grad}\varphi)}$ appears in the second Euler-Lagrange equation. After these, as an Euler-Lagrange equation we arrive at a Maxwell equation with zero charge

$$0 = \mathrm{div}\left(\frac{\partial\vec{A}}{\partial t} + \mathrm{grad}\varphi\right) = \mathrm{div}\vec{E} \tag{27}$$

In general, the Hamilton density function of the field can be introduced as

$$\mathcal{H} = \mathcal{P}\frac{\partial\overrightarrow{A^{\blacksquare}}}{\partial t} - \mathcal{L} \tag{28}$$

where

$$\overrightarrow{A^{\blacksquare}} = \left(A_x, A_y, A_z, \frac{i}{c}\varphi\right) \tag{29}$$

is the four-potential, and $\mathcal{P}$ is the canonically conjugated variable (four-mometum) to $\overrightarrow{A^{\blacksquare}}$:

$$\mathcal{P} = \frac{\partial\mathcal{L}}{\partial\frac{\partial\overrightarrow{A^{\blacksquare}}}{\partial t}} = -\varepsilon_o\left(-\frac{\partial\vec{A}}{\partial t} - \mathrm{grad}\varphi\right) \tag{30}$$

After all, after using the Lagrange density function in equation (17) the Hamilton density function can be written for pure radiation field:

$$\mathcal{H} = \frac{\mathcal{P}^2}{2\varepsilon_o} - \mathcal{P}\mathrm{grad}\varphi + \frac{1}{2\mu_o}\left(\mathrm{rot}\vec{A}\right)^2 \tag{31}$$

The application of scalar and vector potentials is not the only choice to describe the electromagnetic field. Another representation also exist which is especially suitable to solve certain problems (e.g. dipol radiation) [4, 5].

The so-called Hertz vector $\overrightarrow{\Pi}$ can be introduced to the vector potential as:

$$\vec{A} = \varepsilon_o\mu_o\frac{\partial\overrightarrow{\Pi}}{\partial t} \tag{32}$$

It can be seen that the relation

$$\mathrm{div}\vec{A} = -\varepsilon_o\mu_o\frac{\partial\varphi}{\partial t} = \varepsilon_o\mu_o\frac{\partial}{\partial t}\mathrm{div}\overrightarrow{\Pi} \tag{33}$$

is completed, in which the Lorenz condition can be recognized. As direct consequence of this that the scalar field φ can be deduced by $\overrightarrow{\Pi}$:

$$\varphi = -\mathrm{div}\overrightarrow{\Pi} \tag{34}$$

Then the measurable field variables can be given by this only one potential field:

$$\vec{B} = \varepsilon_o \mu_o \frac{\partial}{\partial t} \mathrm{rot} \overrightarrow{\Pi} \tag{35}$$

$$\vec{E} = -\varepsilon_o \mu_o \frac{\partial^2 \overrightarrow{\Pi}}{\partial t^2} + \mathrm{grad}\, \mathrm{div} \overrightarrow{\Pi} \tag{36}$$

It can be checked by a short calculation that the Hertz vector $\overrightarrow{\Pi}$ completes the wave equation:

$$\frac{\partial}{\partial t}\left(\Delta \overrightarrow{\Pi} - \varepsilon_o \mu_o \frac{\partial^2 \overrightarrow{\Pi}}{\partial t^2}\right) = 0 \tag{37}$$

3 Hamilton-Lagrange Formalism for Dissipative Processes

3.1 The Simplest Transport Process: Linear Heat Conduction

The simplest pure dissipative process is the Fourier heat conduction, which is described by a parabolic differential equation. This differential equation is formulated for the classic temperature, applying the local equilibrium hypothesis. This can be done – and it gives a good description – if the energy transport is rather slow [1, 2, 3, 6, 7, 8].

Let us consider the linear Fourier heat conduction as the simplest dissipative process which can be written by the equation

$$\frac{\partial T}{\partial t} - \frac{\lambda}{c_v} \frac{\partial^2 T}{\partial x^2} = 0 \tag{38}$$

The process is described by parabolic partial differential equation in general where $T(x,t)$ means the local equilibrium temperature. The Lagrange density function cannot be directly formulated by temperature T, since the time derivative is not self-adjoint operator. This difficulty can be resolved with the introduction of four times differentiable scalar field $\varphi(x,t)$ that generates the measurable field:

$$T = -\frac{\partial \varphi}{\partial t} - \frac{\lambda}{c_v} \Delta \varphi \tag{39}$$

Then by the potential φ the Lagrange density function of Fourier heat conduction can be expressed:

$$L = \frac{1}{2}\left(\frac{\partial \varphi}{\partial t}\right)^2 + \frac{1}{2}\frac{\lambda^2}{c_v^2}(\Delta \varphi)^2 \tag{40}$$

In the sense of Hamilton's principle the action has an extremum during the motion

$$S = \int L \, \mathrm{d}V \mathrm{d}t = \mathrm{extremum} \tag{41}$$

i.e., the variation of action is zero:

$$\delta S = \int \left(\frac{1}{2} \left(\frac{\partial \varphi}{\partial t} \right)^2 + \frac{1}{2} \frac{\lambda^2}{c_v^2} (\Delta \varphi)^2 \right) \mathrm{d}V \mathrm{d}t = 0 \tag{42}$$

The resulting Euler-Lagrange differential equation gives the equation of motion (field equation) for the problem. The Euler-Lagrange equation for the Lagrange density function given by equation (40) is:

$$\frac{\partial^2 \varphi}{\partial t^2} + \frac{\lambda^2}{c_v^2} \Delta \Delta \varphi = 0 \tag{43}$$

which yields the heat equation (38), taking into account the definition equation (39) we use:

$$\frac{\partial T}{\partial t} - \frac{\lambda}{c_v} \Delta T = 0 \tag{44}$$

This equation describes slow energy transport, however, the finite speed of propagation of action is not required.

3.1.1 Poisson Bracket, Hamilton Equations

It is worth to define the Poisson bracket expression of variables, since it can be proved that the Poisson bracket of the Hamilton density function with a certain variable gives its time derivative [9]:

$$[F, \mathcal{H}] = \frac{\delta F}{\delta \varphi} \frac{\delta \mathcal{H}}{\delta p} - \frac{\delta F}{\delta p} \frac{\delta \mathcal{H}}{\delta \varphi} \tag{45}$$

$$\dot{F} = [F, \mathcal{H}] \tag{46}$$

As a consequence, the Euler-Lagrange equation can be substituted by these two Hamilton equations:

$$\frac{\partial \varphi}{\partial t} = [\varphi, \mathcal{H}] \tag{47}$$

$$\frac{\partial P}{\partial t} = [P, \mathcal{H}] \tag{48}$$

The advantage of this description is the appearance of the lower order differential equation rather than the more complex Euler-Lagrange equation; the solutions may thus, be easier. At the same time, the introduction of the canonical variables gives the chance to construct the phase field.

For a compete and consistent theory, not all of the Lagrange functions are suitable for a developed equation of motion, that can be deduced using variations. Thus, the necessary condition for the suitable Lagrange density function, is to ensure the Euler-Lagrange equation, as the equation of motion, deduced from the Hamilton's principle, but this is not enough for a complete and consistent theory. The necessary condition and the relevant choice of the Lagrange density function is required so that the canonical equations meet the equations of motion of the problem. Thus, the canonical equations are essential for the consistent elaboration of the formulization.

3.1.2 Canonically Conjugated Momentum and Hamilton Function

If we know the Lagrange density function, the canonically conjugated momentum to the field variable can be introduced which can be obtained after the general variation:

$$P = \frac{\partial L}{\partial \dot{\varphi}} = \dot{\varphi} = \frac{\partial \varphi}{\partial t} \tag{49}$$

After then the Hamilton function (the Legendre transformed of the Lagrange function) can be written [9]:

$$\mathcal{H}(P, \varphi) = \dot{\varphi} P - \mathcal{L} = \frac{1}{2}\left(\frac{\partial \varphi}{\partial t}\right)^2 - \frac{1}{2}\frac{\lambda^2}{c_v^2}(\Delta\varphi)^2 = \frac{1}{2}P^2 - \frac{1}{2}\frac{\lambda^2}{c_v^2}(\Delta\varphi)^2 \tag{50}$$

The fulfillment of the canonical equations can be easily checked by a short calculation:

$$\frac{\partial \mathcal{H}}{\partial P} = \frac{\partial \varphi}{\partial t} \tag{51}$$

$$-\Delta\frac{\partial \mathcal{H}}{\partial \Delta\varphi} = \frac{\partial P}{\partial t} \tag{52}$$

4 Elaboration of Formalism and the Directions of Applications

Considering the Lagrange density function can be formulated by non-exclusive self-adjoint operators, suddenly, more directions open in relation to the study of such processes, which were not possible with the Hamiltonian method. Previously, this effective theory and the related methods cannot be part of the developments and deep understanding of various physical processes.

The study of certain nonlinear processes can be part of the themes for potential Lagrangian theories. Such an example, for this description, is the Fourier heat conduction, with space, time and temperature – dependent conducting coefficients [10, 11]. Since the Lagrange formalism means a unified frame, there is the possibility to couple the different processes in a natural way. Here, not exclusively, the classic transports can be coupled with other areas of physics. This may have a special importance, because the concept of dissipation will be an integral part of the formulation. The coupling of the thermal and the cosmologic expanding field, the inflation field is a good example for this, in which, the expanding process is thermodynamically interpretable [12, 13]. The greatest challenge is also the most exciting field, the realization of coupling with the quantized processes.

One of the methods to involve the finite action speed in the theory, is to write the process with a telegrapher hyperbolic equation. Such equations are formulated for

transport processes given by linear differential equations. The consistent general canonical theory has been elaborated in [6, 14]. This also means that the same type of equations can be coupled.

An older tough problem was on how to formulate the transport equations in a Lorentz invariant form, in order to embed them in the theory of special relativity. The general solution is not known at the time of this writing. We can conclude that there is a successful description for the thermal energy propagation with the help of the potential theory [15, 16]. Accordingly, this means that the laws of thermodynamics are completed in the Lorentz invariant formulation [17, 18]. Naturally, the developed method involves the classical heat propagation as a limit.

The examination of general variations of action reveals the geometric – time and space shift, spatial rotation – and the dynamic symmetries of the studied theory. Each geometric invariance, leads to a conservation law, namely, energy, momentum and angular momentum. The dynamic invariances express a special, property dependent conservation (e.g. electric but in general any charges, lepton and barion number). The symmetry of the Lagrange density function of coupled linear transports, verify the validity of the Onsager's reciprocity relations [8, 11, 19, 20]. This poof is important because it is based on field theoretical considerations.

An essential result of the described formalism for the dissipative processes is that the concept of phase space can be constructed; the Liouville equation can be expanded for telegrapher equations. As a consequence of this, it can be shown that the Chapman-Kolgomorov equation is fulfilled, the Onsager's regression hypothesis is valid and the system fluctuations can be handled [21].

The introduction of the constructed canonically conjugated quantities of thermal energy propagation opens the way towards the adoption and application of quantum theoretical methods for the case of dissipative processes. In this way, a kind of excitation of thermal processes (hotons) can be understood. These excitations are particularly interesting since these are deduced from a real thermodynamic background [22, 23, 24, 25]. If this field can be successfully coupled with other fields in a consistent frame, then it is expected that the required criteria of thermodynamics – especially the second law of thermodynamics – will be the part of the description. This may clearly mean that the concepts of dissipation and irreversibility become directly interpretable, in cases of open quantum systems.

An interesting quantum thermodynamic application is the study the quantum properties; the examination of non-extensive interacting boson systems, and the treatment of Bose condensation. During the examination a relation can be recognized between the non-extensive nature and the interaction parameter. Furthermore, it can be shown how the commutation relation is modified, due to the internal interaction of the system [26, 27].

The existence of a description for the dissipative processes has a possible significance to retrieve the dissipation free case, as a limit. It involves and operates two kinds of propagation mechanisms at the same time. Thus, if the process dynamics change, e.g. the wave-like or ballistic propagation turns into diffusive, then, this transition can be correctly discussed. This may yield remarkable progress, mainly in the interpretation of complex processes [14, 28].

Summary

We have shown how to construct a Lagrange density function for those processes which are described by differential equations involving non-selfadjoint operators. Potentials (scalar and vector fields) can be introduced to the measurable physical fields, by which, a complete description and solution of the processes is possible. We have shown this method on a well-known example in electrodynamics. Then, we applied the theory to a simple dissipative process, the Fourier heat conduction. The advantage of theory is that the canonical formalism yields further directions in the study. Finally, we described more possibilities for further research areas. Until now, there are very promising results in these themes, but further studies are needed.

References

[1] F. Márkus, K. Gambár, A Variational Principle in Thermodynamics, *J. Non-Equilib. Thermodyn.* **16** (1991) 27

[2] K. Gambár, *„Least Action Principle for Dissipative Processes"* in Variational and extremum principles in macroscopic systems (eds.: S. Sieniutycz, F. Farkas), Oxford: Elsevier, 2005. pp. 245-266

[3] K. Gambár, F. Márkus, B. Nyíri: A Variational Principle for the Balance and Constitutive Equations in Convective Systems, *J. Non-Equilib. Thermodyn.* **16** (1991) 217

[4] K. Simonyi, *„Elméleti villamosságtan",* Budapest, 1981(*in Hungarian*)

[5] W. Gough, An Alternative Approach to the Hertz Vector, *Process In Electromagnetics Res.,* **12** (1996) 205

[6] D. Jou, J. Casas-Vázquez, G. Lebon: „*Extended Irreversible Thermodynam-ics*", Springer, Berlin, Heidelberg, 1993, 1996, 2001

[7] G. Lebon, D. Jou, J. Casas-Vázquez: *"Understanding Non-Equilibrium Thermodynamics"*, Springer, Berlin, Heidelberg, 2008

[8] S. Sieniutycz: *"Conservation Laws in Variational Thermo-Hydrodynamics"*, Kluwer, Dordrecht, 1994

[9] F. Márkus, K. Gambár: Hamilton's Canonical Equations and the Entropy Production, *J. Non-Equilib. Thermodyn.,* **18** (1993) 288

[10] K. Gambár, F. Markus, Hamilton's Principle for a Set of Nonlinear Heat Conduc-tion, *Open Sys. & Inf. Dyn.*, **12** (2005) 239

[11] P. O. Kazinski: Stochastic Deformation of a Thermodynamic Symplectic Structure, Phys. Rev. E **79** (2009) 011105

[12] F. Márkus, K. Gambár: Derivation of the Upper Limit of Temperature from the Field Theory of Thermodynamics, *Phys. Rev. E* **70** (2004) 055102

[13] F. Márkus, F. Vázquez, K. Gambár: Time Evolution of Thermodynamic Temperature in the Early Stage of Universe, *Physica A* **388** (2009) 2122

[14] K. Gambár, F. Márkus: A Simple Mechanical Model to Demonstrate a Dynamical Phase Transition, *Rep. Math. Phys.,* **62** (2008) 219

[15] F. Márkus, K. Gambár: Wheeler Propagator of the Lorentz Invariant Thermal Energy Propagation, *Int. J. Theor. Phys.* **49 (**2010) 2065

[16] F. Márkus: *„Can a Lorentz Invariant Equation Describe Thermal Energy Propagation Problems?"* in Heat Conduction (ed.: V. Vikhrenko), Rijeka: InTech, 2011. pp. 155-176

[17] T. Szőllősi, F. Márkus: Searching the Laws of Thermodynamics in the Lorentz-Invariant Thermal Energy Propagation Equation, *Phys. Lett. A* **379** (2015) 1960

[18] K. S. Glavatskiy: Lagrangian Formulation of Irreversible Thermodynamics and the Second Law of Thermodynamics, J. Chem. Phys. **142** (2015) 204106

[19] K. Gambár, K. Martinás, F. Márkus: F: Examination of Phenomenological Coefficient Matrices within the Canonical Model of Field Theory of Thermodynamics *Phys. Rev. E* **55** (1997) 5581

[20] K. Gambár, F. Márkus: On the Global Symmetry of Thermodynamics and Onsager's Reciprocity Relations, *J. Non-Equilib. Thermodyn.* **18** (1993) 51

[21] F. Márkus, K. Gambár, F. Vázquez, J. A. del Río: Classical Field Theory and Stochastic Properties of Hyperbolic Equations of Dissipative Processes *Physica A,* **268** (1999) 482

[22] F. Márkus, *„Hamiltonian Formulation as a Basis of Quantized Thermal Processes"* in Variational and Extremum Principles in Macroscopic Systems (eds.: S. Sieniutycz, F. Farkas), Oxford: Elsevier, 2005. pp. 267-291

[23] K. Gambár, F. Márkus: On an Unusual Quantization Procedure of Heat Conduction, *Open Sys. & Inf. Dyn.*, **83** (2001) 69

[24] F. Márkus, K. Gambár: Quasiparticles in a Thermal process, *Phys. Rev. E* **71** (2005) 066117

[25] F. Vázquez, F. Márkus, K. Gambár: Quantized Heat Transport in Small Systems: A Phenomenological Approach, *Phys. Rev. E* **79** (2009) 031113

[26] F. Márkus, K. Gambár: Q-boson System below the Critical Temperature *Physica A* **293** (2001) 533

[27] F. Márkus: On the Nonextensivity of Entropy of the Weakly Interacting Bose Systems, *Physica A*, **274** (1999) 563

[28] K. Gambár, F. Márkus: A Possible Dynamical Phase Transition between the Dissipative and the Non-Dissipative Solutions of a Thermal Process, *Phys. Lett. A* **361** (2007) 283

Extending System Capabilities with Multimodal Control

Gregor Rozinaj, Marek Vančo, Ivan Minárik, Ivan Drozd, Renata Rybárová

Institute of Telecommunications, Slovak University of Technology, Bratislava, Ilkovičova 3, 81219 Bratislava, Slovakia, gregor.rozinaj@stuba.sk, marek_vanco@stuba.sk, ivan.minarik@stuba.sk, ivan.drozd@stuba.sk, renata.rybarova@stuba.sk

Abstract: Multimodal interface (MMI) is the first layer from a user point of view to interact with most IT systems and applications. MMI offers natural and intuitive interface for user identification and system navigation. Typical features of multimodal control contain user identification based on face recognition and speaker voice recognition, system control based on voice commands and gesture recognition. Several examples show typical applications with MMI like voting or direct shopping while watching TV.

