Determination of the Load Acting on the Axial Bearing of a Slewing Platform Drive in Hydraulic Excavators

Vesna D. Jovanović[1], Dragoslav B. Janošević[1], Dragan Z. Marinković[2]

[1] University of Niš, Faculty of Mechanical Engineering, A. Medvedeva 14, 18000 Niš, Serbia, vesna.nikolic@masfak.ni.ac.rs, janos@masfak.ni.ac.rs

[2] Technical University Berlin, Institute of Mechanics, 17. Juni 135, 10623 Berlin, Germany, Dragan.Marinkovic@TU-Berlin.de

Abstract: The paper presents a general selection procedure for an axial bearing of a slewing platform drive in hydraulic excavators based on the spectrum of equivalent bearing loads. A mathematical model of an excavator, with a backhoe and a shovel attachment, is defined to determine the spectrum of bearing loads on the basis of possible digging resistances specified in the entire working range of the excavator. As an example, by using the developed software, the size of an axial bearing was selected for a slewing platform drive in a hydraulic excavator with the mass of 100,000 kg, according to the spectrums of equivalent bearing loads obtained from the analysis of an excavator with a backhoe and shovel attachment.

Keywords: axial bearing; hydraulic excavators; slewing platform

1 Introduction

Hydraulic excavators perform their primary function of digging through a general configuration of the kinematic chain which consists of the support and movement mechanism L_1, Figure 1b,c, slewing platform L_2, and changeable multi-member manipulators L_m, which can be equipped with numerous tools in the form of buckets, claws, grapples, tillers, hammers, hooks, shown in Figure 1a. For digging operations below the ground level, the toward oneself technology (in relation to the excavator operator) is employed and a backhoe attachment is used, shown in Figure 1b. For digging operations above the ground level the away from oneself technology and a shovel attachment are used, shown in Figure 1c. Hydraulic excavators perform spatial manipulation using the slewing platform L_2, Figure 1b,c, which is attached to the support and movement mechanism L_1 by way of a

rotary joint, of the fifth class, in the form of an axial bearing. The slewing drive mechanism of the platform consists of a hydraulic motor 1 as shown in Figure 1d, a reducer 2 coupled over an output gear 2.1 with a ring gear of an axial bearing 3.1. By rule, an axial bearing 3 in Figure 1d consists of an inner ring gear 3.1, which is bolted to the support and movement mechanism L_1, and a toothless outer ring 3.2, which is bolted to the slewing platform L_2. Rolling elements (balls, rollers) are positioned between the rings in one or more races. The synthesis of the complete drive mechanism of a hydraulic excavator slewing platform is performed by the following procedure: a) selection of the concept drive solution, b) selection of the axial bearing, c) definition of attachment elements and elements of the support structure to which the bearing is attached, d) selection of the hydraulic motor and slewing drive reducer.

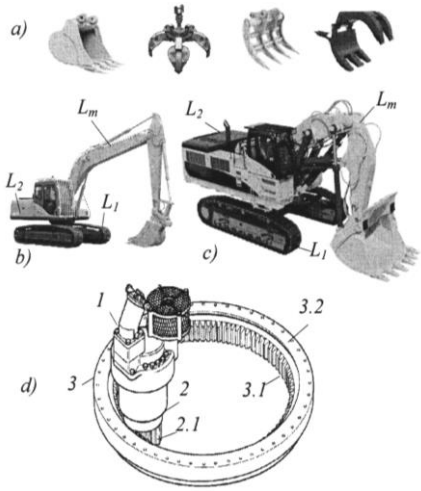


Figure 1
Hydraulic excavators: a) working tools, b) with backhoe attachment
c) with shovel attachment, d) drive mechanism of a hydraulic excavator slewing platform

In the design of the basic excavator systems, research were conducted into: a) analytical modelling and experimental determination of load during the digging process [1] and [2], b) development of mathematical models for kinematic and dynamic excavator analysis [3] and [4], c) development of drive mechanisms and control systems [5] and [6], and d) definition of indicators for analysis and evaluation of excavator digging efficiency [7]. Research into slewing drive mechanisms of excavator platforms deal with: a) loads of axial bearing rolling elements [8] to [10], b) analysis of axial bearing loads in excavators with excavating manipulators [11], and c) regulation of angular velocity of a slewing platform [12] to [14]. This paper provides a selection procedure for the size of an axial bearing of a slewing platform drive in hydraulic excavators with a backhoe and a shovel attachment, based on the spectrum of equivalent bearing loads.

2 Mathematical Model of the Excavator

The mathematical model of the excavator comprises the model of the kinematic chain and the mathematical models of excavator drive mechanisms. The mathematical model encompasses a five-member configuration of the excavator kinematic chain comprising: support and movement mechanism L_1 in Figure 2, slewing platform L_2, and a three-member planar attachment with: boom L_3, stick L_4 and bucket L_5. The space of the excavator model is determined with an absolute coordinate system *OXYZ* and unit vectors. The excavator support surface lies in the horizontal plane *OXZ* of the absolute coordinate system, while the vertical axis *OY* of the same system overlaps with the axis of the axial bearing of the slewing platform drive mechanism. Members of the excavator kinematic chain compose kinematic pairs of the fifth class - rotary joints with one degree of freedom. The centre of joint O_2 of the kinematic pair composed of the support and movement mechanism and the slewing platform is the point of perpendicular intersection of the vertical axis of the joint through the horizontal plane where the centres of rolling elements of the slewing platform drive mechanism axial bearing are positioned. The centres of manipulator joints O_i are points of intersection of the horizontal axis of joints through the plans of symmetry of the excavator manipulator kinematic chain.The intersection of the bucket cutting edge through the plane of manipulator represents the centre of the bucket cutting edge *Ow*.

The mathematical model of an excavator with a backhoe attachment is defined in the following section.

Each member of the excavator kinematic chain L_i is determined, in its local coordinate system $O_i\, x_i\, y_i\, z_i$, with a set of quantities:

$$L_i = \left\{ \hat{\vec{e}}_i, \hat{\vec{s}}_i, \hat{\vec{t}}_i, m_i \right\} \tag{1}$$

where: $\hat{\vec{e}}_i$ - the unit vector of joint O_i axis which connects member L_i to the previous member L_{i-1}, $\hat{\vec{s}}_i$ - the vector of the position of joint O_{i+1} centre which is used to connect the chain member L_i to the next member L_{i+1} (vector si magnitude represents the kinematic length of the member), $\hat{\vec{t}}_i$ - the vector of the position of the member mass centre, m_i - the member mass. Quantities marked with a 'cap' above the symbol are determined in the local coordinate system of the member.

The mathematical model of the excavator drive system encompasses the drive mechanisms of manipulator boom, stick, and bucket, which have two-way hydraulic cylinders c_3, c_4, and c_5 as actuators in Figure 2. Each drive mechanism C_i of the excavator manipulator is determined using a set of quantities:

$$C_i = \left\{ d_{i1}, d_{i2}, c_{ip}, c_{ik}, \hat{\vec{a}}_i, \hat{\vec{b}}_i, m_{ci}, n_{ci} \right\} \quad \forall i = 3,4,5 \tag{2}$$

where: d_{i1},d_{i2} - the diameter of the piston and piston rod of the hydraulic cylinder, c_{ip},c_{ik} - the initial and final length of the hydraulic cylinder, where: d_{i1},d_{i2} - the diameter of the piston and piston rod of the hydraulic cylinder, c_{ip},c_{ik} - the initial and final length of the hydraulic cylinder, $\vec{a}_i$,$\vec{b}_i$ - the vectors of the position of joint centres where the hydraulic cylinder is connected to the kinematic chain members, m_{ci} - the mass of the hydraulic cylinder, n_{ci} - the number of hydraulic cylinders of the drive mechanism.

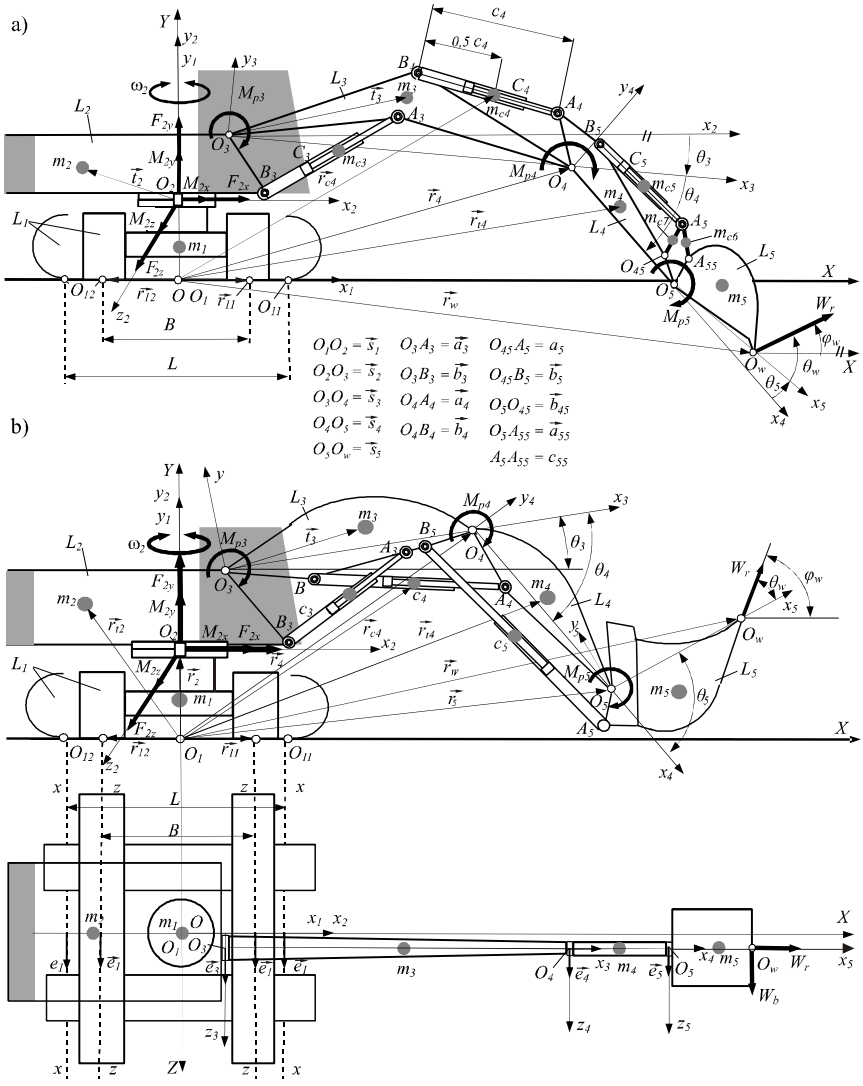


Figure 2
Determining the load of axial bearing slewing platforms of hydraulic excavator with a) backhoe attachment; b) shovel attachment

The subset of transmission parameters of the drive mechanism of the bucket C_5 is determined by the set:

- the vectors of the position of joint centres where the hydraulic cylinder is connected to the kinematic chain members, m_{ci} - the mass of the hydraulic cylinder, n_{ci} - the number of hydraulic cylinders of the drive mechanism.