Keywords: multimodal interface; gesture control; gesture recognition; voice navigation

1 Introduction

At the turn of the 21st Century people were thrilled when they could get connected to the Internet on monochromatic displays of their newest flagship mobile phones. Similarly, the possibility to browse the teletext was considered a high level of interaction with the classic TVs. The past 15 years lead to a stunningly extensive progress of a human – computer interaction (HCI) in all possible kinds of electronic devices.

HCI means interaction between humans and machines via all possible input and output interfaces, i.e. keyboard, mouse, pen, gesture, speech, face, iris, etc. Multimodal interface represents all input and output interfaces based on human senses and inputs from a user, and so creates a natural way of communication and control for the user. The term modality refers to a human sense or a form of user input, for example face recognition is based on vision, speech recognition is based on speech, and gesture recognition is based on movement.

This article offers a proposal of multimodal interface focused on a system navigation. Section 2 includes a short introduction to the MMI. In Section 3 some

of the most common techniques and algorithms of voice and gesture recognition are described together with implemented algorithms for the gesture recognition in our system. Results for the tested algorithms are presented in this section, too. Section 4 contains the architecture of our proposed MMI application. At the end, in Section 5, the possible use of the proposed MMI is explained in various scenarios.

2 Multimodal Interface: Natural Entry to the System

Currently, the most widely used input devices for human–computer communication are keyboard, mouse, or touch tablet. These devices are far from the idea of natural communication with a computer, and rather represent human adaptation to computer limitations. In the last few years a requirement began to pop up that humans need to communicate with machines in the same way as they do with each other: by speech, mimics or gestures, since these forms conceive much more information than traditional peripheral devices are able to acquire. Our system focuses on this need by implementing and interconnecting several modalities to achieve a more natural control. This leads us to the term Multimodal interface [1].

The first step of communication with the Multimodal Interface (MMI) starts with user identification and authorization. Devices are aware of their legitimate users continuously and either adapt to them accordingly, or deny access to unauthorized users. Multiple modalities are available to control the system, each customized to user's personal preferences and habits.

User identification is typically based on a user name and password. Within the context of HBB-Next [2], a European research project, new standard [3] was developed where face [4], [5] and voice recognition are used as main identification approaches. However, other modalities, such as fingerprint recognition, iris recognition, etc., open the possibility of multi-level identification and authentication. System control, the second main part of the MMI, includes voice command navigation, gesture recognition, eye tracking, etc. Several examples show possible applications of MMI, such as voting or direct shopping while watching TV.

3 System Navigation

Focusing on system navigation, gesture and voice command recognition are the key modalities that allow for more natural interactions between humans and computers as they are relatively well examined and easy to implement from the

practical point of view. However, there are a few considerations that have to be taken into account.

The present gesture sets are based on physical input devices used with computers. Simply said, they try to "remove" the device, but keep the same usage patterns, mostly in order to avoid the learned gesture problem. In order to come closer to a natural (touch-less) gesture-based operation, the concept has to change so that gesture sets are designed bottom-up, like if there were no other devices than gestural sensors. Our team has examined several gesture recognition approaches, each serving a different purpose. By combining them, we aim to use the most suitable method for specific situations.

The one feature that is more obvious and expected by general users of an intelligent multimedia system is the voice navigation. Just like the gesture navigation, the voice navigation represents a natural interface between computers and humans. And just like the gesture navigation, the voice navigation's first requirement is to be intuitive and comfortable. This task seems less demanding when compared to gesture recognition, especially since voice recognition is not influenced by any device or sensor used to acquire the voice.

3.1 Intuitive and Natural Gesture Navigation

One of the greatest drawbacks of wider use of natural user interfaces is their lack of usability and human-centred design. While other modalities (i.e. the voice command navigation) seem to adapt rather quickly, the gesture recognition still cannot deliver truly natural experience, especially on touch-less devices. There are several factors that determine whether the gesture recognition is a natural and intuitive process. Firstly, there are the hardware limitations that limit sensor algorithm's ability to recognize more specific details in a gesture performance. This causes gestures to be recognized incorrectly and forces users to perform gestures that are not intuitive, require plenty of effort and lack comfort. System designers tend to overcome the sensor limitations by introducing gestures that are easily recognizable but are often far from simple.

Gestures can be divided into two basic categories by user experience. Innate gestures are based on the general experience of all users such as to move an object to the right by moving hand to the right, catch an object with closed fingers, etc. Naturally, the innate gestures can be affected by habits or culture. With the innate gestures there is no need for a user to study them in order to get good gesture experience, they just need to be shown to him. The second category is learned gestures, which need to be learned.

The gestures can also be divided into three categories based on the notion of motion [6]. Static gestures represent shapes created by gesturing limbs, which carry a meaningful information. The recognition of each gesture is ambiguous due to the occlusion of the limb's shape and, on the higher level of recognition, the

actual meaning of the gesture based on local cultural properties. The second category, continuous gestures serve as a base for an application interaction where no specific pose is recognized, but a movement alone is used to recognize the meaning of a gesture. Dynamic gestures consist of a specific, pre-defined movement of the gesturing limb. Such gesture is used to either manipulate an object, or to send out a control command. There is a problem with humans' inherent inability to perform a gesture in exactly the same dynamics, distance and manner. Additionally, these three groups can be combined in different ways, for example the static posture of a hand with the dynamic movement of an arm.

The general idea behind combining gesture methods is to utilize the best method for each individual action. Where a swiping is a natural approach, trajectory tracking should not be used, and where a simple static gesture serves well, a dynamic gesture would be a waste of resources. The time spent with gesture control relates to the amount of the energy spent, so if the application control requires more energy, then users will use less gestures and more traditional forms of control.

Neural networks and genetic algorithms were mostly used in the beginnings of gesture recognition. These methods had an acceptable recognition rate, but the greatest drawback was the amount of a necessary computing power and time needed for the training of neural networks which were significant, and unacceptably high for practical applications. Nowadays, different techniques are used to recognize gestures, since algorithms which do not require neuron networks have been invented, for example the Golden Section Search, the Incremental Recognition Algorithm and probabilistic models like the Hidden Markov Model (HMM). To increase the success rate of these algorithms, machine learning can be used. There are many approaches how to implement the gesture recognition. HMM methods are one of them, the main reason being that HMM approach is well known and used in many areas. One interesting approach how to implement the HMM into gesture recognition is shown in [7] where the author describes his own method step-by-step, which consists of:

- Gesture modelling
- Gesture analysis
- Gesture recognition

The author uses a Kinect v1 sensor as the input device. There are some problems with a centre hand point because of Kinect's inaccuracy. The starting and ending point of a gesture are determined by using a "static state", where the static state is accepted when the hand is kept relatively still. When a person performs a motion, this movement will be recorded and compared against a database. HMM in this paper was used for training and recognition only. Only basic gestures were tested such as "left", "right", "up", "down" and letters "S", "E", "O". The directions had a very good success rate but the remaining three letters had an average success

rate of about 90%, which is disappointing, given by that the database consists of seven gestures only.

Many approaches which use HMM scheme are based on RGB camera sensing. But just in the last three years the researches started to solve depth images using motion sensors [8], [9] very intensively.

Our research focuses on the gesture recognition area, where we want the user to be able to control all room equipment and devices via the gesture-controlled TV application. The gesture recognition should work robustly in changing light conditions. This is achieved by using the IR-based depth camera incorporated in the Kinect sensor as it has been shown it is susceptible only to strong sources of light [10]. In our MMI application three types of gestures are implemented: static gestures, dynamic gestures and swipe gestures which are a subdomain of dynamic gestures but employ a different algorithm. Each of these methods has several unique usages. Static gestures are used as an additional symbol for dynamic gestures, or as a symbol for the start and the end of dynamic gestures: if a user shows five fingers of a hand, the system allows him to perform dynamic gestures. If the user wants to end the dynamic gesture session, the user closes his palm. The static gesture can be used as a volume controller in combination with palm rotation.

3.2 Static Gestures

We researched several static gesture algorithms [10], finding the modification of Part-based Hand Gesture Recognition (HGR) algorithm [12] as being the most reliable. In our approach a binary image of a hand area is adjusted in order to obtain a convexity hull (polygon created by connecting all extremes around the hand) and its defects. The convexity hull determines a border between two different image parts. To accurately determine a centre of the palm, the author of [12] applied an inner circle, which brings several problems, like false detection or higher computational power needed in some hand postures. To avoid this, a circumscription which is more robust against hand tilt is used. To cope with hull shapes with extreme convexity a point onto contour hull is added that belongs to the hand and has the maximal distance from the found defect.

We implemented our own method to omit the forearm area. The circle is created with centre being the centre of a palm. Then, the two intersections are found with the contour closest to the Kinect-detected elbow point, taking the shortest distance between the centre of a palm and the found points as the circle's radius as given by:

$$Min\ (pointA, pointB) \tag{1}$$

Where *pointA* and *pointB* are intersections of the circumscription and the contour.

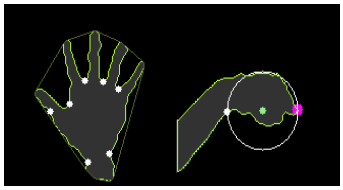


Figure 1
Finding the centre of the palm

The function of two variables is used for the representation of a hand shape. A hand contour is mapped onto X-axis, and Y-axis then describes the relative distance of each point from the centre of the palm (see **Hiba! A hivatkozási forrás nem található.**). Although this implementation is not trivial, its result is easily readable and clearly shows the hand proportions. A search for local maxima and minima is performed as a part of the contour analysis. The first and last local extremes must be local minima; otherwise local maxima at the beginning and end are removed. We modified the original implementation because it caused loss of some important higher relative distance extremes, and was ineffective for lower relative distances.

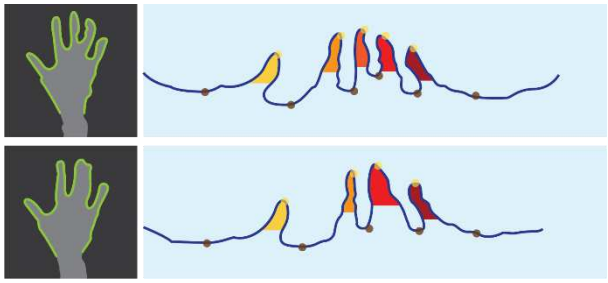


Figure 2
Static gestures are recognized by counting fingertips on a curve spread from the hand contour. Combination of static and dynamic/swipe gestures will let people use more natural gestures over traditional forms of control.

A set of tests was performed for the described method comparing with the methods [13], [14]. We attempted to estimate the complexity of each algorithm in terms of the number of the specific steps needed to obtain a recognized gesture. This information leads to a processing power and delay limitations. The maximum distance from the sensor was measured. Then, the rotation boundaries of the hand were measured in which the algorithms are able to perform reliably. The level of freedom describes the relative distance between the fingers in order to be recognized as separate fingers (0% for max. distance and 100% for joined fingers). This is supported by information about joined fingers detection based on the algorithmic properties of each method. In order to measure the success rate, the gestures were performed 110 cm from the sensor. Four subjects were tasked to

perform 100 gestures by showing the different number of fingers in various hand positions, creating a test set of 400 gestures. The results are summarized in Table 1 and Table 2.

Table 1
Algorithm Comparison [10]

	Convexity Defects	K-Curvature	Part-based HGR
Sensor Distance 80cm to	119 cm	160 cm	175 cm
Algorithm Complexity # specific steps	7	6	7
Joined Fingers Detection	NO	NO	YES
Relative Level of Freedom	40%	70%	95%
Success Rate	80%	92%	90%

Table 2
Algorithm Comparison – Hand Rotation limitations for each algorithm[1] [10]

	Hand Rotation Boundaries					
Convexity Defects	X axis		Y axis		Z axis	
	35°	75°	25°	30°	180°	65°
K-curvature	X axis		Y axis		Z axis	
	35°	75°	25°	40°	175°	170°
Part-based HGR	X axis		Y axis		Z axis	
	50°	85°	25°	40°	150°	125°

Of the three evaluated algorithms, the *Convexity Defects* approach has proved to be the least reliable. Even though the success rate could be considered acceptable, it was very susceptible to a noisy input (given by hand rotation boundaries and a level of freedom). The *K-Curvature* [13] method provided the best results in terms of overall success rate. Additionally, this method is applicable to the widest range of the possible hand rotations from the trio. However, the level of freedom is the bottleneck of the approach. The third analysed method proved to be reliable and is the most robust of the three. With its alternative approach to counting of fingers it is able to distinguish even joined finger given the input image falls within the rotation boundaries. As it was shown each of the methods can be quite reliably used for the static gesture recognition, when in compliance with each method's unique properties.

[1] All angles are relative to default hand position: open palm with index finger in line with Y axis. In X axis, the first angle describes rotation heading front, the other heading back. In Z axis, first angle is rotation counter-clockwise, second angle clockwise (as viewed by the performer). In Y axis, first angle describes rotation to the left, second to the right.

3.3 Dynamic Gestures for Better Interaction

Dynamic gestures are used to provide an authorization to a private content. They are used as a password key. A user can perform a dynamic gesture and the likeliness of the template and the performed gesture are compared via the use of an incremental recognition algorithm proposed by Kristensson and Denby [15], originally designed for digital pen strokes and touch-screen devices. For this approach, a template is defined as a set of segments describing the template gesture. Each segment describes progressively increasing parts of the template gesture so that the first segment is a subset of the second segment, which is a subset of the third segment, etc., and the last segment represents the whole gesture template. Each segment is represented as a series of time-ordered points.

With each new point of the observed gesture the system computes a Bayessian posterior probability that the gesture matches a gesture template, for each template, as given by the formula:

$$P(\omega_j|\mathrm{I}_i) = \frac{\mathrm{P}(\omega_j)P(\mathrm{I}_i|\omega_j)}{\sum_k \mathrm{P}(\omega_k)P(\mathrm{I}_i|\omega_k)} \qquad (2)$$

where $P(\omega_j)$ is the prior probability, $P(\mathbf{I}_i|\omega_j)$ is the likelihood and the denominator is the marginalization term. The prior probability can be used to influence the posterior probability when the distribution of the template probabilities is known. For example, if the probability of each gesture occurrence is known then more precise and successful recognition may be obtained.

The likelihood measure is given as a probability that the part of the observed gesture matches a gesture template:

$$P(\mathrm{I}_i|\omega_j) = P_l(\mathrm{I}_i|\omega_j)E(\mathrm{I}_i|\omega_j), \qquad (3)$$

where $P_l(\mathbf{I}_i|\omega_j)$ is the likelihood of the observed gesture and the respective part of gesture template. It is given as the max of the distance function D taking into account the Euclidean distances between the normalized points of the observed gesture and template segment, and the turning angle between the two point sequences. $E(\mathbf{I}_i|\omega_j)$ is an end-point detection term which serves to favour the complete gestures compared to the parts of the gestures in the case when one full template represents a part of a different template.

The distance function is given by formula:

$$D(I,S) = \exp\left(-\left[\lambda\left(\frac{x_e^2}{\sigma_e^2}\right) + (1-\lambda)\left(\frac{x_t^2}{\sigma_t^2}\right)\right]\right) \qquad (4)$$

The distance function depends on both Euclidean distance x_e between the corresponding points of the recorded trajectory **I** and the known template **S**, and the mean turning angle x_t between the respective line segments of the **I** and **S** sequences. The contribution of the two measures is managed with the variable λ which allows to favour one of the measures against the other.

The posterior probabilities are then filtered using a window over last five predictions to stabilize them. The interested reader may find a more detailed description of the original algorithm in [15].

For our purposes, the algorithm was altered to make use of the depth data provided by the Kinect sensor. The gesture recognition process is triggered by the user's hand movement and the movement's trail is examined in real time by comparing it with the parts of the predefined gesture templates. The set of templates is compared to the performed gesture in real time and templates that do not match the performed sample with pre-set certainty are continuously removed from the set. In this way the algorithm provides a decision on which gesture was performed with decreasing ambiguity, until only one template gesture remains having the highest probability. It is obvious that given a set of gestures which are sufficiently distinguishable from each other the recognition may be successful after only a part of the gesture was performed. An application was created to test the proposed algorithm. The application includes a set of gesture templates which can be extended with custom gestures. Gesture recording works as follows. The Kinect sensor input is used to obtain a trajectory of the performed gesture, consisting of individual points. The trajectory is then reduced in size to fit in 1000x1000 points and saved to the template group.

The default set of gestures consists of capital letters of English alphabet (26 letters). Each gesture was performed five times by four persons, creating a test set of gestures consisting of 520 gestures. It is important to note that not all of the gestures were performed with high accuracy. As opposed, some of the gestures were performed imprecisely and inconsistently with the attempt to examine the gesture variability. On this data set the overall success rate was above 91%. The average distance of users from Kinect, which is an important parameter when considering touch-less environments, was approximately 1.8 meters. Some limits of Kinect sensor by testing were determined. Kinect's accuracy decreases with the growing distance between the sensor and users. This argument is also confirmed in [16], [17]. Smoothing into individual joints was applied to eliminate low accuracy. Then we obtained slightly smooth curves on our imaginary surface. This improvement helped us to achieve more successful results and remove some problems which originated from bad accuracy.

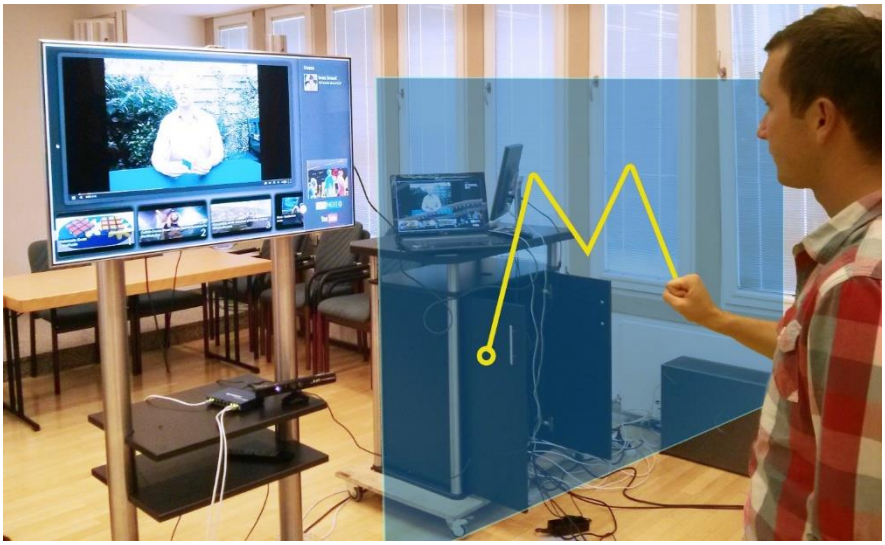


Figure 3
The incremental recognition algorithm recognizes the gesture as it is being performed on a virtual surface, simulating a touch panel of a tablet. System works with the gesture similarly to smartphone unlocking pattern.

The testing equipment consists of the hardware fulfilling the minimal hardware requirements for the Kinect v1 sensor, namely Windows 7 or Windows 8.1 operating systems (tested on both systems), dual-core 2.66 GHz CPU, 2GB of RAM and USB 2.0 connector for the sensor connection. The software used with the Kinect sensor is Kinect SDK 1.8 and EmguCV 2.4.2.

The usage of gestures is extended by swipe gestures. This gesture type brings in a very natural and comfortable approach. Swipe gestures are designed for fast and routine browsing in the menu, programs, or gallery as they consist of 4 directions of hand movement for each hand and combinations of both hands. Our method called Circle Dynamic Gesture Recognition (CDGR) is based on hand detection, speed of movement and distance (see Figure 4). While the hand is in an inner circle, the system stays inactive. After the user crosses the inner circle a short countdown is started. During the countdown the system observes if the user's hand crosses the outer circle. If the countdown reaches a limit before the hand crosses the outer circle, the method will reset and the hand will be again in the center of both circles. So, if the user moves his hand slowly, both circles will follow its joint and no gesture will be recognized. If a human hand executes a faster motion and the inner circle leaves the outer circle, the system processes this motion and determines a gesture.

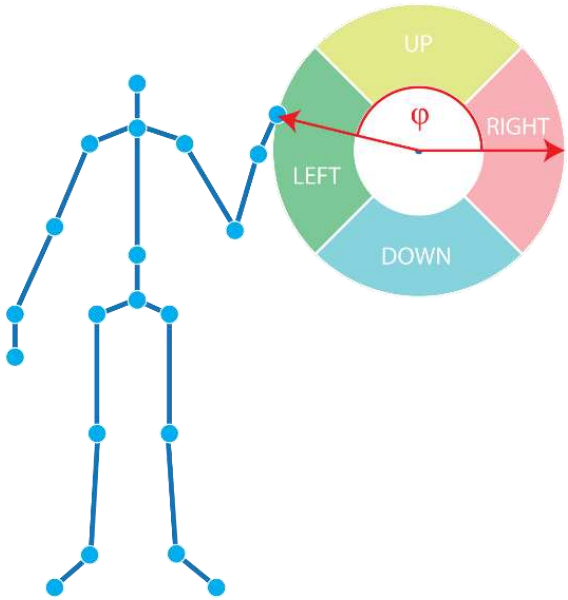


Figure 4
Swipe gestures are fast to perform and reliable to read

The gesture is given by the angle of the performed motion from the middle to the outer circle. Initially, the possible gestures are: swipe left, right, up and down. The gestures can be performed by both hands individually, or as a combination. This allows the user to perform more complex gestures, such as zoom in and zoom out.

During the testing 10 people performed a success test; each person performed 40 gestures, creating a set of 400 performed gestures in total. Our algorithm has proved reliable with 94% success rate for four defined gestures for each hand. The big advantage of the swipe method is its low computational complexity, high precision and easy implementation for many purposes.

3.4 Voice Commands Navigation

Thanks to its complexity and ability to convey deep and thoughtful message in simple form, speech recognition could be one of the most comfortable ways of natural interaction between humans and computers. In the last few years, there has been a considerable leap in development as processing power ceased to be the limiting factor for using advanced algorithms. This is mainly visible in the field of smart personal devices where the biggest players like Google, Apple and Microsoft introduced their personal assistant services. One of the key advancements in their technology, from usability point of view, is shift from command based to conversational based control. In the past, voice navigation was limited by the processing speed of the local device which lead to limited

recognizable vocabulary. This yielded either a few-command set to cover the general domain or a highly limited command domain. Extending the command set usually lead to higher ambiguity given finite set of the describing parameters. This has changed with moving the processing power to the cloud where there are virtually unlimited computational resources as well as storage.

Other restraints of voice commanding include variable and unpredictable acoustic conditions which allow reliable voice command recognition only in a controlled environment (no outside sounds etc.). Also, the success rate of recognition is highly user-dependent: apart from different accents or dictionaries of each individual, even the same person doesn't say the command in exactly the same way twice. This could be avoided partly by training the recognizer by the voice of the person, either before or during usage.

The voice command recognition, just like any other recognition, consists of two principal steps. During learning, the system has to be taught what inputs it may expect and what they mean. Secondly, during recognition, an unknown input pattern is presented and a closest match from the learned pattern set is chosen. Both steps are demanding either in terms of data quality and quantity (learning phase) or quality and speed (recognition phase). Additionally, the concept of continuous learning while recognizing is a logical enhancement of the original 2-step process.

There are currently a number of methods available that can be applied for voice command recognition, which differ greatly in approach as well as complexity. For example, Dynamic Time Warping, which is mentioned further in the text, or Hidden Markov Models that lead to good recognition success rates (up to 98.9 %, as in [18]) without being too demanding no computational power. Other, more complex approaches include neural networks, which experience a renaissance these days as computational power, storage and fast network connection are easily available. Namely, it is techniques like Deep Belief Networks, Convolutional Neural Networks or the late Hierarchical Temporal Memory which aim at modelling and learning relationships between features in the input signal in space and/or time.

With computational options increasing and becoming widely and easily available modern algorithms can look not only at the traditional properties of speech input but also on the more delicate features that had to be omitted before: emotion and context. Both features contain plenty of additional information that give humans the higher idea of the meaning of respective commands. For example, there is difference between prescriptive and calm tone of voice, just as there is difference whether the same word is heard within a fluent speech or the same word is uttered isolated. This is the area where neural networks now play an important role.

In our multimodal interface we implemented voice command recognition based on MFCC and DTW algorithm. Because range of commands used in MMI is not so wide and in different sections of MMI different groups of voice commands are

used (max 10 commands per area), it was not necessary to employ more advanced classification algorithms. We find DTW algorithm to be sufficient for our purposes.

Extraction of best parametric representation of human voice is one of the most important parts to achieve good recognition performance. In our case we decided to use MFCC coefficients. MFCC coefficients are a representation of the short-term powered spectrum on a non-linear mel-scale of frequency. Human auditory system is not linear and mel frequency scale fits it much better than linear frequency scale. The relationship between mel and linear frequency scale is given by (5):

$$F(mel) = 2595 * log(1 + f/700) \tag{5}$$

We used 13 MFCC coefficients plus delta and delta-delta features (39 coefficients together). Since MFCC coefficients represent only power spectral envelope of the time frame, but there is also information in spectral variation, we used delta and delta-delta features. The delta coefficients can be calculated as follows (6):

$$\Delta c(m) = \frac{\sum_{i=1}^{k} i * \left(c(m+1) - c(m-1)\right)}{2 * \sum_{i=1}^{k} i^2} \tag{6}$$

The same formula is used for delta-delta coefficients calculation, where MFCC coefficients are replaced with delta coefficients [19]. On these features DTW algorithm mentioned above was applied. DTW is a computationally inexpensive algorithm to measure the similarity between two temporal sequences which may vary in time or speed. In general, this approach calculates an optimal match between 2 given sequences with certain restrictions.

In our consideration, voice command recognition will be applied maximally on 10 commands in one section. In our testing 10 commands were tested in 200 experiments. The success rate which was achieved was 95%, which is sufficient for our purposes.

4 Multimodal Application

In our application research, we focused on natural multimodal interface and its integration into a multimedia system used on daily basis. The vision of the system is to control the TV and access multimedia content using larger number of modalities. Obviously, the usage of multimodal interface is not limited only to the TV system but has many different applications.

Figure 5
Logic behind the schematic of the multimodal interface shows applications served by the MMI controller which collects recognition information from individual modalities. All of them use input from the Kinect v1 sensor delivered by the MMI input hub.

The block diagram (see Figure 5) shows the concept of the multimodal interface divided into five layers. Physical layer represents hardware input and output devices which enable interaction with the real-world. The input device is currently represented by the Kinect sensor. Kinect is a multifunctional device which can be effectively used by each of the modalities mentioned above, for example, a microphone array for speaker identification, depth camera for gesture recognition, RGB camera for face recognition etc. Multimodal data provided by the Kinect sensor are collected by the HUB which serves as a distribution point of the input data from the Kinect sensor to multiple applications each utilizing different modalities. This is due to the technical limitation that allows the Kinect to communicate with only one application at a time. Modalities described in previous sections are represented as modules with defined APIs. Data obtained from Kinect sensor are then processed in parallel by each module separately. The modular, API-based structure allows to simplify the process of adding new modalities. The MMI controller collects output data from all modules, evaluates and combines it into one output data stream. The stream contains information about recognized users and their requested actions. Applications only depend on MMI controller

output so there is no limit in installing new applications, thus extending overall MMI functionality.

The currently implemented application consists of three micro-applications that cooperate to produce a UI on the TV screen. The first application is designed for video playback and can be easily maximized to the full screen using a simple gesture. The second application, located on the right side of the display, shows a list of users identified by speech or face recognition modules. Only users in this list are permitted to control the TV using predefined set of gestures, voice commands or other modalities, and combinations thereof, such as gestures with voice commands. When the user leaves the room, he/she is automatically removed from the list. The third application displays a list of recommended channels. Depending on user viewing preferences, system provides recommendations that best suit all users in front of the TV. Using swipe gestures a user is able to navigate this list, play or stop the video. To demonstrate the security possibilities of the system, some of the recommended channels are locked. It means that users without permission are not allowed to watch such content until they enter the secret pattern. To enter the secret pattern, we apply dynamic gestures.

In order to make the best use of multimodal interface, it is not always necessary to use touch-less gestures to perform every action. Some actions will always be better executed by using a different modality. I.e. entering text would be difficult, time consuming and by all means uncomfortable using gestures, but can be easily and faster performed with speech recognition. With this in mind, it becomes necessary to introduce an integration platform that will provide applications with requested inputs where the application does not need to know the source modality, if not required explicitly.

Within our research we have designed and implemented a multimedia system making use of several of the modalities mentioned earlier. Namely, the system uses face recognition and speaker identification for user authentication, and swipe gestures, dynamic gestures with static postures and voice command recognition for system control. In order to test the system as a whole, we have devised several use case scenarios where each of the modalities is employed. Thanks to the proposed layered model design, different applications may use different modalities. The modular structure allows for easy deployment of new applications like new ways of TV and room control, multi-device support, controlled access, etc.

5 How to Use It: Scenarios

A system that is aware of its users, knows their habits and interests, can become an intelligent concierge of the household, and can provide advanced interconnections between various services. Here we present only a few ideas of

the possible applications. Some of them are already in use with other being most certainly proposed.

5.1 Shopping while Watching

The dream of teleshopping is becoming true as connected TVs allow to make orders from the TV seat. Going a bit further, the next generation of TV shopping will happen (if not happening already) directly during watching the program. Broadcasters or 3rd party providers annotate the TV program with offerings of products and services related to the program. This additional information displays to the viewers as an optional information giving them the possibility to directly order that nice couch, brand of beer or skirt worn by their favourite actress as they see it in action on screen. Similarly, viewers can schedule for various medical procedures, or apply to subscription services. All is available at a pressing of a button or waving a gesture.

5.2 Smart Household

The integrated TV system can reach beyond the recommendation of TV channels and become the central information and operation hub of the whole household. Family members can get notified of any events that happen in the house, be it washing machine alerts, fridge notifications that the champagne is ready-chilled, or new mail in the mailbox. Additionally, the whole household can be operated from the comfort of the living room, no matter if you want to heat the room up a bit, close curtains or order groceries for delivery. Such system however requires home appliances that are interconnected with the TV system, which should not be a problem in the near future as connected appliances are already in production by several manufacturers.

5.3 Voting

Nowadays, people want to spend their time in front of the TV effectively. They watch preferred channels, programs, TV shows, that are recommended by their friends, family or colleagues. TV programs' or films' ratings are usually available in public databases such as IMDb.com, where the rating is provided by individual viewers. However, to access the rating one has to quit watching the TV and switch to the website in order to rate or obtain the rating of the programme. A more sophisticated system can collect this information immediately after the program finishes, while the viewer executes only a simple gesture. A like/dislike gesture (thumb up/thumb down) or swipe-left/swipe-right gestures may be used to obtain the rating. The main idea is to use very simple and very easily executable gesture, as users do not want to execute complicated gestures while relaxing. After watching the TV, the system automatically invites the viewer to rate the programme, and the viewer can execute the easy gesture in two seconds. The

system collects these opinions and creates global statistics and recommendations for the viewer's friends and family, as well as for the general audience, with the possibility to update the already existing rating systems.

5.4 Digital Doorman

Digital doorman is a feature that helps to keep watching TV a comfortable experience. The typical situation presents a user watching his/her favourite programme while a guest rings at the front door. Usually the user has to interrupt watching, stand up from the sofa and go to check the door phone. It is very likely that he/she will miss a short part of the favourite program. Digital doorman brings the highest comfort utilizing interconnection between multimodal interface and the doorman. In the upper corner of the screen the user can see a live camera stream from the doorbell and immediately can allow or deny the access to the building using an easy gesture or a voice command.

5.5 Phone Pickup

Another application that extends the functionality of the multimodal interface is called phone pickup. It often happens that during watching an exciting TV programme or just having fun with friends a phone suddenly starts to ring, and it is difficult or not comfortable to answer it. Multimodal interface simply enables to pick up or cancel the call in the comfort of the living room, share calls with friends etc. This functionality can be simply achieved by recognizing the phone's owner in front of the camera and offering a remote phone pickup. A phone call can be easily picked up using a voice command or a particular gesture command. The main advantage of this extension is pausing the TV channel playback during the phone call without losing comfort.

Conclusions

In this article we proposed a multimodal interface architecture with implemented voice and face recognition, gesture recognition, and voice command navigation.

Gesture recognition methods, discussed in this article, offer high reliability and can be used in a wide range of applications. The presented results show highly satisfactory recognition efficiency of the third presented method compared to the results of the other two methods for static gesture recognition, and can be applied in practical applications. A very good rate was achieved also by our own method Circle Dynamic Gesture Recognition for swipe gestures. The methods with the highest reliability were implemented into the multimodal interface for system control. We suggest that with the proper configuration of the presented methods a more intuitive gesture navigation can be achieved.

In the section devoted to multimodal applications we introduced a concept of a modular architecture for the multimodal interface. This architecture consists of

five layers with well-defined interfaces between each other. This is in accordance with the Kinect sensor being used as the core device with all the limitations implied by its APIs. Thanks to the modular architecture, the multimodal interface can be easily extended using additional modalities, input devices and micro-applications.

Despite the number of advanced features integrated in the Multimodal Control prototype application, further research is required not only in the area of more sophisticated modalities but also in the implementation of a more complex concept of the whole system. We investigate the possibilities to use the MMI in a complex intelligent room comprising multimodal control of other smart devices like light switches, sockets, air conditioning, etc. The multilevel authorization module needs to be extended for biometric methods and to consider advanced solutions such as identification via mobile devices, NFC tags or RFIDs and many others.

In near future, we plan to extend the whole system with an administration module for an easy and intuitive appearance and personalization of the application.

Acknowledgement

The authors hereby declare that the research leading to his article has been funded by the grant VEGA-1/0708/13 IMUROSA and APVV-0258-12 MUFLON.

References

[1] S. Oviatt.: Multimodal Interfaces. In The Human-Computer Interaction Handbook, Julie A. Jacko and Andrew Sears (Eds.) L. Erlbaum Associates Inc., Hillsdale, NJ, USA 286-304

[2] https://web.archive.org/web/20150711074144/http://www.hbb-next.eu/

[3] ETSI TS 102 796 v1.2.1, Hybrid Broadcast Broadband TV, European Telecommunications Standards Inst., 2012; www.etsi.org/deliver

[4] M. Oravec: Biometric Face Recognition by Machine Learning and Neural Networks, invited paper, The 5th International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014 / International Conference on Prediction, Modeling and Analysis of Complex Systems NOSTRADAMUS 2014, June 23-25, 2014, Ostrava, Czech republic

[5] M. Oravec, J. Pavlovičová, J. Mazanec, Ľ Omelina, M. Féder, J. Ban, M. Valčo, M. Zelina: Face Recognition in Biometrics (Metódy strojového učenia na extrakciu príznakov a rozpoznávanie vzorov 2: Rozpoznávanie tvárí v biometrii), Publisher Felia, Bratislava, 2013, ISBN 978-80-971512-0-1

[6] Vanco, M.; Minarik, I.; Rozinaj, G., Dynamic Gesture Recognition for Next Generation Home Multimedia, in ELMAR, 2013 55th International Symposium, pp. 219,222, 25-27 Sept. 2013

[7] Y. Wang, C. Yang, X. Wu, S. Xu and H. Li: Kinect Based Dynamic Hand Gesture Recognition Algorithm Research, Proceedings of the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012, Vol. 1, pp. 274-279

[8] K. Lai, J. Konrad and P. Ishwar: A Gesture-driven Computer Interface using Kinect, Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium, 2012, pp. 185-188

[9] N. Villaroman, D. Rowe and B. Swan: Teaching Natural user Interaction using OpenNI and the Microsoft Kinect Sensor, Proceedings of the 2011 Conference on Information Technology Education, 2011, pp. 227-232

[10] Khoshelham K, Elberink SO: Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. Sensors (Basel, Switzerland). 2012;12(2):1437-1454. doi:10.3390/s120201437

[11] Vanco, M.; Minarik, I.; Rozinaj, G., Evaluation of static Hand Gesture algorithms, in 2014 International Conference on Systems, Signals and Image Processing (IWSSIP), 2014, Vol., No., pp.83-86

[12] Z. Ren, J. Yuan, J. Meng and Z. Zhang: Robust Part-Based Hand Gesture Recognition Using Kinect Sensor, Multimedia, IEEE Transactions on Vol. 15, No. 5, 2013, pp. 1110-1120

[13] F. Trapero Cerezo: 3D Hand and Finger Recognition using Kinect, available at http://www.scribd.com/doc/161562314/Finger-and-Hand-Tracking-With-Kinect-SDK-3

[14] M. Vančo, I. Minárik and G. Rozinaj: Gesture Identification for System Navigation in 3D Scene. In Proceedings ELMAR-2012: 54th Symposium ELMAR-2012, 12-14 September 2012 Zadar, Croatia. Zadar: Croatian Society Electronics in Marine, 2012, s.45-48. ISBN 978-953-7044-13-8

[15] P. O. Kristensson and L. C. Denby: Continuous Recognition and Visualization of Pen Strokes and Touch-Screen Gestures, Proceedings of the Eighth Eurographics Symposium on Sketch-based Interfaces and Modeling, 2011, pp. 95-102

[16] K. Khoshelham: Accuracy Analysis of Kinect Depth Data, in ISPRS Workshop Laser Scanning, 2011, Vol. 38, p.