The subset of transmission parameters of the drive mechanism of the bucket C_5 is determined by the set:

$$C_{p5} = \left\{ a_5, c_{55}, \vec{\hat{a}}_{55}, \vec{\hat{b}}_{45}, m_{c6}, m_{c7} \right\} \tag{3}$$

where: a_5 - the length of the lever of the bucket cylinder in the transmission part of the drive mechanism, Figure. 2; c_{55} - the length of the link in the transmission part of the drive mechanism; $\vec{\hat{a}}_{55}, \vec{\hat{b}}_{45}$ - the coordinates, of the position of the centre of joints in which transmission levers are connected to the links, m_{c6} - the mass of the link in the transmission part of the drive mechanism, m_{c7} - the mass of the lever of the bucket cylinder in the transmission part of the drive mechanism.

The assumptions of the mathematical model of the excavator are: 1) the support surface and kinematic chain members are modelled using rigid bodies, 2) the first joint of the kinematic chain between the movement mechanism surface represents a polygon bounded by potential longitudinal *x-x,* Figure 2 and transverse *z-z* excavator rollover lines. Inside the polygon, the first joint has the shape of a translatory - sliding joint, while on the edges of the polygon it has the shape of rotary joints O_{11},O_{12}, whose axes represent potential excavator rollover lines, 3) during the manipulation task the work of the excavator is stable, i.e. there are no potential movements in the first joint, 4) during the digging operation the kinematic chain of the excavator has an open configuration subjected to: a) gravitational forces (weights) of: kinematic chain members, members of the drive system and material scooped by the bucket and b) digging resistance W in the centre of the bucket cutting edge O_w, the position of the hydraulic cylinders mass centres is in the middle of the current length of hydraulic cylinders, 5) masses of joint elements belong to the members of the manipulator kinematic chain, 6) the influence of friction is neglected in the kinematic chain joints and drive mechanism joints. The digging resistance vector is determined with the equation:

$$\vec{W} = W_r \cos\varphi_w \vec{i} + W_r \sin\varphi_w \vec{j} + W_b \vec{k} \tag{4}$$

where: W_r - the digging resistance which acts in the plane of the manipulator, W_b - the lateral digging resistance, φ_w - the angle of the direction in which the digging resistance W_r acts in relation to the horizontal OXZ plane of the absolute coordinate system. The direction in which the digging resistance W_r acts in relation to the horizontal OXZ plane of the absolute coordinate system is determined with the angle:

$$\varphi_i = \sum_{3}^{i} \theta_i + \theta_w \quad \forall i = 3,4,5 \tag{5}$$

where: θ_i *(i=3,4,5)* - the angle of the relative position of member L_i in relation to the previous member L_{i-1} upon the rotation around the axis of joint O_i by changing the length c_i of the drive mechanism hydraulic cylinder, θ_w - the angle of the direction in which the digging resistance acts in relation to the positive O_5x_5 axis of the local coordinate system of the bucket L_5.

The magnitude of the digging resistance W_r vector, for a particular direction of action, is defined by the equation [15]:

$$W_r = \min\{W_o, W_1, W_3, W_4, W_5\} \tag{6}$$

where: W_o - the highest boundary digging resistance determined from the excavator non-sliding conditions in the plane of the support surface, W_1 - the highest boundary digging resistance determined from the given excavator stability conditions for potential rollover lines, W_3, W_4 ,W_5 - are the highest boundary values of the digging resistance which can be overcome by the drive mechanisms of manipulator boom, stick, and bucket at the maximum pressure of the excavator hydraulic system.

The unit vector of the digging resistance W_r:

$$ort\vec{W}_r = \cos\varphi_w \vec{i} + \sin\varphi_w \vec{j} \tag{7}$$

The boundary digging resistance W_o bounded by the force of adherence of the excavator to the support surface is determined from the balance conditions of the sliding part of the first joint, i.e. from the condition that the support and movement mechanism of the excavator will not slide, during digging, along the support surface:

$$W_o = \frac{mg \cdot \mu_p}{|\cos\varphi_w|} \tag{8}$$

where: m - the total mass of the excavator, μ_p - the coefficient of adherence of the excavator movement mechanism to the support surface.

Depending on the position of the kinematic chain of the excavator and, the boundary digging resistance W_1, which is limited by the static stability of the excavator, is determined from the balance conditions for one of the rotary joints O_{11},O_{12}, whose axes represent the potential excavator rollover lines, Figure 2:

$$W_1 = \begin{cases} W_{11} = \dfrac{-M_{o11}}{\left(\left(\vec{r}_w - \vec{r}_{11} \times ort\vec{W}_r\right)\cdot \vec{e}_1\right)}, \\ \forall y_w > 0,\ \varphi_{12} > \varphi_w > \left(\varphi_{11} + 180^\circ\right), \\ \forall y_w < 0, \varphi_{11} > \varphi_w > \left(\varphi_{12} + 180^\circ\right), \\ \\ W_{12} = \dfrac{-M_{o12}}{\left(\left(\vec{r}_w - \vec{r}_{12} \times ort\vec{W}_r\right)\cdot \vec{e}_1\right)}, \\ \forall y_w > 0 \left(\varphi_{12} + 180^\circ\right) > \varphi_w > \varphi_{11,} \\ \forall y_w < 0 \left(\varphi_{11} - 180^\circ\right) > \varphi_w > \varphi_{12}, \end{cases} \tag{9}$$

where: $\vec{e}_1 = \{0,0,1\}$ - the unit vector of the first rotary joint (for the longitudinal *x-x* or transverse *z-z* excavator rollover line), M_{o11}, M_{o12} - the gravitational moments for potential excavator rollover lines, i.e. rotary joints O_{11}, O_{12}, $\vec{r}_w$ - the vector of the position of the bucket cutting edge centre, $\vec{r}_{11}, \vec{r}_{12}$ - the vectors of the position of the centre of the appropriate first rotary joint O_{11}, O_{12}, y_w -the vertical coordinate of the bucket top, φ_{11}, φ_{12} - the angles of the position of vectors and in relation to the horizontal plane *OXZ*, determined by the equations:

$$\begin{aligned} \varphi_{11} &= \arccos\left(\frac{\left(\vec{r}_w - \vec{r}_{11}\right)\cdot \vec{i}}{\left|\vec{r}_w - \vec{r}_{11}\right|}\right), \\ \varphi_{12} &= \arccos\left(\frac{\left(\vec{r}_w - \vec{r}_{12}\right)\cdot \vec{i}}{\left|\vec{r}_w - \vec{r}_{12}\right|}\right) \end{aligned} \tag{10}$$

Gravitational moments for potential excavator rollover lines, i.e. rotary joints O_{11}, O_{12}:

$$M_{o1} = \begin{cases} M_{o11} = -g\sum_{k=1}^{k=5} m_k \left(\left(\vec{r}_{tk} - \vec{r}_{11}\right) \times \vec{j}\right)\cdot \vec{e}_1 - g\sum_{k=3}^{k=7} m_{ck}\left(\left(\vec{r}_{ctk} - \vec{r}_{11}\right)\times \vec{j}\right)\cdot \vec{e}_1, \\ M_{o12} = -g\sum_{k=1}^{k=5} m_k \left(\left(\vec{r}_{tk} - \vec{r}_{12}\right) \times \vec{j}\right)\cdot \vec{e}_1 - g\sum_{k=3}^{k=7} m_{ck}\left(\left(\vec{r}_{ctk} - \vec{r}_{12}\right)\times \vec{j}\right)\cdot \vec{e}_1 \end{cases} \tag{11}$$

where: m_k - the mass of the kinematic chain members, m_{ck} - the mass of the drive mechanism members, $\vec{r}_{tk}$ - the vector of the position of the mass centre of kinematic chain members, $\vec{r}_{ctk}$ - the vector of the position of the mass centre of drive mechanism members.

Boundary digging resistances Wi (i=3,4,5) which can be overcome by the drive mechanisms of the manipulator, for the known and the position of the excavator

kinematic chain upon the action of the maximum drive moments Mpi, are determined from the balance conditions for the manipulator joints Oi axes, Fig. 2:

$$W_i = \frac{-M_{pi} - M_{ri}}{\left((\vec{r}_w - \vec{r}_i) \times ort\vec{W}_r\right) \cdot \vec{e}_i} \quad \forall i = 3,4,5 \tag{12}$$

where: M_{pi} - the maximum drive moments of manipulator mechanisms for both directions in which they act (upon piston pushing and piston pulling in the hydraulic cylinder), M_{ri} - the moment of gravitational forces of the kinematic chain members, members of the excavator drive mechanisms, and the mass of soil scooped by the full bucket, for certain axes of joints O_i, $\vec{r}_i$ - the vector of the position of the joint centre in the excavator kinematic chain, $\vec{e}_i = \{0,0,1\}$- the unit vector of the joint axes in the manipulator kinematic chain.

The maximum drive moments of manipulator mechanisms for both directions in which they act (upon piston pushing and piston pulling in the hydraulic cylinder):

$$M_{pi} = \begin{cases} M_{pi1} = sign \quad \left(\dot{\theta}_i\right) \cdot r_{ci} \cdot n_{ci} \cdot \left[\dfrac{d_{i1}^2 \pi}{4} p_m - \dfrac{\left(d_{i1}^2 - d_{i2}^2\right)\pi}{4} p_o \right] \cdot \eta_{ci} \\ \forall \ i = 3,4,5; \ \dot{\theta}_3 > 0, \ \dot{\theta}_4 < 0, \ \dot{\theta}_5 > 0 \\ M_{pi2} = sign \quad \left(\dot{\theta}_i\right) \cdot r_{ci} \cdot n_{ci} \cdot \left[\dfrac{\left(d_{i1}^2 - d_{i2}^2\right)\pi}{4} p_m - \dfrac{d_{i1}^2 \pi}{4} p_o \right] \cdot \eta_{ci} \\ \forall \ i = 3,4,5; \ \dot{\theta}_3 < 0, \ \dot{\theta}_4 > 0, \ \dot{\theta}_5 < 0, \end{cases} \tag{13}$$

where: $\dot{\theta}_i$ - the angular velocities of the kinematic chain members, r_{ci} - the transmission function of the drive mechanism which depends on the length of the hydraulic cylinder and the vector, i.e. coordinates, of the position of the joint centres where hydraulic cylinders are connected to the members of the drive mechanism kinematic chain, p_m - the maximum duct pressure during the extension stroke of the hydraulic cylinder, p_o - the maximum duct pressure during the retraction stroke of the hydraulic cylinder, η_{ci} - the mechanical degree of the hydraulic cylinder efficiency.

The moment of the gravitational forces of the kinematic chain members, members of the excavator drive mechanisms, and the mass of soil scooped by the full bucket, for certain axes of joints O_i, is determined by the equation:

$$M_{ri} = M_{oi} - gm_z \left((\vec{r}_{t5} - \vec{r}_i) \times \vec{j}\right) \cdot \vec{e}_i \quad \forall i = 1,3,4,5 \tag{14}$$

where: M_{oi} - the moment of the gravitational forces of the kinematic chain members and members of the excavator drive mechanisms for certain axes of joints O_i, m_z - the mass of the material scooped with the bucket, where it is assumed that the centre of the scooped material mass overlaps with the centre of the bucket mass.

The moment of the gravitational forces of the kinematic chain members and members of the excavator drive mechanisms for certain axes of joints O_i, when the bucket is empty, is determined by the equation, Figure 2:

$$M_{oi} = -g\sum_{k=i}^{k=5} m_k \left(\left(\vec{r}_{tk} - \vec{r}_i\right)\times \vec{j}\right)\cdot \vec{e}_i + M_{oci} \quad \forall i = 3,4,5 \tag{15}$$

where: M_{oci} - the moment of the gravitational forces of the excavator drive mechanism members for certain axes of joints O_i *(i=3,4,5).*

The moments of the gravitational forces of the excavator drive mechanism members for certain axes of joints O_i *(i=3, 4, 5)* are determined by the following equations:

$$M_{oci} = \begin{cases} M_{oc3} = -g\dfrac{n_{c3}m_{c3}}{2}\left(\left(\vec{r}_{A3} - \vec{r}_3\right)\times \vec{j}\right)\cdot \vec{e}_3 - \\ -g\sum\limits_{k=4}^{k=7} n_{ck}m_{ck}\left(\left(\vec{r}_{ctk} - \vec{r}_3\right)\times \vec{j}\right)\cdot \vec{e}_3 \quad \forall\ i = 3 \\ M_{oc4} = -g\dfrac{n_{c4}m_{c4}}{2}\left(\left(\vec{r}_{A4} - \vec{r}_4\right)\times \vec{j}\right)\cdot \vec{e}_4 - \\ -g\sum\limits_{k=5}^{k=7} n_{ck}m_{ck}\left(\left(\vec{r}_{ctk} - \vec{r}_4\right)\times \vec{j}\right)\cdot \vec{e}_4 \quad \forall\ i = 4 \\ M_{oc5} = -g\dfrac{m_{c6}}{2}\left(\left(\vec{r}_{A5} - \vec{r}_5\right)\times \vec{j}\right)\cdot \vec{e}_5 \quad \forall\ i = 5 \end{cases} \tag{16}$$

where: $\vec{r}_{A3}, \vec{r}_{A4}, \vec{r}_{A5}$ - the coordinates of joints where hydraulic cylinders are connected to the kinematic chain members, Figure 2.

Depending on the position of the bucket, the mass of the material scooped by the bucket is defined by the expression:

$$m_z = \begin{cases} \rho_z \cdot V \cdot \left|\cos\varphi_5\right| & \forall\, 270^\circ \geq \varphi_5 \geq 90^\circ \\ 0 & \forall\ 270^\circ < \varphi_5 < 90^\circ \end{cases} \tag{17}$$

where: ρ_z - the density of the material, V - the volume of the bucket. The value of the lateral digging resistance W_b, for a particular position of the excavator kinematic chain, is defined by the equation:

$$W_b = \frac{m\cdot g\cdot L}{4\cdot x_w}\mu_o \tag{18}$$

where: m - the mass of the excavator, L - the length of the continuous tracks footprint in Figure 2, μ_o - the coefficient of the turning resistance of the tracks against the excavator support surface, x_w - the horizontal coordinate of the bucket cutting edge centre.

3 Bearing Loads

The fictive interruption of the kinematic chain of the excavator in the joint O_2 of the slewing platform L_2 and the reduction of all loads, of the removed part, into its centre, yield:

- the resulting force which subjects the axial bearing to loading:

$$\vec{F}_2 = \vec{W} - g\sum_{i=2}^{5} m_i \vec{j} - g\sum_{i=3}^{7} m_{ci} \vec{j} - g m_z \vec{j} \tag{19}$$

- and the resulting moment which subjects the axial bearing to loading:

$$\begin{aligned} \vec{M}_2 = &\left((\vec{r}_w - \vec{r}_2)\times \vec{W}\right) - g\sum_{i=2}^{5} m_i \left((\vec{r}_{ti} - \vec{r}_2)\times \vec{j}\right) - \\ &- g m_z \left((\vec{r}_{t5} - \vec{r}_2)\times \vec{j}\right) - g\sum_{i=3}^{i=7} n_{ci} m_{ci} \left((\vec{r}_{cti} - \vec{r}_2)\times \vec{j}\right) \end{aligned} \tag{20}$$

where: $\vec{r}_2$ - the vector of the position of the joint centre (axial bearing) O_2.

Components of force F_2 of joint O_2 along the coordinate axes:

$$F_{2x} = \vec{F}_2 \cdot \vec{i}, \quad F_{2y} = \vec{F}_2 \cdot \vec{j}, \quad F_{2z} = \vec{F}_2 \cdot \vec{k} \tag{21}$$

Components of moment M_2 of joint O_2 along the coordinate axes:

$$M_{2x} = \vec{M}_2 \cdot \vec{i}, \quad M_{2y} = \vec{M}_2 \cdot \vec{j} \quad M_{2z} = \vec{M}_2 \cdot \vec{k} \tag{22}$$

Components of axial bearing loads of the excavator slewing platform are:

- axial force:

$$F_{2a} = F_{2y} \tag{23}$$

- radial force:

$$F_{2r} = \left(F_{2x}^2 + F_{2z}^2\right)^{0,5} \tag{24}$$

- and moment:

$$M_{2r} = \left(M_{2x}^2 + M_{2z}^2\right)^{0,5} \tag{25}$$

Moment M_{2r}, whose vector lies in the horizontal plane, subjects the axial bearing to loading, while moment M_{2y}, whose vector direction matches the bearing axis, balances the drive moment of the platform rotation mechanism. The size of the bearing is selected on the basis of the determined equivalent spectrum of bearing loads and diagrams of bearing loading capacity (curves I, II, III, IV and V, Fig. 4), which are provided by the specialized bearing manufacturers [16].

The equivalent spectrum of bearing loads consists of an equivalent force and an equivalent bearing load moment determined by the equations for:

- equivalent force F_e:

$$F_e = \left(a \cdot F_{2a} + b \cdot F_{2r}\right) f_s \tag{26}$$

- and equivalent moment M_e:

$$M_e = f_s \cdot M_{2r} \tag{27}$$

where: a - the factor of the axial force influence, b - the factor of the radial force influence, f_s - the factor of the bearing working conditions. Values of factors a,b,f_s are provided by the bearing manufacturers depending on the type of bearing (single-row, multi-row, ball, roller), type and size of machines and their working conditions.

4 Selection of Bearings

For a reliable selection of an axial bearing of a slewing platform in a hydraulic excavator, of a certain size, it is necessary to determined the spectra of bearing loads for all possible configurations of kinematic chains that the excavator is equipped with. These possible configurations of kinematic chains differ from the variants of support and movement mechanisms, then the variants of manipulator members, and the tools which the excavator uses. Furthermore, it is also necessary to determine the spectrum of bearing loads for the same configuration of the excavator kinematic chain in as many positions of the entire working range of the excavator as possible, having in mind that each position of the kinematic chain carries the possibility of the action of the digging resistance in various directions depending on the excavator working conditions. To satisfy all of the above requirements, on the basis of the previously given calculation procedure, a computer programme was developed to determine the loading spectrum and select the axial bearing of the platform rotation drive in hydraulic excavators.

During the analysis the following is set at the programme input: L_i - parameters of the members of the excavator kinematic chains, C_i - parameters of the drive mechanisms of the excavator manipulator, p_m - the maximum pressure of the hydraulic static system of the excavator, p_o - the pressure in the retraction duct of the hydraulic static system of the excavator, N_3 - the desired number of the manipulator boom positions in its range of movement, N_4 - the desired number of the stick positions in its range of movement for a certain position of the manipulator boom, N_5 - the desired number of the bucket positions in its range of movement for a certain position of the manipulator stick, N_w - the desired number of changes in the angle θ_w of the directions in which the digging resistance acts for a certain position of the bucket, θ_{wp} - the initial angle of the direction in which the

digging resistance acts, θ_{wk} - the final angle of the direction in which the digging resistance acts, ρ_z - the density of the scooped material, V_z - the volume of the bucket, μ_p - the coefficient of adherence, a - the factor of the influence of the axial bearing force, b - the factor of the influence of the radial bearing force, f_s - the factor of the bearing working conditions.

Based on the input values, and through the cyclic change of the given numbers N_w, N_5, N_4 and N_3, Figure 3, the programme determines: a) geometric values (θ_i, r_i, r_{ti}, r_w) which define the position of the joint centres and mass centres of the excavator kinematic chain, b) loading moments (M_{oi}, M_{ri}) and drive moment (M_{pi}) of drive mechanisms, c) boundary digging resistances (W_o, W_1, W_3, W_4, W_5), for the entire working range of the excavator, d) components of axial bearing loads (F_{2a}, F_{2r}, M_{2r}), and e) equivalent axial bearing loads (F_e, M_e).