[17] B. Molnár, C. K. Toth, and A. Detrekoi: Accuracy Test of Microsoft Kinect for Human Morphologic Measurements, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 3, 2012, pp. 543-547

[18] J. Kacur and V. Chudy: Topological Invariants as Speech Features for Automatic Speech Recognition, International Journal of Signal and Imaging Systems Engineering, Vol. 7, No. 4, 2014, pp. 235-244

[19] Sreenivasa Rao, K.; Nandi, D.: Language Identification Using Excitation Source Features, Springer International Publishing, 2015, ISBN 978-3-319-17725-0

An Extension of Maximal Covering Location Problem based on the Choquet Integral

Aleksandar Takači[1], Ivana Štajner-Papuga[2], Darko Drakulić[3], Miroslav Marić[4]

[1]Faculty of Technology, University of Novi Sad, Bulevar Cara Lazara 1, 21000 Novi Sad, Serbia; atakaci@uns.ac.rs
[2] Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg D. Obradovića 4, 21000 Novi Sad, Serbia; ivana.stajner-papuga@dmi.uns.ac.rs
[3]Faculty of Philosophy, University of East Sarajevo, Alekse Šantića 1, 71420 Pale, Bosnia and Herzegovina; ddrakulic@ffuis.edu.ba
[4] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Beograd, Serbia; maricm@matf.bg.ac.rs

Abstract: The aim of this paper is to demonstrate the applicability of the Choquet integral, a well-known fuzzy integral, in the Maximal Covering Location Problem (MCLP). Possible benefits of the used integral, which is based on monotone set functions, include the flexibility of a monotone set function, which is in the core of the Choquet integral, for modeling the Decision Maker's behavior. Various mathematical models of the Maximal Covering Location Problem are given. The approach, based on the Choquet integral versus the standard approach, is thoroughly discussed and illustrated by several examples.

Keywords: Maximal Covering Location Problem; monotone set function; Choquet integral

1 Introduction

Many problems from the real world contain different uncertainties, ambiguities, and vagueness, so their mathematical models obtained with the classical mathematical techniques are not fully accurate. Fuzzy sets and different probabilistic methods are the most frequently used techniques for modeling problems from the real world. This study introduces a new method for modeling the Maximal Covering Location Problem (MCLP) by using a well-known fuzzy integral, the Choquet integral.

The Maximum Covering Location Problem (MCLP) was defined by Church and ReVelle in [5] and it represents a very important class of problems in operations research. They defined MCLP as follows: "Maximize the coverage within a

desired service distance *S* by locating a fixed number of facilities". In other words, the aim of MCLP is locating facilities on a given network in such a manner that they cover as many locations as possible. This class has a decisive role in many real world problems, such as locating shops, gas stations, bus stations, hospitals and other emergency services. Similar classes of problems include the Location Set Covering Problem (LSCP) and Minimal Covering Location Problem (MinCLP). The aim of LSCP is to cover all locations with as few facilities as possible, and the aim of MCLP is to cover as many locations as possible with a fixed number of facilities. In all models, the networks are represented by distances between locations (or travel times between them). Location coverage depends on the distance (or travel time) to the nearest facility and it depends on the given value called the coverage radius. In the classical case, those values are represented by real numbers, but in the "real-world" problems, those values are not fully determined, and they can contain different levels of vagueness, e.g., "the coverage radius is between 10 and 20 kilometers", "the travel time is around 20 minutes" or "it is pretty close". These linguistic ambiguities can be modeled by using different types of fuzzy numbers. Many authors developed different fuzzy MCLP models (FMCLP) and the most common approach is using fuzzy numbers for the radius of coverage. In the classical model of MCLP, each location is either covered or uncovered, while in FMCLP models, the locations could also be partially covered. The main question of FMCLP is how to treat partially covered locations. Depending on the nature of the problems, the degree of location coverage could be calculated using *t*-norms and *t*-conorms ([16]). This study takes into consideration another issue, namely the interaction between facilities which need to be optimally arranged. Now, the Choquet integral is being used in order to take into the account the different interactions between facilities which should yield a better quality solution.

This paper is organized as follows: in Section 2, a brief literature overview related to MCLP, FMCLP and the usage of fuzzy sets in location problems is presented. Section 3 includes certain basic mathematical notions, such as fuzzy sets, fuzzy numbers and the Choquet integral are given. Section 4 contains a new model of FMCLP based on the Choquet integral, while the last section offers some concluding remarks.

2 Literature Overview

As mentioned above, MCLP was developed by Church and ReVelle (1974) [5]. Different MCLP models were presented in the following years, like MCLP on the plane (Church, 1984 [6]), capacitated MCLP (Current and Storbeck, 1988 [7]), probabilistic MCLP (ReVelle and Hogan, 1989 [20]) and implicit MCLP (Murray et al. 2010 [17]). An exhaustive review of the covering problems and MCLP can be found in [12].

In recent years, several fuzzy models for the covering location problem have been presented. Darzentas in [8] presented a discrete location problem with fuzzy accessibility criteria and formulated it with an application of the set partitioning type of integer programming. Perez et al. in [19] presented that position of the facility in real applications can be full of linguistic vagueness, and they modeled them by using networks with fuzzy values. These fuzzy values appropriately describe the network nodes, lengths of paths, weight of nodes, etc. Batanovic et al. in [1] described the application of fuzzy sets in modeling the maximum covering location problems for networks in uncertain environments. They modeled distance (traveling times) from a facility site to demand nodes by fuzzy sets. Davari et al. in [9] presented a MCLP model with fuzzy variables for travel times for any pairs of nodes.

3 Definitions and Preliminaries

3.1 Discrete Choquet Integral

Since this discreteness is highly tangible in applications, the short overview of the discrete case, i.e., basic information on the discrete Choquet integral is given in this section.

It has to be emphasized that this type of integral is a highly applicable aggregation operator (see [11,14]). The Choquet integral generalizes the so-called additive operators, e.g., the OWA (the ordered weighted averaging operators, see [23]) and the weighted mean.

The first necessary notion is the one of a fuzzy measure. Firstly, let X be a set of criteria, that is, let it be a set of all input values.

Definition 3.1 *A set function* $\mu: P(X) \longrightarrow [0, \infty)$ *is a fuzzy measure, if the following it satisfied*

- $\mu(\emptyset) = 0$,
- *for arbitrary* $A, B \in P(X)$*, if* $A \subset B$ *then* $\mu(A) \leq \mu(B)$ *(monotonicity).*

Now, the triplet $(X, P(X), \mu)$ is a fuzzy measure space ([2, 3, 13, 22]). In general, instead of $P(X)$ some σ-algebra of subsets of X can be used.

As already mentioned, for the purpose of this research the focus is on a discrete case, i.e., on simple functions **-** functions that can assume only a finite number of values. Therefore, the following form of functions will be observed

$$f: X \longrightarrow \{\omega_1, \omega_2, \dots \omega_n\},$$

where $\omega_i \in [0, \infty)$ and the working assumption, with no influence on generality, is $0 \leq \omega_1 < \omega_2 < \cdots < \omega_n \leq 1$.

Moreover, based on the type of problems that will be investigated in the future, it is sufficient to observe simple functions with values in $[0,1]$ and normalized fuzzy-measures, i.e., further it will be assumed that $\{\omega_1, \omega_2, \dots \omega_n\} \subseteq [0,1]$ and $\mu: P(X) \longrightarrow [0, \infty)$ is a fuzzy measure.

The definition of the Choquet integral for the discrete case follows ([4]).

Definition 3.2 *The Choquet integral of an arbitrary simple function* $f: X \longrightarrow \{\omega_1, \omega_2, \dots \omega_n\}$*, based on a fuzzy measure has the following form*

$$(C)\int_X f \mathrm{d}\mu = \sum_{i=1}^{n} (\omega_i - \omega_{i-1}) \cdot \mu(\Omega_i),$$

where $\Omega_i = \{x | f(x) \geq \omega_i\}$ *and* $\omega_0 = 0$ *and* μ *is a fuzzy measure.*

More on the Choquet integral can be found in [2, 3, 4, 10, 15, 18], to just name a few sources.

In general, the universality of fuzzy integrals as aggregation operators is deducted from the minimal restrictions imposed on set function that is in its core. The fact that the Choquet integral, discussed in this paper, covers many well-known classical aggregation operators can be illustrated by the following example (see [11]).

Example 3.1

- *For the fuzzy measure* $\mu: P(X) \longrightarrow [0,1]$*, given by*
  $$\mu(X) = 1 \text{ and } \mu(A) = 0 \text{ for } A \neq X,$$
  the corresponding Choquet integral coincides with the classical minimum.
- *For the fuzzy measure* $\mu: P(X) \longrightarrow [0,1]$*, given by*
  $$\mu(\emptyset) = 0 \text{ and } \mu(A) = 1 \text{ for } A \neq \emptyset,$$
  the corresponding Choquet integral coincides with the classical maximum.
- *For the fuzzy measure* $\mu: P(X) \longrightarrow [0,1]$*, given by*
  $$\mu(A) = 0 \text{ for } \mathrm{card}(A) \leq n - k \text{ and } \mu(A) = 1 \text{ otherwise,}$$
  the corresponding Choquet integral coincides with the classical k*-order statistic.*
- *For the fuzzy measure* $\mu: P(X) \longrightarrow [0,1]$*, given by*
  $$\mu(A) = \frac{\mathrm{card}(A)}{\mathrm{card}(X)},$$
  the corresponding Choquet integral coincides with the classical arithmetic mean.
- *For the fuzzy measure* $\mu: P(X) \longrightarrow [0,1]$*, given by*
  $$\mu(A) = \sum_{j=0}^{\mathrm{card}(A)-1} w_{n-j},$$
  where w_i *are pre-given weights, the corresponding Choquet integral coincides with the OWA operator.*

The main drawback for the practical use of the Choquet integral is the number of sets that need a predefined value of the fuzzy measure. If the observed function has a range of cardinality *n*, a Decision Maker needs to predefine 2^n values. One of the possible ways for simplifying a Decision Maker's task is to define values only for singletons and to aggregate the remaining values by some aggregating operator. Since monotonicity of measure is essential for this integral, this can be done by a t-conorm, i.e., if the fuzzy measure μ is the so-called *S*-decomposable measure.

Definition 3.3 [18] *A set function* $\mu: P(X) \longrightarrow [0,1]$*that satisfies the following*

- $\mu(\emptyset) = 0$,
- $\mu(A \cup B) = S\big(\mu(A), \mu(B)\big)$ *for* $A \cap B = \emptyset$,

where S is a t-conorm, is called the S-decomposable measure.

A t-conorm is a binary operation $S: [0,1]^2 \longrightarrow [0,1]$, that is commutative, nondecreasing, associative and has zero as the neutral element. Elementary examples of continuous t-conorms are:

- $S_{\mathrm{M}}(x, y) = \max(x, y)$ − maximal,
- $S_{\mathrm{P}}(x, y) = x + y - xy$ − probabilistic,
- $S_{\mathrm{L}}(x, y) = \min(x + y, 1)$ – Lukasiewicz.

where $x, y \in [0,1]$. More on t-conorms and t-norms (dual operations) can be found in [16,18], among others. Also, the sources [11,14] offer more general background on aggregation operators.

Since t-conorms are associative operations, they can easily be extended to *n*-ary operators and used for calculating measures of non-singleton sets. Forms of *n*-ary operators for three previously mentioned basic t-conorms are given by the following example.

Example 3.2 *Let* $\{x_1, x_2, \dots, x_k\}$ *be an arbitrary subset of X. If* μ *is a S-decomposable measure, and values* $\mu(\{x_i\})$ *are predefined, then the value* $\mu(\{x_1, x_2, \dots, x_k\})$ *can be calculated as follows (see [16])*

- *if* $S = S_{\mathrm{M}}$
$$\mu(\{x_1, x_2, \dots, x_k\}) = \max(\mu(\{x_1\}), \dots, \mu(\{x_k\})),$$
- *if* $S = S_{\mathrm{P}}$
$$\mu(\{x_1, x_2, \dots, x_k\}) = 1 - \prod_{i=1}^{k}(1 - \mu(\{x_i\})),$$
- *if* $S = S_{\mathrm{L}}$
$$\mu(\{x_1, x_2, \dots, x_k\}) = \min\left(\sum_{i=1}^{k} \mu(\{x_i\}), 1\right).$$

Due to the nature of the problem that will be investigated further on, the focus of this paper is on the discrete case, i.e., when the observed set of input values is finite $X = \{x_1, x_2, \dots, x_n\}$.

3.2 MCLP – Classical Case

As mentioned in Section 1, MCLP was introduced by Church and ReVelle in 1974 [5], with the following mathematical model:

$$\text{maximize} \quad g = \sum_{i \in I} a_i y_i$$

$$\text{subject to} \quad \sum_{j \in N_i} x_j \geq y_i \,, \forall i \in I$$

$$\sum_{j \in J} x_j = P$$

$$x_j \in \{0,1\}, \forall j \in J$$

$$y_i \in \{0,1\}, \forall i \in I$$

where

I – set of locations (indexed by i)
J – set of eligible facility sites (indexed by j)
S – radius of coverage
d_{ij} – travel time from location i to location j
$x_j = \begin{cases} 1, \text{if facility is located at location } j \\ 0, \text{otherwise} \end{cases}$
a_i – population in node i
P – number of facilities
$N_i = \{j | d_{ij} \leq S\}$ – set of all facilities j which cover location i

In this paper, population in a node a_i is not considered, but it does not reduce the generality of the problem.

N_i is the set of facility sites and it provides location coverage, i.e., location is covered if the distance between it and some facility is less than the predefined radius S, and location is not covered otherwise. A demand node is "covered" when the closest facility to that node is at a distance less than or equal to S. A demand node is "uncovered" when the closest facility to that node is at a distance greater than S. The objective is to maximize the number of people served or "covered" within the desired service distance. Constraints of the type (1) allow $y_i$to equal 1 only when one or more facilities are established at sites in the set N_i(that is, one or more facilities are located within the S distance units of the demand point i). The number of facilities allocated is restricted to equal P in constraint (2). The solution to this problem specifies not only the largest amount of population that can be covered, but the P facilities that achieve this maximal coverage.

This condition is modeled by classical logic and each location could be fully covered or uncovered, and that fact gives motivation for the introduction of fuzzy numbers in modeling MCLP.

3.3 MCLP via Fuzzy Numbers (FMCLP)

The main idea of using fuzzy numbers in modeling MCLP is the introduction of vagueness in location covering. FMCLP is the extension of MCLP, where some conditions are represented with fuzzy numbers and in FMCLP, the location can be covered, uncovered or partially covered ([21]). Depending on the nature of the problem, different aggregation operators (max, arithmetic average, median, min...) can be used to calculate the degree of partial coverage of a location. In the following model of FMCLP, max operator is used, but other operators can be used in a similar way.

$$\text{Maximize} \quad g = \sum_{i \in I} y_i$$

$$\text{subject to} \quad \max x_j \cdot c_{ij} \geq y_i \,, \forall i \in I$$

$$\sum_{j \in J} x_j = P$$

$$x_j \in \{0,1\}, \forall j \in J$$

$$y_i \in [0,1], \forall i \in I$$

where

I – set of locations (indexed by i)
J – set of eligible facility sites (indexed by j)
S – radius of complete coverage

s – fuzzy radius of partial coverage
d_{ij} – travel time from location i to location j

$$x_j = \begin{cases} 1, \text{if facility is located at location } j \\ 0, \text{otherwise} \end{cases}$$

P – number of facilities

$$c_{ij} = \begin{cases} 1, \ d_{ij} \leq S \\ 0, \ d_{ij} \leq S + s \\ e \in (0,1), \ \text{otherwise} \end{cases}$$ - matrix of coverage

The main difference between MCLP and FMCLP lies in the coverage radius. In the presented FMCLP model, the coverage radius is a fuzzy number (right-shoulder fuzzy number) which allows partial coverage. Now, the coverage degree y_i is a number in the unit interval and the coverage matrix determines its value. The exact value of y_i is defined by a membership function and depends on the nature of the problem.

Travel time could also be a fuzzy number (these are usually triangular fuzzy numbers) and that modification results in another FMCLP model. In that model, partial coverage is defined by the intersection of the fuzzy radius (represented by a

right-shoulder fuzzy number) and fuzzy travel time (represented by a triangular fuzzy number). More on this approach and its applications in other location problems can be found in [21].

4 Fuzzy Integral-based Models of Fuzzy Maximal Covering Location Problem

The main motivation for proposing new models is taking into consideration the interaction measure between facilities. In all the existing models, the facilities could not interact with each other and each facility has the same importance. Thus, the level of interaction, or the level of joint importance, is given by a monotone set-function. Together with the usage of the Choquet integral, it forms a new, promising powerful extended model of FMCLP.

The basics of the proposed model are

- P – number of facilities [integer],
- $X = \{L_1, L_2, \dots, L_R\}$ – set of all locations,
- $Y = \{Y_1, Y_2, \dots, Y_P\}$ – set of all facilities,
- $\mu: \mathrm{P}(X) \longrightarrow [0,1]$ – measure of interaction for different facilities modeled by a monotone set function,
- $\omega_{i,j} \in [0,1]$ – degree of coverage for location L_i by the j-th facility,
- A is the intended layout of facilities from Y over the location set X.

The following constitute the proposed model:

MODEL Ch - the Choquet based model

$$f_{L_i}: Y \to \{\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,m}\}, \quad i = \{1, \dots, R\}, \tag{1}$$

$$g(A) = \sum_i (C) \int f_{L_i} \, \mathrm{d}\mu. \tag{2}$$

Namely, the functions (1) give the degree of coverage of each node by the facilities from Y, while formula (2) is the function whose maxima, for different positions of the facilities from Y, is needed. Given this, the layout of the facilities from Y for which (2) is maximal is the optimal layout. The monotone set function μ is predefined by a Decision Maker and can be interpreted as a quality measure of facilities and their interaction. The optimal case is obtained when the Decision Maker is able to provide the values of μ for all subsets of Y. By doing that, the Decision Maker expresses their own opinion on how the facilities in question interact, i.e., how "strong" they are together. However, this means that the Decision Maker should single-handedly provide 2^P values, which would be an unreasonable request. An acceptable solution is to ask for values only for

singletons, and to use an aggregation operator, e.g., a t-conorm, acceptable for the Decision Maker's behavior. The following algorithm is proposed

- STEP I: Acquiring values for $\mu(\{Y_1\}), \mu(\{Y_2\}), \dots, \mu(\{Y_P\})$,
- STEP II: Selection of the appropriate t-conorm:
  - S_{M} – if the strongest facility dominates all others,
  - S_{P} – if facilities complement each other, with overlaps,
  - S_L – if facilities complement each other, with negligible overlaps,
- STEP III: Calculation of values for
  $$\mu\left(\{Y_{j_1}, Y_{j_2}, \dots, Y_{j_k}\}\right), \ \{j_1, j_2, \dots, j_k\} \subseteq \{1,2, \dots, P\},$$
  by formulas from Example 3.2.

Remark 4.1 *Step II offers only three options because they can easily be interpreted by real life concepts such as domination (one facility is much more important to the Decision Maker and its influence is strong enough to overcome influences of other facilities) and negligible overlaps (influences of different facilities can be directed to the same area, however they do not compete with each other). Of course, the set of t-conorms is much wider (see [16]) and some other t-conorms can be chosen depending on the decision maker's preferences.*

The behavior of the proposed model depending on the Decision Maker's personal perceptions of quality and interaction of facilities is illustrated by the following propositions.

Proposition 4.1 Let $X = \{L_1, L_2, \dots, L_R\}$ be the set of all locations, $Y = \{Y_1, Y_2, \dots, Y_P\}$ the set of all facilities, $\omega_{i,j} \in [0,1]$ degree of coverage for location L_i by the *j*-th facility and let A be the intended layout of facilities from Y over the location set X.

If qualities of facilities $Y = \{Y_1, Y_2, \dots, Y_P\}$ are estimated by two different decision makers, i.e., if two S- decomposable measures $\mu_1: \mathrm{P}(Y) \longrightarrow [0,1]$ and $\mu_2: \mathrm{P}(Y) \longrightarrow [0,1]$ based on the same t-conorm S are assigned, such that

$$\mu_1(\{Y_j\}) \leq \mu_2(\{Y_j\}),$$

for all $j \in \{1,2, \dots, P\}$, then the following holds

$$g_{\mu_1}(A) \leq g_{\mu_2}(A).$$

Proof. Since $\mu_1: \mathrm{P}(Y) \longrightarrow [0,1]$ and $\mu_2: \mathrm{P}(Y) \longrightarrow [0,1]$ are S-decomposable measures and since for all singletons $\{Y_j\}$, $j \in \{1,2, \dots, P\}$, holds $\mu_1(\{Y_j\}) \leq \mu_2(\{Y_j\})$, based on monotonicity of t-conorms (see [16]), it follows that $\mu_1(E) \leq \mu_2(E)$ for all $E \in \mathrm{P}(Y)$. Now, based on properties of the Choquet integral (see [2,3]), it holds

$$(C)\int f_{L_i}\,\mathrm{d}\mu_1 \leq (C)\int f_{L_i}\,\mathrm{d}\mu_2,$$

for all corresponding functions $f_{L_i}: Y \to \{\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,m}\}$, $i \in \{1,2, \dots, R\}$. Therefore, the claim holds.

Proposition 4.2 Let $X = \{L_1, L_2, \dots, L_R\}$ be the set of all locations, $Y = \{Y_1, Y_2, \dots, Y_P\}$ the set of all facilities, $\omega_{i,j} \in [0,1]$ degree of coverage for location L_i by the j-th facility and let A be the intended layout of facilities from Y over the location set X.

If interactions of facilities $Y = \{Y_1, Y_2, \dots, Y_P\}$ are estimated by two different decision makers such that two different S-decomposable measures, S_1-decomposable measure $\mu_1: \mathrm{P}(Y) \longrightarrow [0,1]$ and S_2-decomposable measure $\mu_2: \mathrm{P}(Y) \longrightarrow [0,1]$, are assigned in the following manner

$$\mu_1(\{Y_j\}) = \mu_2(\{Y_j\}),$$

for all $j \in \{1,2, \dots, P\}$, and $S_1 \leq S_2$, then the following holds

$$g_{\mu_1}(A) \leq g_{\mu_2}(A).$$

Proof. Measures $\mu_1: \mathrm{P}(Y) \longrightarrow [0,1]$ and $\mu_2: \mathrm{P}(Y) \longrightarrow [0,1]$ are S-decomposable measures, therefore, based on the starting assumption $S_1 \leq S_2$ ($S_1(x,y) \leq S_2(x,y)$ for all $x, y \in [0,1]$, see [16]), it holds $\mu_1(E) \leq \mu_2(E)$ for all $E \in \mathrm{P}(Y)$. Now, due to properties of the Choquet integral (see [2,3]), analogous to the proof of the previous proposition, the claim holds.

Remark 4.2 *Since for three proposed t-conorms holds $S_M \leq S_P \leq S_L$, it is obvious that for the resulting mark for a certain layout A holds $g_{\mu_{S_M}}(A) \leq g_{\mu_{S_P}}(A) \leq g_{\mu_{S_L}}(A)$. That is, if facilities complement each other, instead having one that is dominant, the resulting mark is higher.*

Additionally, although at first glance the introduction of μ seems to increase the computational complexity, this can be avoided, because in the implementations only few subsets are connected to a single node.

Proposition 4.3 The algorithm for calculation of the function $g(A) = \sum_i (C) \int f_{L_i} d\mu$ has the maximal complexity of $O(RP \log P)$, where P is the number of the given facilities and R is the number of the observed locations.

Proof. The worst case, i.e., the maximal complexity, is reached when each location has a different deegree of coverage for all available facilities, that is when the range of function f_{L_i} has exactly P different elements, for all $i = \{1, \dots, R\}$. In that case, $\int f_{L_i} \mathrm{d}\mu = \sum_{k=1}^{P} (\omega_{i,k} - \omega_{i,k-1}) \cdot \mu(\Omega_{i,k})$ has P summands. Before calculation of this sum, it is necessary to sort elements from $\{\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,P}\}$, i.e., to sort the set of all deegrees of coverage. This can be done in $O(P \log P)$ steps (by using, for example, Merge Sort). With sorted elements, computational complexity of this sum depends on the complexity of computing measures $\mu(\Omega_{i,1}), \mu(\Omega_{i,2}), \dots, \mu(\Omega_{i,P})$. From the definition of the S-measure μ and properties of t-conorms in general, follows that

$$\mu(\Omega_{i,j}) = \mu(\{Y_j, Y_{j+1}, \dots, Y_P\}) = S\left(\mu(\{Y_j\}), \mu(\{Y_{j+1}, \dots, Y_P\})\right) = S\left(\mu(\{Y_j\}), \mu(\Omega_{i,j+1})\right),$$

which insures that integral $(C)\int f_{L_i} \mathrm{d}\mu$ (with sorted elements of $\{\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,P}\}$) can be computed with coplexity $O(P)$. Since there are R summands in the function $g(A)$, the total complexity is $O(RP \log P)$.

Remark 4.3 *In order to simplify the computational complexity and bring this concept closer to the Decision Maker, functions (1) can have linguistic values, i.e.,*

$$f_{L_i}: Y \to \{\text{none}, \text{poor}, \text{fair}, \text{good}, \text{full}\}. \quad (3)$$

If $f_{L_i}(Y_j)$=none*, then node (location)* L_i *is not in range of the facility* Y_j *for the observed layout, etc. Of course, later on, linguistic values can be appropriately coded. In this case, the exact values of elements in sets* $\{\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,P}\}$ *are known in advance and sorting can be done in* $O(P)$ *steps (by using, for example, Counting sort). Therefore, the total complexity is* $O(PR)$*.*

4.1 Examples

The proposed model can be illustrated by the following simple setting. Let the assumption be that there are 6 locations with distances as in Figure 1 and two facilities to be located. Now, the set of locations is $X = \{L_1, \dots, L_6\}$ and set of facilities is $Y = \{Y_a, Y_b\}$. However, since facilities will be positioned on certain locations, the notation will be $Y = \{Y_{aj}, Y_{bk}\}$ where $\{j, k\} \subset \{1,2,\dots,6\}$ depends on the intended position of a facility.

Let the assumption be that two layouts are under the consideration:

A: $Y = \{Y_{a1}, Y_{b6}\}$,
B: $Y = \{Y_{a2}, Y_{b5}\}$,

i.e., facilities $Y_a$and Y_b are located on locations L_1 and L_6, and L_2 and L_5, respectively. The first calculation is the implementation of the classical case, the second one is done via fuzzy numbers, while the third one is based on the model proposed in this paper. Since the quality (or influence) of facilities in question is the same for the first two approaches (given by examples 4.1 and 4.2), for the sake of simplicity, the following notations will be used:

A: $Y = \{Y_1, Y_6\}$,
B: $Y = \{Y_2, Y_5\}$,

which is the standard in MCLP problems. However, for the third approach, the quality of facility is relevant and this more complex notation will be used.

Example 4.1 *First, a classical MCLP problem without any fuzzy coverage will be used. Let it be supposed that the coverage radius is 5 km, i.e., the function is defined in the following way*

$$f_{L_i}(L_j) = \begin{cases} 1, if\ \mathrm{d}(L_i, L_j) \leq 5\ km, \\ 0, otherwise. \end{cases}$$

If the facilities are located in L_1 and L_6 then all six locations are covered. On the other hand, if the facilities are located in L_2 and L_5, only four locations are covered, the locations L_3 and L_6 are not covered by this solution. Therefore, the optimal solution is option A.

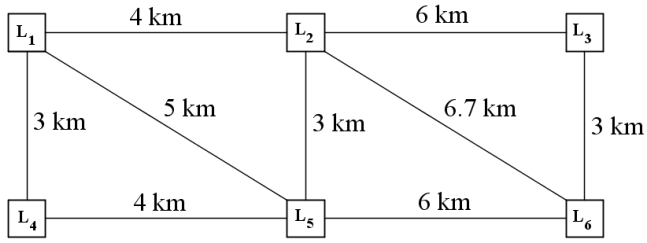


Figure 1
Location setting

Example 4.2 *It will now be supposed that the location can be partially covered, i.e., FMCLP will be considered. The coverage radius for this approach is defined by the following function (see [21])*

$$f_{L_i}(L_j) = \begin{cases} 1, & if\ \mathrm{d}(L_i, L_j) \leq 3km, \\ -\frac{1}{4}\mathrm{d}(L_i, L_j) + \frac{7}{4}, & 3\ km < \mathrm{d}(L_i, L_j) \leq 7km, \\ 0, & otherwise. \end{cases}$$

For option A, the facilities are located in L_1 and L_6 and they are marked as Y_1 and Y_6 and

- $f_{L_1}: \{Y_1, Y_6\} \to \{\omega_{1,1}, \omega_{1,2}\}$, $f_{L_1}(Y_1) = 1$, $f_{L_1}(Y_6) = 0$;
- $f_{L_2}: \{Y_1, Y_6\} \to \{\omega_{2,1}, \omega_{2,2}\}$, $f_{L_2}(Y_1) = 0.75$, $f_{L_2}(Y_6) = 0.075$;
- $f_{L_3}: \{Y_1, Y_6\} \to \{\omega_{3,1}, \omega_{3,2}\}$, $f_{L_3}(Y_1) = 0$, $f_{L_3}(Y_6) = 1$;
- $f_{L_4}: \{Y_1, Y_6\} \to \{\omega_{4,1}, \omega_{4,2}\}$, $f_{L_4}(Y_1) = 1$, $f_{L_4}(Y_6) = 0$;
- $f_{L_5}: \{Y_1, Y_6\} \to \{\omega_{5,1}, \omega_{5,2}\}$, $f_{L_5}(Y_1) = 0.5$, $f_{L_5}(Y_6) = 0.25$;
- $f_{L_6}: \{Y_1, Y_6\} \to \{\omega_{6,1}, \omega_{6,2}\}$, $f_{L_6}(Y_1) = 0$, $f_{L_6}(Y_6) = 1$.

Therefore, the coverage of location L_1 is $\max(1,0) = 1$, *for L_2 is* $\max(0.75,0.075) = 0.75$, *L_3 is* 1, *L_4 is* 1, *L_5 is* 0.5 *and L_6 is* 1. *Now, the coverage degree of the option A is*

$$g(A) = 1 + 0.75 + 1 + 1 + 0.5 + 1 = 5.25.$$

For layout B the following holds

- $f_{L_1}: \{Y_2, Y_5\} \to \{\omega_{1,1}, \omega_{1,2}\}, \quad f_{L_1}(Y_2) = 0.75, \quad f_{L_1}(Y_5) = 0.5;$
- $f_{L_2}: \{Y_2, Y_5\} \to \{\omega_{2,1}, \omega_{2,2}\}, \quad f_{L_2}(Y_2) = 1, \quad f_{L_2}(Y_5) = 1;$
- $f_{L_3}: \{Y_2, Y_5\} \to \{\omega_{3,1}, \omega_{3,2}\}, \quad f_{L_3}(Y_2) = 0.25, \quad f_{L_3}(Y_5) = 0.075;$
- $f_{L_4}: \{Y_2, Y_5\} \to \{\omega_{4,1}, \omega_{4,2}\}, \quad f_{L_4}(Y_2) = 0.5, \quad f_{L_4}(Y_5) = 0.75;$
- $f_{L_5}: \{Y_2, Y_5\} \to \{\omega_{5,1}, \omega_{5,2}\}, \quad f_{L_5}(Y_2) = 1, \quad f_{L_5}(Y_5) = 1;$
- $f_{L_6}: \{Y_2, Y_5\} \to \{\omega_{6,1}, \omega_{6,2}\}, \quad f_{L_6}(Y_2) = 0.075, f_{L_6}(Y_5) = 0.25.$

and

$$g(B) = 0.75 + 1 + 0.25 + 0.75 + 1 + 0.25 = 4.$$

Again, layout A is optimal.

The following example illustrates the proposed model based on the Choquet integral. In this case, the quality of facilities, at least the Decision Maker's perception of that quality, influences the result.

Example 4.3 *Let one consider option A. The values that describe the quality of each facility are $\mu(\{Y_{a1}\})$ and $\mu(\{Y_{b6}\})$, and they are provided by a Decision Maker. Their joint quality, i.e., the measure of how much they complement each other is $\mu(\{Y_{a1}, Y_{b6}\})$, can be obtained, as presented in Example 3.1. It is assumed that the measure of an empty set is zero.*

The next step is the calculation of the Choquet integral for each function according to the measure μ. That is "coverage of the location L_i":

- $(C)\int f_{L_1} \mathrm{d}\mu = (1-0)\cdot\mu\left(\{y | f_{L_1} \geq 1\}\right) = \mu(\{Y_{a1}\})$,
- $(C)\int f_{L_2} \mathrm{d}\mu = (0.075-0)\cdot\mu\left(\{y | f_{L_2} \geq 0.075\}\right) + (0.75-0.075)\cdot \mu\left(\{y | f_{L_2} \geq 0.08\}\right) = 0.075\mu(\{Y_{a1}, Y_{b6}\}) + 0.675\mu(\{Y_{a1}\})$,
- $(C)\int f_{L_3} \mathrm{d}\mu = (1-0)\cdot\mu\left(\{y | f_{L_3} \geq 1\}\right) = \mu(\{Y_{b6}\})$,
- $(C)\int f_{L_4} \mathrm{d}\mu = (1-0)\cdot\mu\left(\{y | f_{L_4} \geq 1\}\right) = \mu(\{Y_{a1}\})$,
- $(C)\int f_{L_5} \mathrm{d}\mu = (0.25-0)\cdot\mu\left(\{y | f_{L_5} \geq 0.25\}\right) + (0.5-0.25)\cdot \mu\left(\{y | f_{L_5} \geq 0.5\}\right) = 0.25\mu(\{Y_{a1}, Y_{b6}\}) + 0.25\mu(\{Y_{a1}\})$,
- $(C)\int f_{L_6} \mathrm{d}\mu = (1-0)\cdot\mu\left(\{y | f_{L_6} \geq 1\}\right) = \mu(\{Y_{b6}\})$.