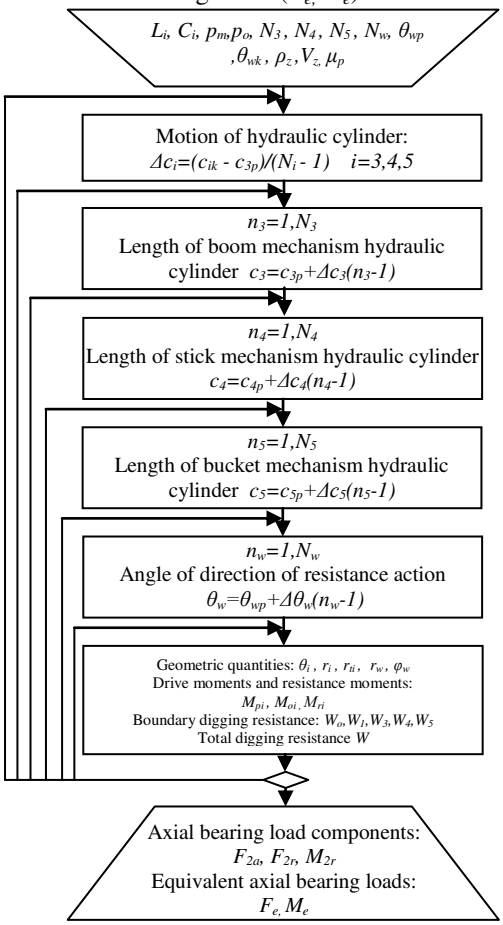


Figure 3
Algorithm of the programme for the analysis of axial bearing loads of a slewing platform drive in hydraulic excavators

The programme output yields a spectrum of bearing loads for the entire working range of the excavator which consists of equivalent bearing loads for each position of the excavator kinematic chain and for each given direction in which the digging resistance acts. By comparing the spectrum of the equivalent bearing loads obtained through analysis with the permitted bearing loads provided in the bearing loading capacity diagrams, the reliably necessary size of the bearing is selected.

5 Results and Discussion

By using the developed software, the analysis was conducted for the axial bearing load of a slewing platform drive in a hydraulic excavator with the mass of 100000 kg and the power of 400 kW with a backhoe and a shovel attachment.

The kinematic chain of the selected size of the excavator with a backhoe attachment can be equipped with: three different support and movement mechanisms L_1, Figure 2, three booms L_3 and four sticks L_4 of various lengths, and twelve backhoe attachments L_5 of various volumes for the digging of materials of different characteristics in different working conditions

The kinematic chain of the selected size of the excvator with a shovel attachment can be equipped with two different support and movement mechanisms L_1 and six shovel attachments L_5 of various volumes. As an example, the spectrum of axial bearing load of a slewing platform drive was determined for two possible variants A1 and A2 of the kinematic chain of the excavator with a backhoe attachment and for two possible variants B1 and B2 of the kinematic chain of the excavator with a shovel attachment.

Variant A1 consists of a support and movement mechanism with the footprint length of L=4.64 m, Figure 2, track distance of B=3.6 m, boom with the length of s_3=7.2 m, stick with the length of s_4=2.9 m, and a backhoe attachment with the volume of V= 4.8 m^3 for the digging of a material with the density of ρ_z = 2200 kg/m^3.

Variant A2 consists of a support and movement mechanism with the footprint length of L=5.035 m, track distance of B=3.6 m, boom with the length of s_3=10.5 m, stick with the length of s_4=5.8 m, and a backhoe attachment with the volume of V= 2.0 m^3 for the digging of a material with the density of ρ_z = 1800 kg/m^3.

Variant B1 consists of a support and movement mechanism with the footprint length of L=4.64 m, track distance of B=3.6 m, boom with the length of s_3=4.65 m, stick with the length of s_4=3.4 m, and a shovel attachment with the volume of V= 4.4 m^3 for the digging of a material with the density of ρ_z = 2200 kg/m^3.

Variant B2 consists of a support and movement mechanism with the footprint length of L=5.035 m, track distance of B=3.6 m, boom with the length of s_3=4.65

m, stick with the length of s_4=3.4 m, and a shovel attachment with the volume of V= 6.5 m^3 for the digging of a material with the density of ρ_z = 1650 kg/m^3.

The input file of the programme contains the parameters shown in Table 1 which are the same for all variants A1, A2, B1, and B2 of the kinematic chain of the excavator.

Table 1
Values of the input parameters of the programme

Parameters	p_m	p_o	N_3	N_4	N_5	N_w	μ_p	a	b	f_s
Values	32 MPa	1.2 MPa	30	20	10	10	0.85	1	2.05	1.45

The determination of the spectrum of bearing loads of variants A1 and A2 includes the range of the change in the acting angle of a possible digging resistance between Θ_{wp}=30° and Θ_{wk}=150°. The determination of the spectrum of bearing loads of variants B1 and B2 includes the range of the change in the acting angle of a possible digging resistance between Θ_{wp}=200° and Θ_{wk}=300° .

On the basis of the set input parameters, using the developed programme, the spectrum of axial bearing loads was determined for the slewing platform drive for variants A1,A2, B1 and B2 of the excavator kinematic chain. The programme output provided, among other things, the equivalent force and the equivalent bearing load moment for variants A1,A2, B1 and B2 of the excavator kinematic chain, determined for 6000 (N_3xN_4xN_5) positions in the entire working range and in each position for 10 different directions in which potential digging resistances could act.

The obtained values of the equivalent forces and moments are shown in the form of a diagram shown in Figure 4, Figure 5 as a spectrum of bearing loads for the longitudinal (x-x) and transverse (z-z) potential rollover lines for variants A1,A2, B1 and B2 of the excavator kinematic chain. All of the diagrams of the spectra of axial bearing loads for variants A1, A2, B1 and B2 of the excavator kinematic chain show diagrams of the permitted loading capacity of the five same bearing which differ in size are shown in Figure 4, curves I, II, III, IV, V) [16]. The diagram of the permitted loading capacity represents the dependency of the allowed axial bearing moment and force. The slewing platform drive corresponds to that size of the bearing whose permitted loading capacity is closest yet larger than the potential values of the spectrum of the equivalent bearing loads.

For example, by comparing the spectra of the equivalent bearing loads for variant A1 of the excavator kinematic chain with the loading capacity diagrams it can be noticed that the potential longitudinal rollover line (x-x), Figure 4a corresponds to the bearing size V, while for variant A2 the transverse rollover line (z-z), Figure 4b corresponds to the bearing size II.

As far as variant B1 of the excavator kinematic chain is concerned, the potential longitudinal rollover line (x-x), Figure 5c, corresponds to the bearing size IV, while for variant B2 the transverse rollover line (z-z), Figure 5d, corresponds to the bearing size II.

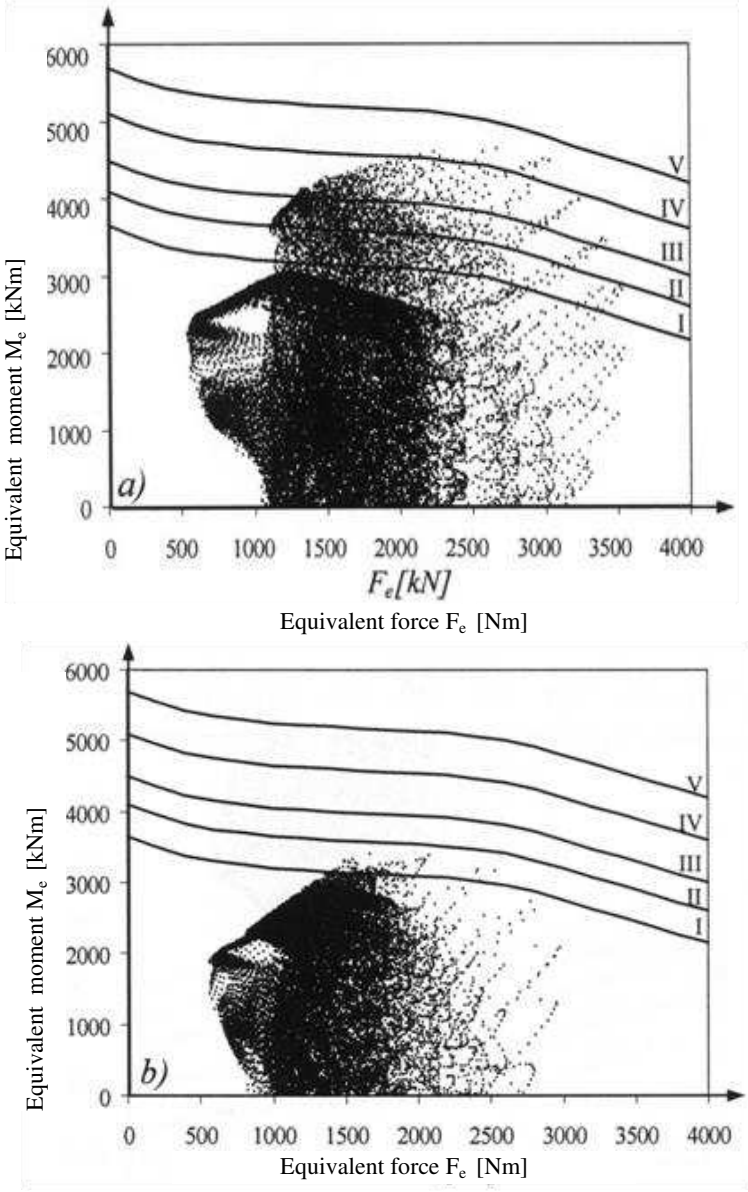


Figure 4

Spectrums of axial bearing load of a slewing platform drive of hydraulic excavators: a) variant A1 - longitudinal rollover line x-x, b) variant A2 - transverse rollover line z-z