The coverage degree of the $L_1 - L_6$ layout is given by

$$\begin{gathered} g(A) = \textstyle\sum_i (C)\int f_{L_i} \mathrm{d}\mu = \\ 2.925\mu(\{Y_{a1}\}) + 2\mu(\{Y_{b6}\}) + 0.325\mu(\{Y_{a1}, Y_{b6}\}). \end{gathered} \tag{4}$$

Similarly, for layout $L_2 - L_5$, i.e., for option B, the coverage degree is

$$\begin{gathered} g(B) = \sum_i (C)\int f_{L_i} \mathrm{d}\mu \\ = 0.425\left(\mu(\{Y_{a2}\}) + \mu(\{Y_{b5}\})\right) + 3.15\mu(\{Y_{a2}, Y_{b5}\}). \end{gathered} \tag{5}$$

Let it be assumed that the qualities of two facilities in question are graded with, e.g., 0.6 *and* 0.8 *and if overlaps are negligible, the* S_L *can be used as the aggregation operator. For option A, if the facility of quality* 0.6 *is placed on location* L_1*, the following holds*

$$\mu(\{Y_{a1}\}) = 0.6,\ \mu(\{Y_{b6}\}) = 0.8,\ \mu(\{Y_{a1}, Y_{b6}\}) = 1 \text{ and } g(A)=3.68.$$

On the other hand, for option B, if the facility of quality 0.6 *is placed on location* L_2*, the following holds*

$$\mu(\{Y_{a2}\}) = 0.6,\ \mu(\{Y_{b5}\}) = 0.8,\ \mu(\{Y_{a2}, Y_{b5}\}) = 1 \text{ and } g(B)=3.745.$$

Now, since the quality of facilities is taken into account, the result is different and the optimal solution is layout B.

While in MCLP and FMCLP the quality of facilities is not taken in to consideration, it has a high influence on the result in the proposed model. The flexibility of the proposed model can be additionally illustrated by the following example, that is the continuation of the previous one.

Example 4.4 *If the positions of facilities in option A are inverted, i.e., if the layout is* A: $Y = \{Y_{b1}, Y_{a6}\}$*, the following holds*

$$\mu(\{Y_{b1}\}) = 0.8,\ \mu(\{Y_{a6}\}) = 0.6,\ \mu(\{Y_{b1}, Y_{a6}\}) = 1 \text{ and } g(A)=3.865.$$

That is, now this layout is better than layout B.

Remark 4.4 *If the assumption is that all facilities are of the same quality, e.g., quality* 1*, the proposed model coincides with FMCLP.*

As seen from the previous examples, the new model allows the quality of facilities, given by the measure μ, to influence the final decision. All four examples are summarized in Table 1. The optimal option is marked with *.

Table 1

Comparison of coverage degrees

	MCLP	FMCLP	MODEL Ch, I	MODEL Ch, II
option A	6*	5.24*	3.68	3.865*
option B	4	4	3.745*	3.745

Remark 4.5 *If there is no other facility (e.g. hospital) near* Y_1 *(*$L_1 - L_6$ *layout), as illustrated in the previous example, the coverage degree of the location* L_1 *where is located* Y_1 *corresponds to* $\mu(\{Y_1\})$*, more precisely, it corresponds to the quality of* Y_1*. On the other hand, if the layout* $L_2 - L_5$ *is observed, hospitals are close, thus the coverage of the location* L_2 *corresponds to* $\mu(\{Y_2, Y_5\})$*, i.e., to the joint measure of facilities* Y_2 *and* Y_5*.*

Conclusion

This paper presents a generalization of the MCLP obtained by the incorporation of the Choquet integral into FMCLP. The nature of the observed integral takes into consideration the joint influence of each facility combination, which has not been done in any type of location problem before. The introduction of fuzzy integrals into the FMLCP makes the model more flexible and adaptable to real life problems. As it can be seen from (4), expert opinion of a Decision Maker given through set-function μ has a direct influence on the result. Thus a practical need for a new type of location problem is justified, and will further be called Extended FMCLP.

Acknowledgement

The authors would like to thank Humberto Bustince for providing the idea to use fuzzy integrals in FMCLP. This was suggested during a discussion at the FSTA 2014 conference.

This work was supported by the Ministry of Science and Technological Development of Republic of Serbia.

References

[1] V. Batanovic, D. Petrovic, R. Petrovic: Fuzzy Logic-based Algorithms for Maximum Covering Location Problems, Information Sciences 179 (1-2) (2009) 120-129, DOI: 10.1016/j.ins.2008.08.019

[2] P. Benvenuti, and R. Mesiar: ”Integrals with Respect to a General Fuzzy Measure”, Fuzzy Measures and Integrals (M. Grabisch, T. Murofushi, M. Sugeno eds.) Physica-Verlag (Springer-Verlag Company), Heidelberg 2000, 205-232

[3] P. Benvenuti, R. Mesiar, D. Vivina: ”Monotone Set Functions-based Integrals”, Handbook of Measure Theory (E. Pap ed.), Elsevier, Amsterdam, 2002, 1329-1379

[4] G. Choquet, ”Theory of Capacities”: Annales de l’Institut Fourier 5 (1953), 131-295

[5] R. Church and C. ReVelle: Maximal Covering Location Problem, Papers of the Regional Science Association 32 (1974) 101-118

[6] R. Church: The Planar Maximal Covering Location Problem, Journal of Regional Science 2(24) (1984) 185-201

[7] J. R. Current and J. E. Storbeck: Capacitated Covering Models, Environment and Planning B: Planning and Design 15(2) (1988) 153-163

[8] J. Darzentas: A Discrete Location Model with Fuzzy Accessibility Measures, Fuzzy Sets and Systems 23 (1987) 149-154, DOI: 10.1016/0165-0114(87)90106-0

[9] S. Davari, M. H. F. Zarandi, A. Hemmati: Maximal Covering Location Problem (MCLP) with Fuzzy Travel Times, Expert Systems with Applications 38 (12) (2011) 14535-14541, DOI: 10.1016/j.eswa.2011.05.031

[10] D. Deneberg: Non-Additive Measure and Integral, Kluwer Academic Publishers, Dordrecht-Boston-London, 1994

[11] M. Detyniecki: Fundamentals on Aggregation operators, http://www.cs.berkeley.edu/˜marcin/agop.pdf

[12] R. Z. Farahani, N. Asgari, N. Heidari, M. Hosseininia, M. Goh: Covering Problems in Facility Location: A Review, Computers and Industrial Engineering 62 (2012) 368-407, DOI:10.1016/j.cie.2011.08.020

[13] M. Grabisch: k-additive Fuzzy Measures, 6th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Granada, Spain, July 1996

[14] M. Grabisch, J. Marichal, R. Mesiar, and E. Pap: Aggregations Functions, Cambridge University Press, 2009

[15] M. Grabisch, H.T. Nguyen, and E.A. Walker: Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference, Kluwer Academics Publishers, Dordrecht, 1995

[16] E. P. Klement, R. Mesiar and E. Pap: Triangular Norms, Series: Trends in Logic, Kluwer Academic Publishers, Vol. 8, Dordrecht 2000

[17] A. T. Murray, D. Tong, and K. Kim: Enhancing Classic Coverage Location Models, International Regional Science Review 33(2) (2010) 115-133, DOI: 10.1177/0160017609340149

[18] E. Pap: Null-Additive Set Functions, Kluwer Academic Publishers, Dordrecht, 1995

[19] J. A. M. Perez, J. M. M. Vega, J. L. Verdegay: Fuzzy Location Problems on Networks, Fuzzy Sets and Systems 142 (2004) 393-405, DOI: 10.1016/S0165- 0114(03)00091-5

[20] C. ReVelle, and K. Hogan: The Maximum Availability Location Problem, Transportation Science 23 (1989) 192-200

[21] A. Takaci, M. Maric, D. Drakulic: The Role of Fuzzy Sets in Improving Maximal Covering Location Problem (MCLP). Procc. of SISY 2012, Subotica, Serbia 103-106

[22] Z. Wang, and G. J. Klir: Generalized Measure Theory, Springer, 2000

[23] R. R. Yager: ”On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making,” IEEE Transactions on Systems, Man and Cybernetics 18(1988), 183-190

Improving Automation and Process Control of an Indirect Gravure (pad) Printing Machine

Arash Hakimi Tehrani, Edgar Dörsam

Technische Universität Darmstadt, Institute of Printing Science and Technology, Magdalenenstr. 2, 64289 Darmstadt, Germany,
e-mail: hakimi_a@idd.tu-darmstadt.de, doersam@idd.tu-darmstadt.de

Jann Neumann

PERFECTA Cutting Systems GmbH, Schäfferstraße 44, D-02625 Bautzen, Germany, e-mail: neumann@perfecta.de

Abstract: Because pad printing can be used on 3-D substrates, it has attracted the attention of many researchers in the field of printed electronics. This paper presents developments in the automation of a pad printing machine that improve its specifications for use in scientific fields and equip it with some unique features. Users of this machine can obtain graphs of printing force and printing step durations for tracing and analyzing the printing process. Here, to explain the design method, the printing technique features, the pad printing working process and related machine parts, as well as the development and design process, are described. In this section, some hardware, such as National Instruments CompactRIO, as well as software (LabVIEW) and data transferring under the EtherCAT protocol will also be discussed. Finally, the machine user interface and some analytical graphs of the machine will be explained.

Keywords: Automation pyramid; Mechatronic system structure; Indirect gravure printing; Pad printing control system; LabVIEW

1 Introduction

Indirect gravure printing is the collective name for an indirect printing process having one transferring part (pad) and one gravure printing form. In many cases, it is referred to as pad printing [5, 10]. Pad printing has some advantages over other printing methods. Because it is a gilt-edged technique for printing on non-smooth objects having concave and convex surfaces, it has a competitive advantage for work with 3-dimensional substrates, which have differing shapes, thicknesses and

dimensions. This capability makes pad printing suitable for a wide range of uses in a variety of production processes, such as medical instruments, electrical devices, automotive parts and printed electronic devices. This technique has been applied in the medical imaging field to print piezoelectric thick-films on a curved substrate [15]. It has also been used in the production processes of gas sensors [4], solar cells [6, 13], UHF RFIDs [16], OLEDs, biomedical sensors [24], mobile phone antennas [27], and microelectronic circuits [11]. Moreover, it can also be combined with other methods, such as screen printing. For example, in [14], some research success was achieved by combining pad and screen printing to produce an ultrasonic transducer in high frequency scales. Because of the importance of pad printing usage in such scientific fields, this paper focuses on the development of control and automation for pad printing machines, with a focus on their usage in scientific fields. According to scientific research, a more highly automated pad printing machine represents a new demand. Here, automation should provide high accuracy, high value of data transfer and management and high controllability of printing parameters. So, the main contribution of this work is an improvement of the automation level of pad printing machines.

This paper is organized as follows: the next section, introduces basic concepts of mechatronics, automation and pad printing machine. Second, the development concept for the pad printing automation system is presented. Third, the various components of the pad printing machine and its structure are described. Afterward, a flowchart diagram of the machine's working process is offered. Next, we focus on the development of the machine structure, automation level, data flow and software designing process as parts of the development process for the pad printing machine. Then, as examples, some reports of the system are mentioned. Finally, the paper is concluded with a brief summary.

1.1 Mechatronic System Description

To achieve this goal, the pad printing machine is developed in the three fields of electronics, control systems, and mechanics. As shown in Figure 1 (a), these fields are all related to a mechatronic system [8]. Therefore, to provide a better description, the pad printing machine is considered as a mechatronic system.

Figure 1 (b) shows the structure of a mechatronic system, which consists of four units: control, sensors, actuators and mechanics [9].

The level of controlling and processing and the user interface is related to the level of system automation, which is described in the next section.

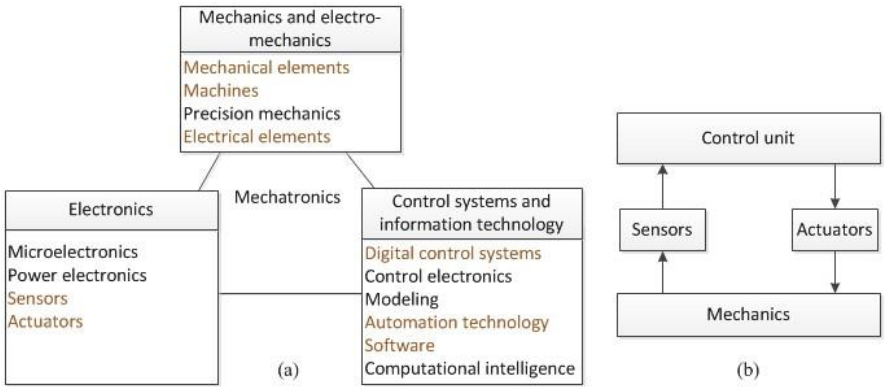


Figure 1

(a): Mechatronics field concept. The connected lines represent the synergistic integration of these three scientific fields in generating the mechatronics concept. The orange color defines the developed parts that are described in this paper. (b): The mechatronic system structure [9], with its four units of control, sensors, actuators and mechanics. The arrows show the direction of information flow.

1.2 Automation Description

Automation is the application of a control system in a process toward the end of reducing human intervention, improving process throughput, and decreasing production losses [1]. The automation pyramid in Figure 2 shows different levels of automation.

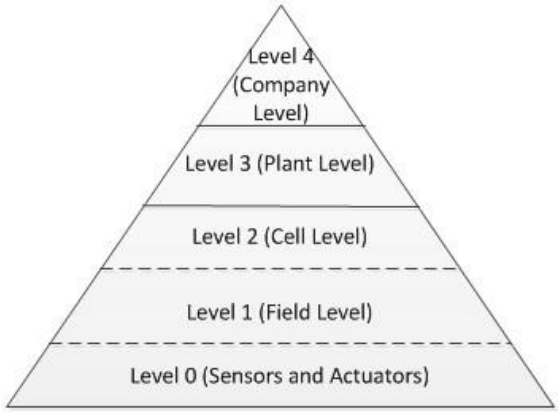


Figure 2

The automation pyramid. The automation development direction is from bottom (low level) to top (high level). In some cases, levels 0, 1 and 2 are considered as one group; hence, the dashed lines [9, 23, 26]

The possibility of passing through each level to a higher level varies from system to system. In some cases, achieving a particular level is very difficult, or even impossible. Level 0 is achieved merely by use of sensors and actuators to control the (mechatronic) system. Level 1 involves controlling and processing the signals of the system and is called the field level. Level 2 (the cell level) depends on user interface and process monitoring. Level 3 (the plant level) involves optimal scheduling and maintenance, as provided by the manufacturing execution system (MES) and the management information system (MIS). Level 4 (the company level) involves enterprise resource planning (ERP) and the programming and production control of an entire company [9, 23, 26].

1.3 Description of Pad Transfer Printing

Pad printing machines come in two types: pad transfer printing and rotary pad transfer printing. This paper focuses on pad transfer printing. A schematic diagram of pad printing is given in Figure 3. As shown, the pad and printing form are initially located in their reference positions (red dash). The printing form table then advances (step 2). Next, the pad comes down to pick up the ink (step 3), a movement reversed in step 4. In step 5, the printing form table returns to its reference position. Finally, the pad descends to transfer the ink film to the substrate [5] (step 6), and then returns to its reference position in step 7.

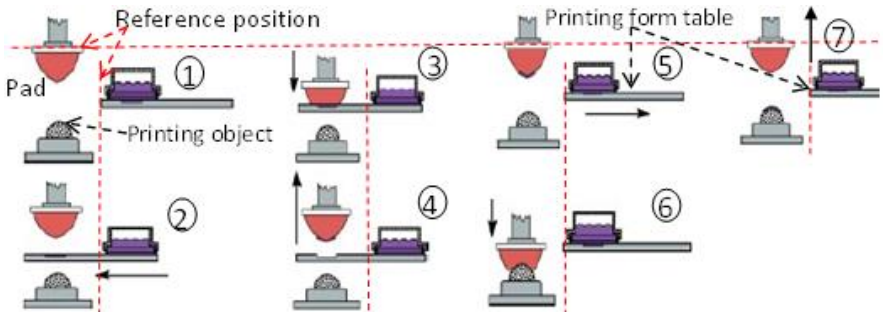


Figure 3
Pad transfer printing schematic diagram showing the pad, printing form table and printing object. The 1st picture shows the start of the process and the 7th picture shows the end, where pad, table and object return to their original position. The arrows define the movement direction [2].

2 Concept of an Automated Pad Printing Machine

Over time, there have been some improvements in pad printing machines. Today, most well-known models have features such as variable speed control, printed-pieces counter and variable pad position [17, 18, 25]. The pad printing machine

developed in this project has 4 extra capabilities: printing process force tracing on the substrate and printing forms; position control in both the X and Y directions; velocity control in both the X and Y directions and user-defined contact time. Within each printing cycle, these parameters can be changed and/or saved for future batches and are traceable in different graph formats for each printed sample. These parameters have been classified in Table 1. As indicated, the force as a printing parameter can be controlled and traced when the pad is pressed on either the printing form or the substrate. The position and velocity of the pad and printing form axes are controlled and traced in two directions: X and Y. The contact time of the pad on the printing form (for obtaining ink) and on the substrate (for transferring ink) can also be controlled and traced.

Table 1
Categorization of printing parameters according to their controllability and traceability (C&T) at different axis directions in newly developed pad printing machine

Printing parameters	C&T at Y (pad) axis direction	C&T at X (printing form) axis direction
Force	Pad on printing form	NO
	Pad on substrate	
Position	Completely	Completely
Velocity	Completely	Completely
Contact time	Pad on printing form	Pad on printing form
	Pad on substrate	

3 Categorization of Pad Printing Machine Components according to a Mechatronic System Structure

The most important movable parts of a pad printing machine are its axes (See Figure 4). These are referred to as the printing form axis and the pad axis. The printing form axis movement is in the forward (X) direction. The movement along the pad axis (Y) is downward, so it has negative values compared to the reference coordinate system.

As shown in Figure 1 (b), a mechatronic system has four parts: control unit, sensors, actuators and mechanics [9]. Thus, the pad printing machine structure can be illustrated as in Figure 5.