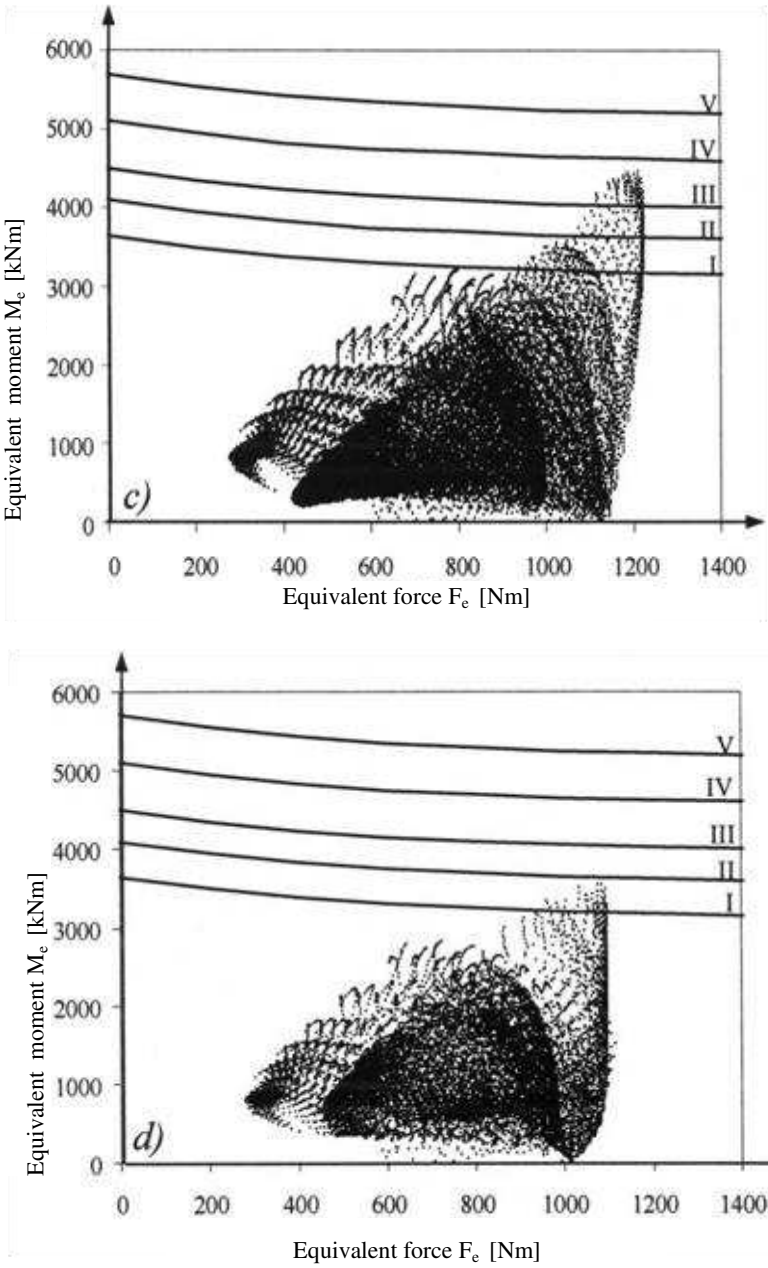


Figure 5

Spectrums of axial bearing load of a slewing platform drive of hydraulic excavators: c) variant B1 - longitudinal rollover line x-x, d) variant B2 - transverse rollover line z-z

Finally, on the basis of the obtained loading spectra and diagrams of permitted bearing loading (Figs. 4, 5), the analyzed excavator variants A1, A2 and B1, B2 correspond to the bearing size V which belongs to the KD 800 series of single-row roller slewing bearings with internal toothing manufactured by Rothe Erde [16].

Conclusions

The synthesis procedure for the slewing platform drive of hydraulic excavators should, among other things, provide for the selection of the axial bearing size. It is characteristic for all sizes of hydraulic excavators that the same model of excavator can have different configurations of kinematic chains equipped with various working tools. Furthermore, an excavator with the same configuration of the kinematic chain has a number of different positions and working conditions during its operation in its working range. For these reasons, this paper defines a mathematical model upon which software was developed that enables a comprehensive analysis of equivalent axial bearing loads of an excavator slewing platform which allow for a reliable selection of the bearing size. The comprehensive analysis using the developed software includes the determination of the spectra of equivalent axial bearing loads for a slewing platform in each possible configuration of the excavator kinematic chain for a desired number of working positions in the entire working range and in every position for a desired number of different working manners and conditions.

Acknowledgement

The paper was done within the project TR 35049 financed by the Ministry of Education and Science of the Republic of Serbia.

References

[1] Maciejewski, J., Jarzebowski, A., Trampczynski, W. (2004) Study on the Efficiency of the Digging Process using the Model of Excavator Bucket, *Journal of Terramechanics*, Vol. 40, pp. 221-233

[2] Tanasijevic, M., Ivezic, D., Ignjatovic, D. and Polovina, D. (2011) Dependability as Criteria for Bucket Wheel Excavator Revitalization, *Journal of Scientific & Industrial Research (JSIR)*, Vol. 70, pp. 13-19

[3] Gu, J., Taylor, J., Seward, D. (2007) Modelling of an Hydraulic Excavator using Simplified Refined Instrumental Variable (SRIV) Algorithm, *Journal of Control Theory and Applications,* No. 5-4, pp. 391-396

[4] Hall, A. S., McAree, P. R. (2005) Robust Bucket Position Tracking for a Large Hydraulic Excavator, *Mechanism and Machine Theory,* Vol. l40, pp. 1-16

[5] Plonecki, L., Trampczynski, W., Cendrowicz, J. (1998) A Concept of Digital Control System to Assist the Operator of Hydraulic Excavators, *Automation in Construction,* Vol. 7, pp. 401-411

[6] Geu, F., Kecskemethy, F., Pottker, A. (2007) Workspace Analysis and Maximal Force Calculation of a Face-Shovel Excavator using Kinematical Transformers, *12th IFToMM World Congress, Besancon*, pp. 18-21

[7] Sung-Uk, L., Chang, P. H. (2002) Control of a Heavy-Duty Robotic Excavator using Time Delay Control with Integral Sliding Surface, *Control Engineering Practice* 10 697-711

[8] Tadeusz, S., Damian, D., Mariusz, S., (2008) Evaluation of Load Distribution in the Superstructure Rotation Joint of Single-Bucket Caterpillar Excavators, *Automation in Construction*, Vol. 17-3, pp. 218-223

[9] Jose, I. A., Xabie, S., Jorge, D. (2003) Load Distribution in a Four Contact-Point Slewing Bearing, *Mechanism and Machine Theory,* Vol. 38-6, pp. 479-496

[10] Hedrih, K., Veljović, Lj. (2013) New Vector Description of Kinetic Pressures on Shaft Bearings of a Rigid Body Nonlinear Dynamics with Coupled Rotations Around No Intersecting Axes, *Acta Polytechnica Hungarica,* Vol. 10-7, pp. 151-170

[11] Janosevic, D., Nikolic, V., Petrovic, N. (2012) Determining the Load Spectrum of Axial Bearing Slewing Platforms of Hydraulic Excavator, *XX International Conference on "Material Handling, Constructions and Logistics", Belgrade*, pp. 177-180

[12] Bin, Y., Jiao, Z., Douglas, K., John, L. (1998) High Performance Swing Velocity Tracking Control of Hydraulic Excavators, *American Control Conference, Philadelphia,* pp. 818-822

[13] Jianqi, L. (1992) *An Energy-Saving Device Applied to the Swing System of Hydraulic Excavator, Presented at the International Fluid Power Exposition and Technical Conference, Milwaukee,* p. I92-3.2

[14] Jacek, K. (2005) Swing-Free Stop Control of the Slewing Motion of a Mobile Crane, *Control Engineering Practice,* Vol. 13-4, pp. 451-460

[15] Janosevic, D. (1997) Optimal Synthesis of Drive mechanisms in Hydraulic Excavators. *PhD Thesis,* Faculty of Mechanical Engineering, University of Nis, Serbia

[16] Slewing Bearings, Rothe Erde GmbH,D-44137 Dortmund (2007) catalog. Available on http://www.thyssenkrupp-rotheerde.com/gb/ produkte_gwl.shtm

Biomass-fired Boiler Control Using Simulated Annealing Optimized Improved Varela Immune Controller

Alexander Hošovský

Department of Mathematics, Informatics and Cybernetics, Faculty of Manufacturing Technologies with seat in Prešov, Technical University of Košice, Bayerova 1, 08001 Prešov, Slovakia, e-mail: alexander.hosovsky@tuke.sk

Abstract: Water temperature control of biomass-fired boilers represents a process with extreme delay which makes it quite difficult to stabilize it using conventional PID controller tuned according to commonly used PID tuning rules. It is proposed here to use a controller based on Varela-Countinho second generation immune network which was shown to have very good anti-delay capabilities. Since there are currently no tuning rules for this type of controller, simulated annealing algorithm is used for optimizing the set of controller parameters to achieve good performance according to IAE criterion. The resulting controller is shown to offer stable performance even for such a long time-delay which is also robust for up to 30% variations in system time constant compared to a nominal case.

Keywords: B-cells; antibody; temperature; optimization; controller

1 Introduction

An effective control of biomass combustion and biomass-fired boilers in general in regard to meeting ever stricter requirements on emission levels remains a challenging task [1], [2], [3]. In works [2] and [3] the aspects of biomass-fired boiler operation in terms of carbon monoxide and oxygen concentration relationship are analyzed. Despite the complexities of processes associated with biomass combustion, the commonest controller type applied to this task is still a conventional PID controller [4], [5], [6]. In work [5] the ways of improving the performance of current control techniques are inspected while in [6] Smith predictor is applied to compensate for the performance deterioration due to long time delay of a boiler. The prevalence of PID control can almost certainly be attributed not that much to PID's satisfying performance in that task but more probably to its ease of use, multitudes of tuning rules and good comprehensibility. Nevertheless, ambitions for improving the performance of biomass-fired boiler control using more advanced techniques are still present, e.g. in works [5], [7], [8].

The works [7] and [8] present a preliminary analysis of possible intelligent techniques applied to the control of biomass-fired boilers. In [6] a control design model for medium scale biomass-fired boiler was developed. Precisely, this kind of plant should be treated as multivariable system with several inputs and several outputs that are cross-linked. However, based on the measured data it was shown that water temperature control could be treated as single input single output (SISO) system with fuel feed as an input variable and water temperature as an output variable. An identified time-delay for this system is 480 s (for increasing fuel feed), which makes it difficult to be stabilized using standard tuning rules for PID controllers. A method of time-delay compensation using Smith predictor (e.g. [9]) is proposed – this, however, requires the model to be equivalent to the plant, which might not always be fulfiled in practice and this may lead to deterioration of controller's performance. It is therefore proposed to use Varela immune controller which was shown [10], [11] to have very good anti-delay capabilities. Since there are no tuning rules for this kind of controller and the effect of its parameters on controller's performance is less intuitive than in case of PID controller, it is also proposed to optimize its parameters using simulated annealing. It is shown that this combination provides a stable and robust performance even for a system with such long time-delay.

2 Used Methods

2.1 Improved Varela Immune Controller

Varela immune controller belongs to the group of nonlinear controllers which make use of some of natural immune system paradigms. The theory of immune controllers was extensively studied by Chinese researchers [10], [11], [12]. Varela immune controller is based on the idea of second generation immune networks proposed by Varela and Countinho in [13]. This model takes into account the interactions between B-cells and antibodies present as free soluble molecules. The central idea lies in the relationship between B-cell maturation (only fully mature B-cell can produce antibodies) and proliferation probability and cumulative receptor occupancy (or sensitivity). This characteristic has the form of two mutually shifted bell functions from which the basic fact that B-cells are activated only at intermediate receptor occupancy can be observed (Fig. 1).

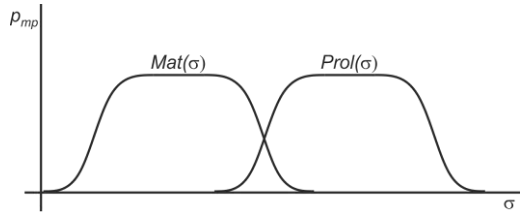


Figure 1
Probability functions for B-cell maturation and proliferation in relation to network sensitivity

The dynamics of second generation immune network can be described by two differential equations [13], [14]:

$$\frac{df_i}{dt} = -f_i(k_1 + k_2\sigma_i) + k_3 b_i Mat(\sigma_i) \tag{1}$$

$$\frac{db_i}{dt} = b_i(k_4 Prol(\sigma_i) - k_5) + Meta[i] \tag{2}$$

where k_1 – antibody natural death rate, k_2 – antibody interaction death rate, k_3 – antibody production rate by fully matured B-cells, k_4 – B-cell proliferation rate, k_5 – B-cell death rate, f_i – free soluble antibody with i idiotype, b_i – B-cell with i idiotype, $Meta[i]$ – additional cells from the resting pool, $Mat(\sigma_i)$ – maturation probability function, $Prol(\sigma_i)$ – proliferation probability function. The symbol σ_i is the sensitivity of network to i idiotype defined as:

$$\sigma_i(t) = \sum_{j=1}^{N} m_{ij} f_j \tag{3}$$

where m_{ij} is affinity between i and j idiotype.

It is clear from (1) and (2) that the dynamics of free antibodies as well as B-cells with specific idiotype are related to the network sensitivity to this idiotype (i.e. receptor occupancy).

It is possible to construct a controller which uses the foundations of second generation immune networks on the condition that some special assumptions relevant to control theory are made. The theory of second generation immune networks describes the activity of immune system in the absence of antigens [10]. The adaptation of immune system theory to automatic control, however, uses the notion of antigen (as a foreign element that needs to be eliminated) to represent a control error. It is therefore necessary to incorporate the effects of antigen presence into the immune controller model. According to [15], the rate of antigens can be given by the following equation:

$$\dot{a}_i = k_{ag} a_i - k_e f_i a_i \tag{4}$$

where k_{ag} – multiplication rate of antigens, k_e – elimination rate of antigens, a_i – antigen with i idiotype. Moreover, the *Meta* term in equation (2) might be modified in order to account for B-cells supplied from bone marrow modeled by constant k_{bm} and B-cells proliferation due to antigen presence modeled by factor $k_{bag}a_i$. The final model used for constructing Varela controller can be summarized as follows:

$$\begin{aligned}
\frac{da_i}{dt} &= k_{ag}a_i - k_e f_i a_i \\
\frac{df_i}{dt} &= -f_i(k_1 + k_2\sigma_i) + k_3 b_i Mat(\sigma_i) \\
\frac{db_i}{dt} &= b_i(k_4 Prol(\sigma_i) - k_5) + k_{bm} + k_{bag}a_i \\
\sigma_i(t) &= \sum_{j=1}^{N} m_{ij} f_j
\end{aligned} \tag{5}$$

The four equations (5) constitute the biological model of immune network dynamics according to Varela-Countinho with addition of antigen dynamics equation (first equation in (5)). This biological model has to be transformed into a model that would use control variables needed for its function as a controller. The interactions in (5) are shown in Fig. 2 (with + sign denoting stimulation effect and – sign denoting suppression effect) together with certain analogies between biological variables and control variables.

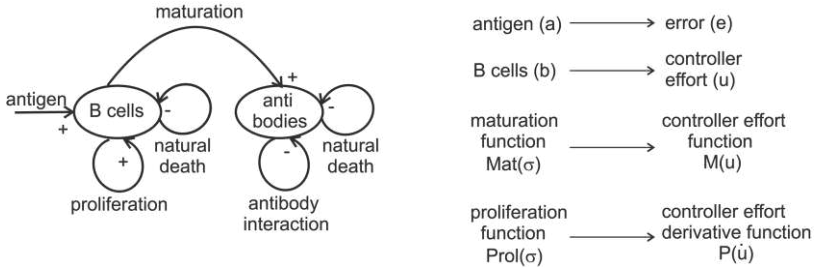


Figure 2

Interactions in biological model of Varela immune network with inclusion of antigen effect (left) and some analogies between biological and control model (right)

According to [12], two assumptions are made in regard to the implementation of Varela controller when compared to its biological model shown in (5) –

1) Antigen is capable of replication in human body, which is modeled by the first term in the first equation of biological model. As was mentioned above, antigen is treated as a control error in immune controllers but there is no analogy for its replication in control systems and thus it is not considered in the development of Varela immune controller.

2) Also, in contrast to B cell and antibody concentration, controller effort and control error can assume negative values. Maturation and proliferation functions were thus modified in the following way:

$$M(u) = k_M\,(e^{-\alpha_1|u|} - e^{-\alpha_2|u|})\,sign(u)$$
$$P(\dot{u}) = k_P\,(e^{-\alpha_1|\dot{u}|} - e^{-\alpha_2|\dot{u}|})\,sign(\dot{u}) \tag{6}$$

where α_1, α_2 – negative constants and $\alpha_2 < \alpha_1$, and k_M, k_P – positive constants. Also the network sensitivity σ was replaced with u in maturation function and with $\dot{u}$ in proliferation function. The shape of this function is depicted in Fig.3 for normalized values of respective variables.

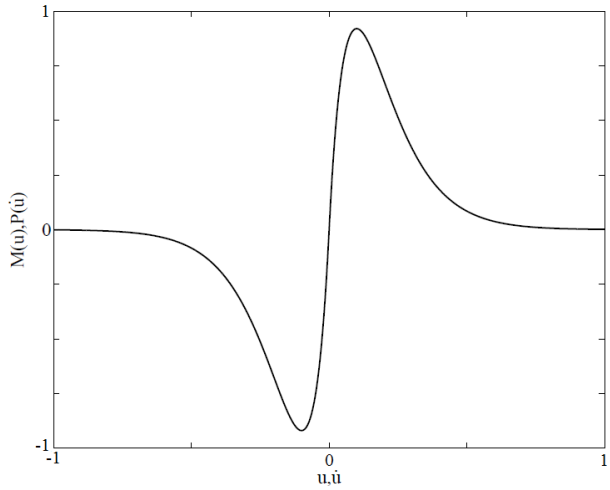


Figure 3
Maturation and proliferation function in Varela immune controller

Using the abovementioned assumptions and analogies shown in Fig. 2 we may rewrite the equations (5) to the following form:

$$\frac{de}{dt} = -k_e f(t) e(t)$$
$$\frac{df}{dt} = -k_f f(t) + k_3 M(u) u(t) \tag{7}$$
$$\frac{du}{dt} = u(t)(k_4 Prol(\dot{u}(t)) - k_5) + k_{bm} + k_{bag} e(t)$$

where in the first equation the self-replication term of antigen (now control error) has been removed and the term $(k_1 + k_2\sigma)$ has been replaced with k_f. It can also be seen that the antibody concentration f has been preserved also in Varela immune controller model as there is no analogy for this variable in controller design

(actually *f* is one of the three state variables of Varela controller, other two being *u* and *du/dt*). In order to implement a practical controller, it was proposed in [11] to further modify equations in (7) by taking time derivative of the last equation, neglecting the second term in derivative of $u(t)k_4 Prol(\dot{u}(t))$ and substituting the control error rate equation into this equation to get:

$$\begin{aligned} \frac{df}{dt} &= -k_f f(t) + k_3 M(u)u(t) \\ \frac{d^2u}{dt^2} &= \dot{u}(t)k_4 Prol(\dot{u}(t)) - k_5\dot{u}(t) + k_I f(t)e(t) \end{aligned} \tag{8}$$

where $k_I = k_e k_{bAg}$. This is the final model of Varela immune controller used for experimentation.

2.2 Simulated Annealing

Simulated annealing is an optimization technique based on the idea of annealing process in metalurgy. This process is described as a way of obtaining low energy state of a solid in heat bath [16]. First a solid is heated to a high temperature at which the atoms in lattice are arranged randomly and then it is cooled in a controlled manner so that a regular crystal lattice is formed – this state corresponds to a state with minimal energy. If the temperature is decreased to quickly, a state with higher than minimal energy (corresponding to a polycrystalline structure) is attained. Inventors of simulated annealing algorithm used an analogy of this process for optimization technique where the state of a solid corresponds to a problem solution, energy associated with a current state corresponds to a fitness function value and the temperature represents a control parameter. The points within the vicinity of current solution are generated using the following function [17]:

$$g(\mathbf{X_C}, \mathbf{T}) = \mathbf{X_c} + \mathbf{T_c} \frac{y(x)}{\|y(x)\|_2}; y(x) = \frac{1}{2\pi} \exp^{\frac{-x^2}{2}} \tag{9}$$

where $\mathbf{X_c}$ – *n*-dimensional current point vector, $\mathbf{T_c}$ – current temperature vector, *y(x)* – pseudorandom generator function, *x* – a number taken from the default random stream and $\|.\|_2$ - 2-norm. In this implementation, the perturbation of current solution depends on current temperature (represented as a vector, i.e. can be different for every element of solution vector). The acceptance function decides whether the generated solution is accepted or not (if the new solution decreases the fitness function value it is certainly accepted while if it increases the fitness function value it can still be accepted with a probability given below):

$$\Delta f = f_n - f_c < 0, P(\mathbf{X_n} = \mathbf{X_c}) = 1$$
$$\Delta f = f_n - f_c > 0, P(\mathbf{X_n} = \mathbf{X_c}) = \frac{1}{1 + \exp^{\frac{\Delta f}{\max(\mathbf{T_c})}}} \tag{10}$$

where f_n – cost function for a new point, f_c – cost function for a current point, *P(.)* – probability function.