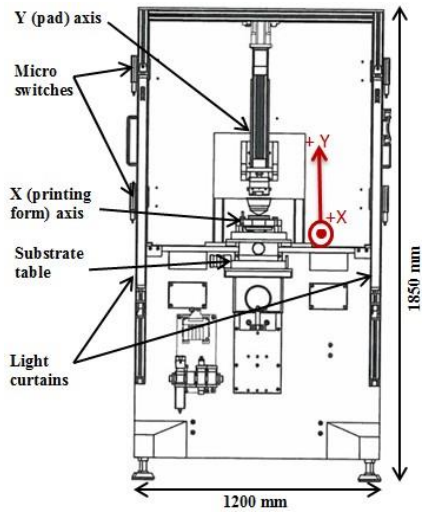


Figure 4

Pad printing machine schema. The dimensions of the machine are 1200 by 980 by1850 mm. The Y positive direction is upward and X positive direction is forward [18]. These elements are highlighted to promote a better understanding of the printing machine structure given in Figure 5.

According to the mechatronic structure of Figure 1 (b), the four general units of the new pad printing machine structure and their relationship to each other are described in Figure 5. All processing and logical calculations happen in the control unit. This unit consists of the machine's real-time embedded industrial controller and electric servo drives. As shown, the vertical and horizontal drives receive control data from the real-time controller and send the results back to it. The control unit has another element, identified as software. This element receives the demands of the user in the UI (user interface) and sends them to the Main software block. After processing, these demands are then sent to the real time controller. Further, the controlling program is located at the Main block. The actuators receive the commands of the controlling unit and execute them on the mechanical parts. For example, the vertical drive controls the vertical servomotor (actuator). This actuator moves the vertical axes (mechanical unit), which ultimately moves the pad as a printing unit. The sensors unit is another part of system that measures some parameters of the mechanics unit and then sends them to the control unit for processing. The mechanics unit has the role of mechanically executing user commands. The most important elements of the mechanics unit are shown in Figure 4 and Figure 6 (a). The printing unit is located in the mechanics unit and consists of the pad unit, the inking unit and the object-holding unit [10]. They can be called the operational, input and output units respectively, according to the operating maintenance model for printing [7]. Each of these has special parts, which are shown in Figure 6 (a).

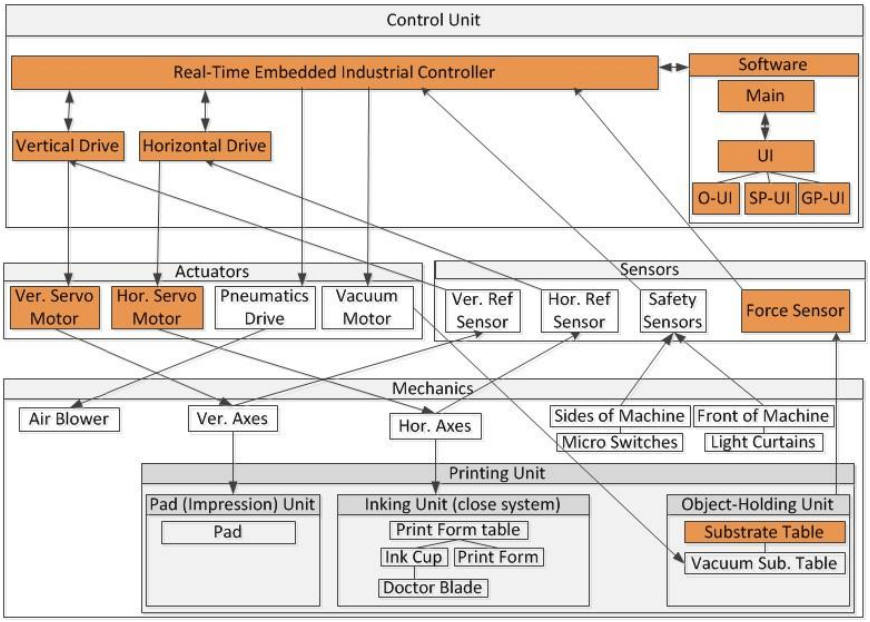


Figure 5

The pad printing machine structure. It is divided into control, actuators, sensors and mechanics units, according to the mechatronic system structure. The arrow directions indicate the machine data flow. The orange color defines the developed parts. The printing unit is considered a part of the mechanics unit and consists of pad, inking and object-holding units. It is classified to help illustrate the development points of the machine.

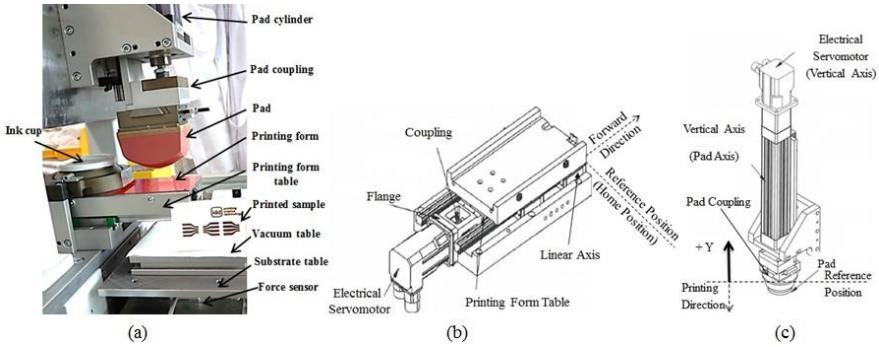


Figure 6

(a): The pad printing machine unit. The component parts are labeled in the (a) segment. (b): The printing form (X) axis elements [18]. The reference position line and forward direction of movement have been defined here. The printing form table is located at the reference position when the table tip is positioned at the reference position. The printing form receives the ink when moved in the forward direction. (c): The pad (Y) axis elements [18]. The reference position line and printing direction have been defined. The printing direction is in the -Y direction. The pad and pad axis are located at the reference position when the pad coupling is at the reference position line.

The main material of the pad is silicon. The pad shape has the two important parameters of pad angle (from side to print area) and pad printing surface [22]. The pad's role is to transfer ink from the printing form to the substrate. The printing form is an etched plate of print motif. Pad printing is divided into two main types: closed and open inking systems. In open inking systems, there is no cover on the ink trough, whereas in closed systems, the ink trough is sealed. Thus, there is more solvent evaporation in open systems than in closed [5, 10]. The inking system discussed in this paper is a closed system. The ink cup (the closed system ink trough) is an ink storage device located on the printing form that delivers the ink to the printing form in each printing cycle [6]. The substrate is an object located on the substrate table on which the printing process is performed. More detailed pictures of the printing form and pad axes from Figure 6 (a) are given in Figure 6 (b) and Figure 6 (c), respectively. The reference (home) position and printing direction of each axis are shown in these figures. The reference position of the printing form table and its forward direction are shown in Figure 6 (b). The printing form table has an electrical servomotor that has been connected to the linear axis by means of a coupling and flange. These accessories allow forward and backward movement of the printing form table. The reference position and Y direction of the pad axis are shown in Figure 6 (c) (vertical direction). The Y axis positive direction is upward. In other words, the printing process is performed in the negative direction. The reference position is the connection point of the pad with the pad coupling. The tip of the pad is not selected as the reference point, since pads with different heights may be used at this location.

4 Pad Printing Working Process

The pad printing working process is described in Figure 7. The inputs of the flowchart are printing form, ink, pad and substrate (2D or 3D).

Initially, the printing form and pad goes to their reference positions. In the next step, the substrate is fixed on its table. The substrate fixer could be a vacuum table. Briefly, according to this flowchart, the printing form gets the ink from the ink cup and then moves forward. Next, the pad descends and receives the ink from the printing form. The pad then goes up again and the printing form table comes backward. The pad then comes down and is pressed onto the substrate, thus transferring the ink. Finally, the pad ascends, and the printing project is completed.

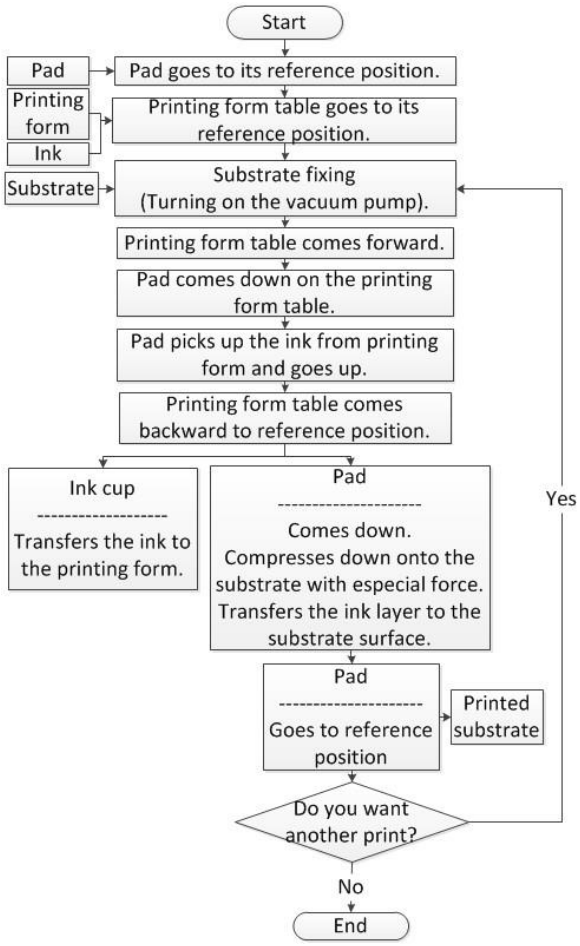


Figure 7

Pad printing working process flowchart. The different steps of the printing process are described here from start to end. The pad, printing form, ink and substrate are input materials and the printed substrate is the output.

5 Development Process of Pad Printing Machine

The hardware type, software design and automation level of the system are related to the machine's application.

The hardware for use in scientific fields should be capable of highly accurate control. Then, pursuant to its application, one designs the desired data flow

between hardware, software and final user. Next, one selects appropriate software according to the hardware and data flow system. Eventually, the programing process is started. In the following, the development process of the machine structure is described. Then, the development of the system's automation level and its data transferring route are discussed. Afterward, the uniquely designed software is presented, with concentration on the user interface (UI).

5.1 Development of the Machine Structure

The parts of the machine units we developed are shown in Figure 5. As the most important control unit element, the National Instrument CompactRio (cRIO) 9074 has been used. The cRIO is a real-time, embedded industrial machine controller with additional monitoring capabilities. Its specifications have been described in [20]. With this device, we were able take advantage of features such as a high speed, real-time processor, the ability to add measurement devices as I/O modules and the ability to expand external devices through networking. So, by utilizing these advantages and software features, an on-line controlling and data mining system with data measurement and processing capabilities could be created. Two Kollmorgen AKD servo drives were used as the vertical and horizontal drives of the controlling unit in Figure 5. These drives are capable of multi-axis programmable motion. Moreover, they can measure and control the speed, acceleration, position, torque and current of servomotors. Their response to mechanical load changes is immediate, thus allowing for an appropriate control level. Motor control is possible in three operation modes: torque, velocity and position. In torque mode, the motor current is controlledand the current loop is updated every 0.67 microseconds to achieve an accurate control system. The drives used in the scientific pad printing machine support the EtherCAT protocol for data transferring [12, 19]. An EtherCAT connection was used as a network protocol in this project [21]. By using EtherCAT, the contact time between the process steps and the CPU load is decreased [3]. It has high-speed performance with an accurate synchronization of less than 1 microsecond between master and slave, a feature important for coordinated motion between the motion axes. Because of its features, EtherCAT is used in machine design, motion control and measurement equipment applications [3, 21]. To ensure precise, delicate motion control, a quick and synchronized data transferring system is needed. Therefore, in our scientific pad printing machine, the EtherCAT has Kollmorgen drives connected to the cRIO.

An overview of industrial communication systems is shown in Table 2. The maximum bit rate of the EtherCAT data transferring system is 100 Mbit/s, which compares favorably with other methods [9]. The communication relationship of the EtherCAT protocol is master/slave. This means that a device has one-sided control over one or many devices. In the system described in this paper, the CompactRIO hardware, as a real-time, embedded industrial controller, has the role of master for controlling the horizontal and vertical AKD servo drives (slaves).

Table 2
Different industrial communication systems [9]. The most important parameters of industrial communication systems are the Max. bit rate, Max. number of nodes and the communication relationship. According to the communication specifications of the system hardware, the EtherCAT system is one of the best candidates for the new pad printing machine, since it has a high bit rate (100Mbit/s) and master/slave communication (because in this case two drives are to be controlled by means of a real-time controller).

System	Max. bit rate	Max. Nodes No.	Communication
AS-i	167 kbit/s	124	Master/Slave
CAN	1 Mbit/s	127	Publisher/Subscriber
PROFIBUS DP	12 Mbit/s	125	Master/Slave
DeviceNet	0.5 Mbit/s	64	Master/Slave
INTERBUS	2 Mbit/s	512	Master/Slave
SERCOS	16 Mbit/s	255	Master/Slave
PROFINET	100 Mbit/s	Unlimited	Master/Slave
EtherCAT	100 Mbit/s	65535	Master/Slave
Powerlink	100 Mbit/s	254	Publisher/Subscriber

As shown in Figure 5, a force sensor has been appended to the sensors unit. A single-point load cell with a maximum capacity of 100 kg and a safe load limit of 150 kg at a maximum eccentricity of 150 mm and accuracy class C3 has been used as a force sensor. In addition, a force measurement capability has been added to the printing form table. Ultimately, all of these forces are measured and controlled at a high accuracy level (Min. LC verification interval of 0.1961 N for max. capacity of 980.665 N) as part of the effective parameters of printing quality. In addition to controlling the printing process, the user can store these data for off-line data analysis. These capabilities, along with high-speed data transfer over the EtherCAT protocol, validate this machine as a scientific pad printing machine.

5.2 Development of the Automation Level

In accordance with the automation pyramid (Figure 2) and the specifications of conventional pad printing machines, most well-known pad printing machines [17, 18, 25] have normal sensors and actuators and controller devices, such as PLCs. Therefore, based on their features, they are located at level 0 or 1. Although many of them have an input system for entering printing parameters, they do not have process monitoring on their user interface, and thus do not advance to level 2. The newly designed and automated pad printing machine has reached the second automation level due to its on-line monitoring of the printing process and the parameters on the user interface. Moreover, the ability to alter printing parameters for the next printing sample according to the monitored data and the ability to store, handle and trace data with DIAdem (Version 14, National Instruments) is an

improvement in the pad printing information management system that could lift the automation level of the new pad printing machine to level 3.

5.3 Data Flow of the Pad Printing Machine Structure

The data flow between different units of the pad printing machine is illustrated in Figure 5. In this picture, the arrows point in the direction of the pad printing machine data flow. The user interface (UI) receives demands from the user and sends them to the software Main program of the control unit. After that, at the same unit, these data are transferred to the real-time controller for processing and translating into machine language. Then, the electrical command is sent to the horizontal and vertical drives. Then the data are sent to the actuator unit and, finally, executed by means of printing units in the mechanics unit.

The outputs of this newly designed machine are divided into the two categories of printed objects (printed outputs) and printing specification reports (software outputs). The printed substrate could be a 3-D object (e.g., printed electronic devices on 3-D surfaces). The other output is software based and produces such useful machine reports as printing force or inking force. It should be noted that the data flow route of the machine parts takes place via the EtherCAT protocol with a high-speed data transfer rate (max. bit rate of 100 Mbit/s) [9, 21].

5.4 Software Design

In this work, the LabVIEW (Version 13, National Instruments) was used to program the embedded FPGAs. The programing procedure of this new machine has been classified into two parts. One part is the Main program and the other part has been designed as a machine user interface (UI). The Main program has various block diagrams. This part has been designed for the control and data processing of different machine units (Control, Actuators, Sensors, Mechanics) according to user demands and the working process flowchart of Figure 7. The LabVIEW programing structures of the Main program were defined pursuant to the flowchart steps of Figure 7 and related working functions were then programmed inside each program structure. This program also processes all measured machine data. In another step, all measured data is sent to another program part, where data management is performed. In every 4 milliseconds, a package of data is sent to the data inventory for research and scientific analysis.

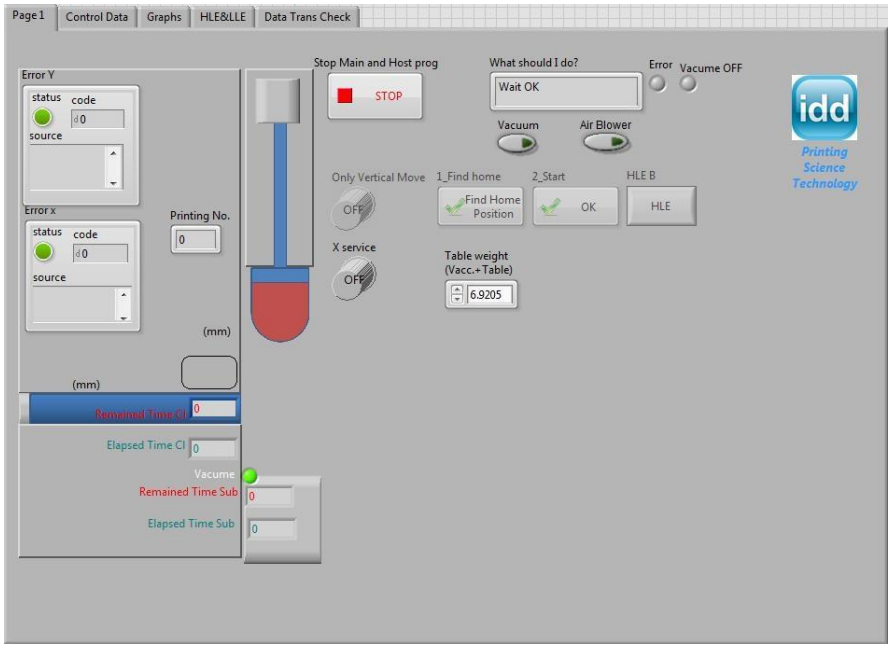


Figure 8

The Operational sub-User Interface (O-UI). The left side of picture shows the on-line printing process graphically. Other parts are the printing operational keys. They are called operational, because of their effect on the operation of the printing process. By pressing the "Find home position", the axes will go to their reference position, and by pressing the "start" button, the printing process will be started.