The performance of simulated annealing is strongly dependent on the selection of cooling schedule which was chosen to be of exponential type. The schedule could be expressed in this form [17]:

$$\mathbf{T_n} = \mathbf{T_c} 0.95^{\mathbf{k}} \tag{11}$$

where $\mathbf{T_n}$ – new temperature vector, $\mathbf{k}$ – annealing parameter vector. The implementation of simulated annealing algorithm had also capability of reannealing (repeated increase in temperature vector values according to the sensitivity of fitness function to the changes in respective parameters) expressed in the form of sensitivity function [17]:

$$\mathbf{S} = \mathbf{R} \frac{\partial f(\mathbf{X_c})}{\partial \mathbf{X_c}}, \mathbf{k} = \left| \ln \left| \frac{\mathbf{T_0}}{\mathbf{T_c}} \frac{\max(\mathbf{S})}{\mathbf{S}} \right| \right| \tag{12}$$

where $\mathbf{S}$ – parameter sensitivity vector, $\mathbf{R}$ – parameter range vector, $\mathbf{T_0}$ – initial temperature vector and max(.) is maximum element in vector.

2.3 Plant

The model of plant used for experimentation was based on the identification carried out for a biomass-fired boiler in [6]. This boiler was a medium-power unit (1 MW) fed by various types of biomass fuels: woodchips, bark, wood waste, agricultural plants and straw. From the control point of view it is a process with very long time delay owing to the fact that it takes approximately 60 seconds to feed the fuel by feeder and some additional time to deliver the fuel by grate into the furnace. Thus, the overall time delay was found out to be around 7 to 8 minutes [6]. The system itself was considered to be of single-input single-output type where the input variable was fuel feed [$kg.h^{-1}$] and the output variable was water temperature [^oC]. It was found out that the system responses to increase and decrease in fuel feed were qualitatively different and therefore two distinct linear transfer functions were identified – one for the case when the fuel feed rate was positive and other one for the case when it was negative (Fig. 4). These two transfer functions ($G_{B+}(s)$ and $G_{B-}(s)$) had the following form:

$$G_{B+}(s) = \frac{0{,}0615}{(200s+1)^4} e^{-480s} \qquad G_{B-}(s) = \frac{0{,}06}{(165s+1)^2} e^{-150s} \tag{13}$$

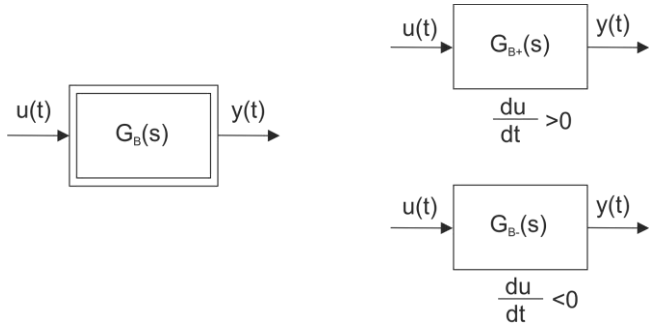


Figure 4
Nonlinear plant transfer function represented by two different linear transfer functions for increasing and decreasing fuel feed (i.e.input variable) (according to [6])

The main disturbances that can be considered for biomass-fired boilers are: thermal losses, fluctuations in heat demand and variable fuel capacity. Of these, the second one is considered also in the presented model. The fluctuations in heat demand will eventually be reflected in the changes of water temperature at the output of boiler.

3 Results and Discussion

3.1 Problem Description and Initialization

From the control point of view, the water temperature control system of biomass-fired boilers is specific especially due to its excessive time-delay which significantly obstructs the process of obtaining good performance. As was shown in [6], standardly it was difficult to stabilize the plant using the experience-based tuning of a controller. Thus, some kind of a controller that could not only stabilize the plant but possibly also achieve robust performance for this kind of plant was needed.

In Fig. 5 the block diagram of water temperature control using simulated annealing optimized Varela immune controller is shown. The following nomenclature is used : *w(t)* – desired water temperature, *e(t)* – water temperature control error, *u(t)* – fuel feed (control action), *v(t)* – heat demand fluctuations (disturbance), *y(t)* – actual water temperature. The control action was upper limited with a value of 350 $kg.h^{-1}$ (represented as saturation block at the controller output). According to equation 8, Varela immune controller contains several free parameters that affect its performance:

k_f – combined parameter of antibody death rate (due to natural death of antibodies as well as due to their interaction)

k_3 – antibody production rate by mature B-cells

k_4 - B-cell proliferation rate

k_5 – B-cell death rate

k_1 – combined parameter of antigen elimination rate and B-cell proliferation rate due to antigen presence.

In addition to these parameters, another three parameters could also be taken into account: initial value of antibody concentration (represented as initial setting of antibody integrator) and scaling coefficients for proliferation and maturation function. These three parameters were not incorporated into optimization and were set in advance based on the trial-and-error results. Furthermore, it was found out that the best results were obtained when k_5 parameter (i.e. B-cell death rate) was set to zero. Even though the number of modifiable parameters was thus reduced to four, it would have been difficult to achieve optimal settings using manual tuning as the interactions between parameters were quite complex.

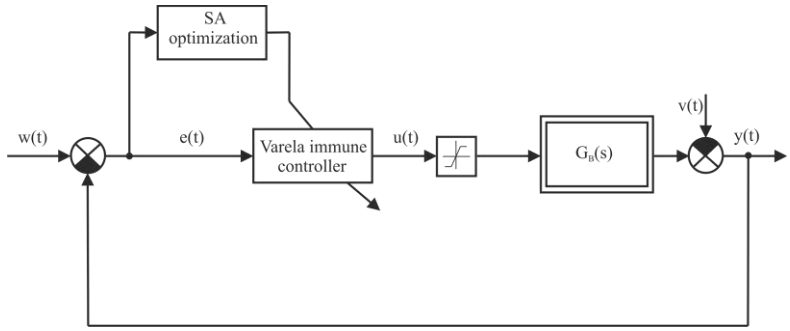


Figure 5
Block diagram of water temperature control using Varela immune controller with simulated annealing optimization

To achieve good Varela controller settings (good in the sense of near optimal solution to the optimization problem), simulated annealing algorithm was used. The fitness function for simulated annealing algorithm was in the form of integral absolute error criterion (IAE):

$$IAE = K_S \sum_{k=1}^{n} \left| T_{rk} - T_k \right| \tag{14}$$

where K_S – scaling constant set to 0.001, T_{rk} – reference temperature for *k-th* sample [°C], T_k – actual temperature for *k-th* sample [°C], n – number of samples.

Several parameters and functions of the simulated annealing algorithm needed to be set before the optimization. The main three functions (generation and acceptance functions and temperature schedule) forming the basis of the algorithm are expressed in equations (9)-(11). The main control parameters were initial temperature, reannealing interval and number of stall generations (stopping criterion) which were respectively set to the following values: T_0 = 150, η_R = 100 it and η_{SG} = 2000 it.

3.2 Optimization and Control Test

Since the transfer functions representing the system for increasing or decreasing fuel feed were quite different, it was necessary to optimize the controller for both situations separately (it would have been difficult to obtain satisfactory performance with only one set of controller parameter values). Nevertheless, the situations for which different gain values should have been used were considered easily discernible (increasing or decreasing fuel feed) and were not therefore viewed as obstruction to the use of Varela controller. Even when treated separately for increasing or decreasing fuel feed, the whole control system had to be considered nonlinear due to the Varela controller's nature causing the system performance to be dependent also on excitation signal magnitude. Moreover, in order to avoid extreme immune response of the Varela controller it was necessary to optimize the performance for several consecutive steps of desired water temperature (using only one step would result in very fast yet undesirable response due to the controller's insensitivity to future changes in reference water temperature).

In Fig. 6 the reference temperature step pattern used for optimization is shown. This pattern consists of four consecutive steps in reference temperature change (denoted as ΔT_r) in the range of 0-8 ^{o}C with the time duration of 15000 seconds for each step. The same step pattern was used for optimization of both increasing and decreasing fuel feed with using transfer functions given in (13). The results of both optimization runs are found in Table 1 with the graphs of best and current fitness values as a function of iteration count shown in Fig. 7. The results obtained using simulated annealing algorithm were improved using a local optimizer (LO).

An exponential cooling schedule was preferred over logarithmic or linear due to being much faster than other two schedules. On the other hand, excessive speed of the temperature decrease might be detrimental to optimality of the obtained result. It was thus helpful to use reannealing interval after which the temperature was increased again so that a probability of accepting worse point was higher (and algorithm could jump out of a local minimum). The reannealing points can be seen as spikes in current function value on fitness-iteration graphs in Fig. 7. In both cases the results could be further improved using Nelder-Mead algorithm.

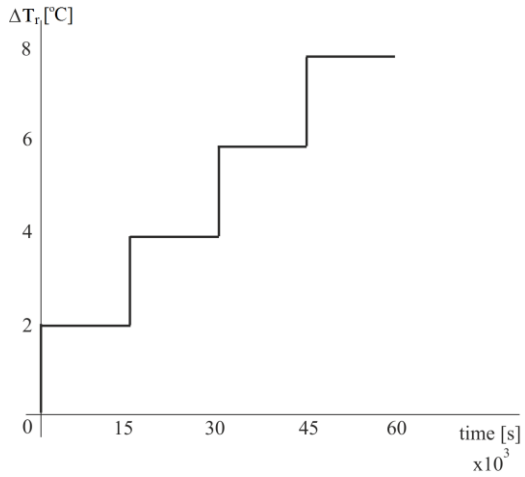


Figure 6
Reference temperature step pattern used for optimization of Varella controller values

Table 1
The results of hybridized SA optimization for increasing and decreasing fuel feed

	du/dt > 0	du/dt < 0
SA Iterations	3375	5290
SA Fitness Value	24.881	10.103
Optimization Time	6400	9889
SA Function Evaluations	3401	5335
LO Iterations	104	80
LO Fitness Value	21.588	7.971
LO Function Evaluations	168	133

As is obvious from Table 1 the results obtained for decreasing fuel feed (i.e. for controlling G_{B-}) are much better compared to G_{B+}. This can be naturally attributed to G_{B+} being a fourth-order system (compared to a second order system for G_{B-}) with almost three times the time-delay of G_{B-} which resulted in much faster responses.