In front of all these complex processes, there is a supporting program with a graphical part as a mask. We refer to this part of the machine as the user interface (UI). It is divided into different sub-UIs, as shown in Figure 8, Figure 9 and Figure 10. The UI receives the demands of the user and translates them into useful parameters for the control program part. Figure 8 shows the operational sub-UI (O-UI), so named because of the execution functional keys located there, such as "start printing process". In a graphical segment of this sub-UI, the machine working process (pad and printing form movement) is shown on-line as a graphical animation.

In Figure 9, the machine set point sub-UI (SP-UI) is shown. In the SP-UI, the values of the printing parameters are defined. For an easy definition of values, all parameters have been categorized according to the printing steps shown in Figure 3 and the speed limitations have been defined here for input parameters. The ability to save and load parameter data has been added in the SP-UI to make it easy to use the machine and print with the same parameters at different times. The user can define the pad printing working process via automatic force control on the printing form or substrate merely by pushing a button on the SP-UI. The

contact time on the printing form and the substrate are two other parameters that the user may choose to influence the printing process. All of these options represent advantages of our pad printing machine over conventional machines.

Figure 9

The Set Point sub-User Interface (SP-UI). The different printing parameters are described here by the user. The printing parameters have been categorized according to the printing step involved; the number and description of each step is described at the left side of picture. These classifications make it easy for the user to operate the machine.

Figure 10 shows the on-line Graph Panel sub-User Interface (GP-UI). This panel displays the on-line graphs of the printing force, the printing form force, the position and velocity of the X-axis, and the position and velocity of the Y-axis over time during the printing process. The user can also save the data of the desired graphs in the host computer for further scientific analysis.

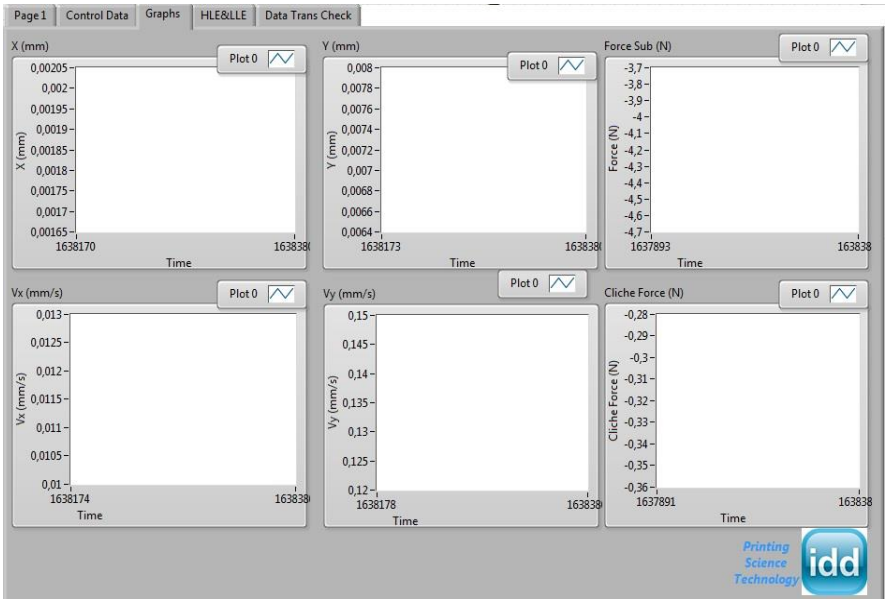


Figure 10

The Graph Panel sub-User Interface (GP-UI). This panel shows the position and velocity of pad and printing form axes, the printing force and ink delivery force (of printing form). All of these 6 diagrams are shown on-line for the duration of the printing process. By means of these graphs, the user can monitor and trace the printing process and make decisions that improve the parameters for subsequent printing iterations.

The main contribution of this work is an improvement of the automation level of pad printing machines. As described in section 5.2, the most well-known pad printing machines [17, 18, 25] are located at the automation level 1 or 2, whereas the presented developed machine has reached the automation level 3. According to this novelty, the printing parameters (Figure 9) can be set with sufficient precision. Afterward, the printing process will be executed by the developed control system and other parts (Figure 5) according to the set values, identically. These parameters are adjusted independently of each other. This feature will cause to better controllability of the printing process. As an example, the printing force, speed and pad position can independently be adjusted and controlled. Then, the effect of these parameters on the printing quality and process can be controlled. Whereas, the conventional pad printing machines can not support such aspects.

The pad printing is a complex process and it is important to investigate its process, systematically. Future work should aim to decrease the complexity of the system and to investigate the unknown behaviors of the system during the printing process. The highly automated pad printing machine developed in this work forms an ideal basis for further researches, since it allows for measurement of the printing parameters, and the ability of the database analysis.

6 First Results

Data related to the force, velocity, position and torque of the printing and printing form axes are saved in every 4 milliseconds as software outputs of the machine. These data have a relationship to some of the parameters affecting the printing quality. Thus, depending on user demands, different types of graphs and analyses may be obtained using these data. For example, using an 85 by 75 by 66 mm pad of hardness 12 Shore A together with a polyethylene terephthalate (PET) substrate of thickness 125 micrometer, a length of 297 mm and a width of 210 mm, the following results were obtained: Figure 11 shows the movement of the printing form (X) and pad (Y) axes according to the printing steps of Figure 3. The X curve (green) describes the position and movement of the printing form axis and the Y curve (red) delineates the pad axis movement versus time. For example, steps 3 and 4 are related to ink pick up. In step 3, the pad goes down to retrieve ink; thereafter, in step 4, the pad goes up again. According to this diagram, you can easily measure the time duration of each printing step.

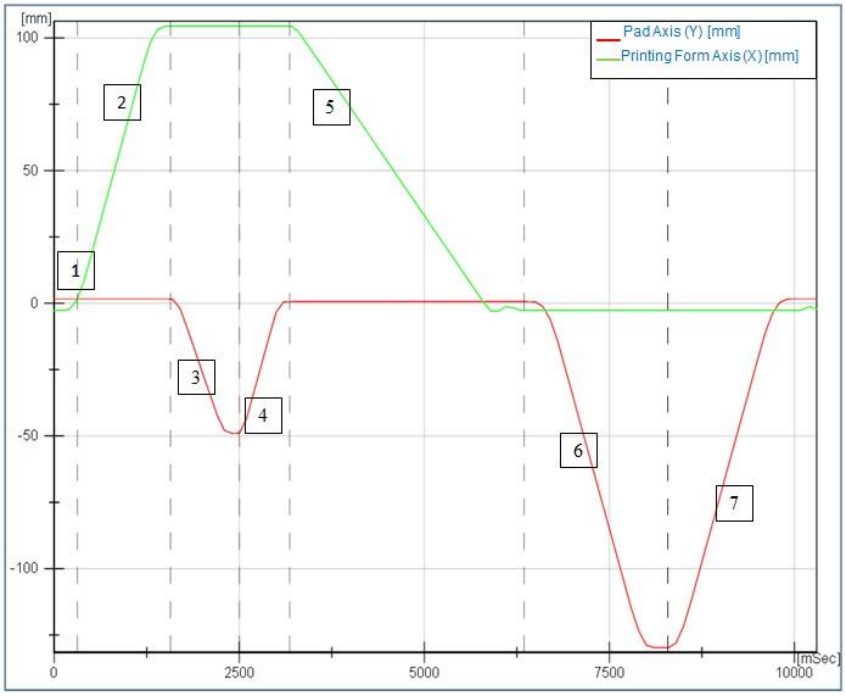


Figure 11

Axes movement diagrams. This graph shows the pad axis movement (red curve) and printing form axis (green curve) over time. The vertical dashed lines show the time duration (millisec) of each printing step, as numbered according to Figure 3. The highly accurate system measurements permit time units of milliseconds (mSec). The values of pad axis position are negative, since the printing direction is along the -Y axis.

One type of scientific graph is shown in Figure 12. This graph shows the pad movement behavior versus force for measured data in the pad printing process. The pad movements can be obtained approximately by calculating the force-Y (pad movement) using equation (1) (See dashed line in Figure 12). The F-Y diagram is divided into two parts: Increasing pad force when the pad moves downward on the printing form (step 6 from Figure 3) and decreasing pad force when the pad moves upward (step 7 from Figure 3).

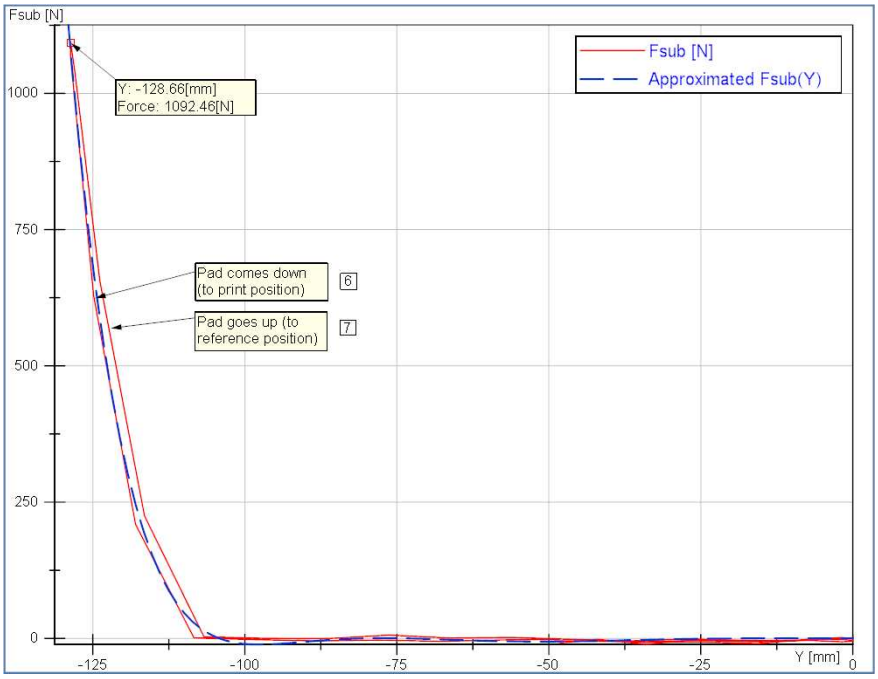


Figure 12

Printing force versus pad movement and its approximate equation diagram. The printing force curve (red) has two parts. The left part is related to pad downward motion for ink transferal and the other is related to pad upward motion after printing on the substrate. They have been signed according to printing steps 6 and 7 (See Figure 3). Because of the printing direction (against Y direction) the pad position values are negative. According to increasing of absolute position value from |-112| mm till |-128.66| mm, the pad is in contact with the substrate and the force is increased. The force-increasing behavior according to pad position in this experiment is an eighth-order equation (equation (1)) whose approximated curve (blue curve) has been derived by means of DIAdem.

Using equation (1), the pad axis movement for printing with a special force (for this type of pad) can be calculated.

$$F(Y) = 2.927E07 * Y^5 + 1.04E - 08 * Y^6 + 1.221E - 10 * Y^7 + 4.723E - 13 * Y^8 \quad (1)$$

Thus, our scientific pad printing machine was able to generate scientific graphs and allow conclusions about printing processes and situations, especially for highly intricate objects, such as printed electronic devices.

7 Summary

This paper has described the process of development at the automation level and the resulting structure of a pad printing machine, with a focus on scientific applications. The mechatronic system structure has been taken as a machine structure model and the printing unit has been incorporated as part of this system. Some developments have been made regarding the machine structure, such as the use of force sensors, National Instrument CompactRio hardware and Kollmorgen servo drives over EtherCAT data transferring protocol. The goals for these devices are highly accurate data measurements, processing and controlling functions, and high speed data transfer. As a validation of our work, we were able to generate various scientific graphs that reflect the printing process for each printed object (2-D or 3-D). Via these graphs, the user can improve the printing quality for future printing iterations and achieve traceable and repeatable printing. The user can follow the printing graphs on-line via the machine user interface or off-line via DIAdem or related software. All of these developments have led to an increase in the automation level of the pad printing machine. For example, the graph tracing the X-Y position movement versus time has been used to determine the duration of different printing steps. The graph of the pad movement versus force for a special pad has been generated and its approximated equation has been derived. This diagram and its related equation are useful for calculating pad axis movement for obtaining a printed object with a special force. The scientific pad printing machine can be used for unique applications, such as printed electronic devices (e. g., OLEDs).

References

[1] I. Azolibe, E. W. McGookin, J. Houston, and C. Winton, "Serving the Data Needs of Multiple Applications with One Data Source: an Industry Application Case Study," in *Symposium on Information Control Problems in Manufacturing (INCOM 2015)/IEEE*, Ottowa, Canada, 2015, pp. 1-6

[2] T. G. DECO*,* (2015)*, Introduction to Pad Printing - Pad Printing 101,* Available: http://www.decotechgroup.com/library/pad-printing/tech-bulletin-pad-print-101/

[3] ET Group, "Ethercat - the Ethernet Fieldbus," ed. Nuremberg, Germany: EtherCAT Technology Group, 2014, pp. 4-9

[4] V. Golovanov, J. L. Solis, V. Lantto, and S. Leppävuori, "Different Thick-film Methods in Printing of One-Electrode Semiconductor Gas Sensors," *Sensors and Actuators B: Chemical,* Vol. 34, pp. 401-406, Aug. 1996

[5] P. Hahne, *Innovative Drucktechnologien: Siebdruck - Tampondruck*. Verlag Der Siebdruck, Germany, 2001

[6] P. Hahne, E. Hirth, I. E. Reis, K. Schwichtenberg, W. Richtering, F. M. Horn*, et al.*, "Progress in Thick-Film Pad Printing Technique for Solar Cells," *Solar Energy Materials and Solar Cells,* Vol. 65, pp. 399-407, Jan. 2001

[7] C. Horváth and Z. Gaál, "Operating Maintenance Model for Modern Printing Machines," *Acta Polytechnica Hungarica,* Vol. 5, pp. 39-47, 2008

[8] R. Isermann, "Mechatronic Systems-Innovative Products with Embedded Control," *Control Engineering Practice,* Vol. 16, pp. 14-29, Jan. 2008

[9] E. Kiel and E. Kiel, *Drive Solutions: Mechatronics for Production and Logistics*. Springer Science & Business Media, 2008

[10] H. Kipphan, *Handbook of Print Media*. Springer, Germany, 2000, pp. 442-449

[11] A. Knobloch, "Mikroelektronikschaltungen aus gedruckten Polymeren," Technische Physik, Friedrich-Alexander-Uni., Germany, Erlangen-Nurnberg, 2003

[12] C. Kollmorgen, "AKD Servo Drive Datasheet," ed. USA: Kollmorgen, 2010

[13] F. C. Krebs, "Pad Printing as a Film Forming Technique for Polymer Solar Cells," *Solar Energy Materials and Solar Cells,* Vol. 93, pp. 484-490, Apr. 2009

[14] M. Lethiecq, R. Lou-Moeller, J. A. Ketterling, F. Levassort, L. P. Tran-Huu-Hue, E. Filoux*, et al.*, "Non-Planar Pad-printed Thick-Film Focused High-Frequency Ultrasonic Transducers for Imaging and HIFU Applications," in *International Symposium on Piezoresponse Force Microscopy and Nanoscale Phenomena in Polar Materials*, 2011, pp. 1-4

[15] F. Levassort, E. Filoux, M. Lethiecq, R. Lou-Moler, E. Ringgaard, and A. Nowicki, "P3Q-3 Curved Piezoelectric Thick Films for High Resolution Medical Imaging," in *Ultrasonics Symposium, IEEE*, 2006, pp. 2361-2364

[16] S. Merilampi, T. Björninen, L. Ukkonen, P. Ruuskanen, and L. Sydänheimo, "Characterization of UHF RFID Tags Fabricated Directly on Convex Surfaces by Pad Printing," *The International Journal of Advanced Manufacturing Technology,* Vol. 53, pp. 577-591, Mar. 2011

[17] C. Micro Print, "Pad Printing Catalog," 3 ed. Switzerland: Micro Print LC GmbH, 2014, p. 72

[18] T. Morlock, "Tampondruck GFG 100," ed. Dornstetten, Germany: ITW MORLOCK GmbH, 2012

[19] T. National Instruments (2012, Aug.), Applications for EtherCAT RIO, *White Papers* [Technical Document], Available: http://www.ni.com/white-paper/14083/en/

[20] T. National Instruments (2014, Oct.), CompactRIO Integrated Systems with Real-Time Controller and Reconfigurable Chassis NI cRIO-907x, [Datasheet], pp. 1-8, Available: http://sine.ni.com/ds/app/doc/p/id/ds-204/lang/en

[21] T. National Instruments (2012, Aug.), NI EtherCAT RIO: Deterministic Expansion for LabVIEW RIO Systems, *White Papers* [Technical Document], Available: http://www.ni.com/white-paper/7299/en/

[22] T. Proell (2014, Oct.), Pad Printing Theory and Practice, *Pad Printing Inks* [Thechnical data], pp. 1-38, Available: http://www.proell.de

[23] T. Robles, R. Alcarria, D. Martın, M. Navarro, R. Calero, S. Iglesias, *et al.*, "An IoT-based Reference Architecture for Smart Water Management Processes," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA),* Vol. 6, pp. 4-23, 2015

[24] D. Sharp, "Printed Composite Electrodes for In-Situ Wound pH Monitoring," *Biosensors and Bioelectronics,* Vol. 50, pp. 399-405, Dec. 2013

[25] C. TAMPOPRINT, "Pad Printing Machines," ed. Korntal-Münchingen, Germany: TAMPOPRINT AG, 2014, pp. 4-11

[26] D. P. Xenos, M. Cicciotti, G. M. Kopanos, A. E. F. Bouaswaig, O. Kahrs, R. Martinez-Botas, *et al.*, "Optimization of a Network of Compressors in Parallel: Real Time Optimization (RTO) of Compressors in Chemical Plants – An Industrial Case Study," *Applied Energy,* Vol. 144, pp. 51-63, Apr. 2015

[27] X. Ye and Q. Zengchao, "Antenna 3D Pad Printing Solution Evaluation," in *IEEE International Symposium on Antennas and Propagation (APSURSI)*, 2011, pp. 2773-2776