The optimization started at initial point that was determined based on previous trial-and-error experimentation. It was found out that in order to avoid the aforementioned "overimmunization" of the controller, it was necessary to use rather low values for k_f and k_1 (for given value of initial antibody concentration). The results obtained for both optimization runs are summarized in Table 2. The most important differences in resulting values can be observed for k_f (combined antibody death rate) and k_4 (B-cell proliferation rate), implying possible stronger correlation of these parameters to the performance related to system order and time-delay.

Table 2
Initial and optimized Varela controller parameter values for increasing and decreasing fuel feed

Bounds	$10^{-7} \le k_f \le 1$	$10^{-7} \le k_I \le 1$	$10^{-7} \le k_3 \le 10$	$10^{-7} \le k_4 \le 2$
Initial values	0.001	0.0001	1.5	0.3
Optimized values (du/dt > 0)	4.682×10^{-4}	2.44×10^{-6}	0.023	0.235
Optimized values (du/dt < 0)	8.3×10^{-3}	2.644×10^{-6}	0.034	1.331

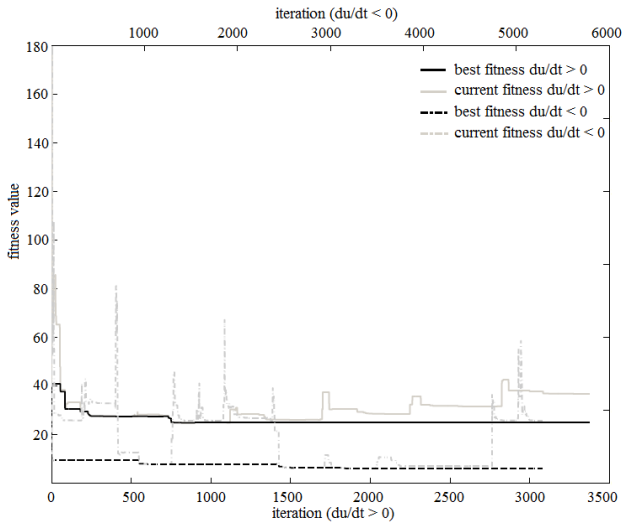


Figure 7
Current and best fitness vs. iteration graphs for both optimization runs (increasing and decreasing fuel feed)

The performance of optimized Varela controller for water temperature control (increasing fuel feed) is shown in Fig. 8. The test was again carried out for four consecutive temperature steps but this time each with different magnitude (changes of 1, 2, 3 and 4 degree Celsius). As can be seen from the figure, Varela controller offers stable (even if a bit sluggish) performance with overshoots ranging from 10.8 to 13.4% (Table 3). It is notable that the settling times for all steps are relatively long (approx. 6900 s for every response) which can be attributed to modest responses of Varela controller to large changes in control error (caused by its specific interactions between antigen, B-cells and antibodies). This might be one of the strongest reasons for its good anti-delay capabilities.

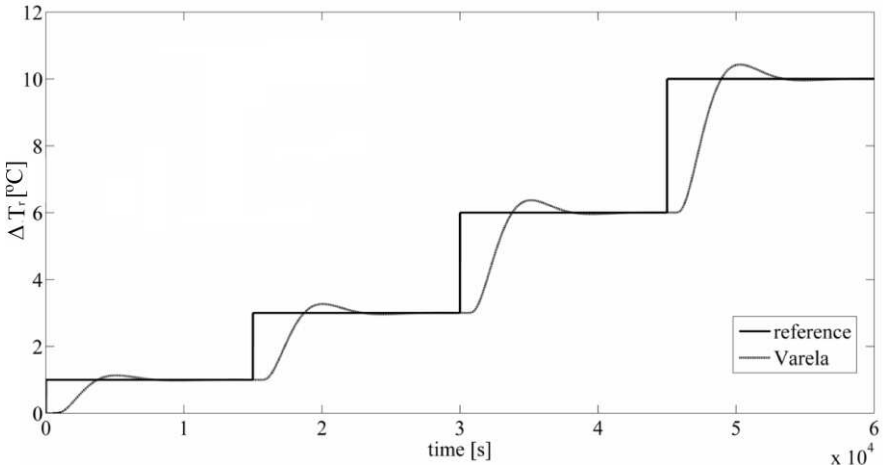


Figure 8
Control test for increasing fuel feed (four consecutive steps in desired temperature with different magnitude)

Table 3
Results of control test for increasing and decreasing fuel feed with parameters from Table 2

	du/dt>0	du/dt<0
Overshoot, M_P	M_{P1} = 13.1%, M_{P2} = 13.4% M_{P3} = 12.4%, M_{P4} = 10.8%	M_{P1} = 17%, M_{P2} = 18.2% M_{P3} = 17.2%, M_{P4} = 15.3%
Rise time(0-100%), t_r	t_{r1} = 4619, t_{r2} = 4582 t_{r3} = 4665, t_{r4} = 4815	t_{r1} = 1692, t_{r2} = 1653 t_{r3} = 1676, t_{r4} = 1724
Settling time (5%), t_s	t_{s1} = 6876, t_{s2} = 6860 t_{s3} = 6903, t_{s4} = 6955	t_{s1} = 2575, t_{s2} = 2539 t_{s3} = 2561, t_{s4} = 2604

A similar test was carried out for decreasing fuel feed the results of which are shown in Fig. 9 (and also Table 3). In this case, the responses are much faster (around 2600 s) due to G_{B-} being of second order and with shorter time delay. Overshoots for all the steps ranged from 15.3% to 18.2%, i.e. slightly greater than for previous case implying higher concentrations of B-cells (higher fuel feed rates) for given control sequence. Due to the integral action of Varela controller in response to invading antigen, in all cases the steady-state control error is zero. Summarizing the results shown in Table 3, it is readily observable that at optimized parameter values and in the range of ±10°C changes the controller offers stable performance with only small differences in performance indicator values.

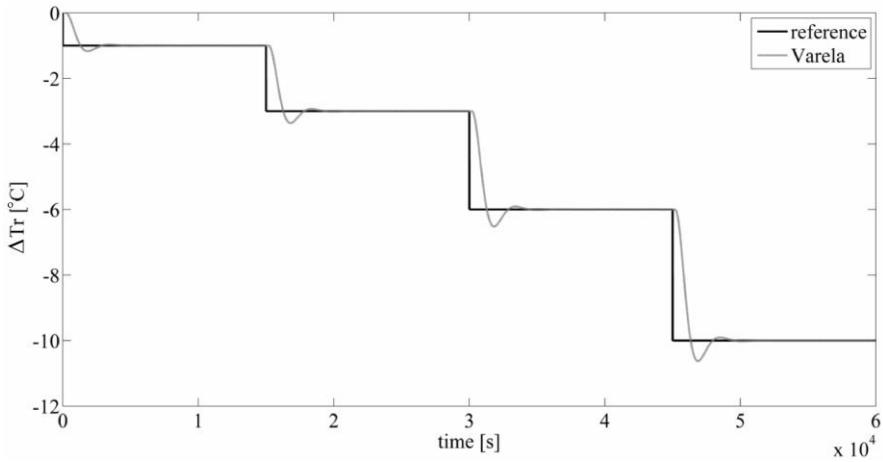


Figure 9
Control test for decreasing fuel feed (four consecutive steps in desired temperature with different magnitude)

The controller was also tested for its load rejection capabilities and robustness with results shown in Fig. 10 and Fig. 11 respectively. As was mentioned before, a disturbance in the form of heat demand fluctuations represented as the drop in water temperature was simulated with two steps of 1 °C and 2 °C at t = 5000 s. The disturbance cause output to fall to disturbance magnitude with consequent correction (rise times for both steps: t_{rv1} = 4641 s and t_{rv2} = 4652). The responses settled to a range within ±5% around desired value in t_{sv1} = 6889 s and t_{sv2} = 6892 s respectively, that is in almost the same time. This test was carried out for G_{B+} tranfer function as the disturbance was assumed to cause a drop in water temperature which was to be compensated by increase in fuel feed.

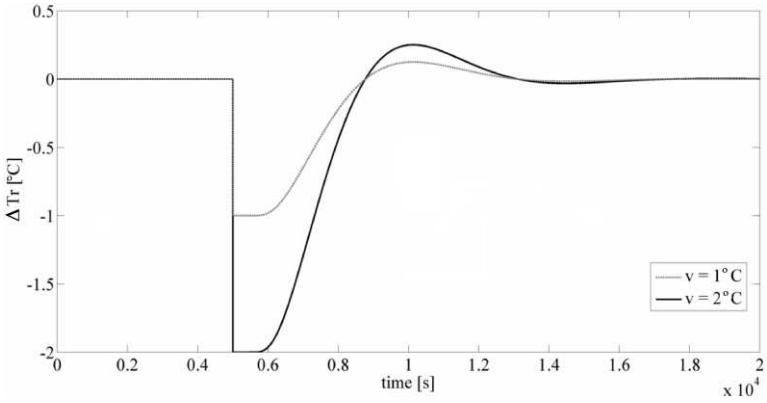


Figure 10
Load rejection test for two step disturbances representing fluctuations in heat demand (and causing the drop in water temperature)

The robustness test was carried out for ±5 °C steps in desired temperature for both G_{B+} and G_{B-}. This value was chosen as a middle value of the considered range of temperature changes (±10 °C). The nominal time constants (T_{GB+} = 200 and T_{GB-} = 165) were increased three times by 10, 20 and 30% respectively with controller set to optimized parameter values for G_{B+} and G_{B-}. As can be seen in Fig. 11, the controller is quite robust to the variations in system time constant as even 30% variations did not destabilize the system and only moderately increased overshoot (from 12.5% to 18.1% for G_{B+} and from 16.4% to 24.1%) and slowed the response.

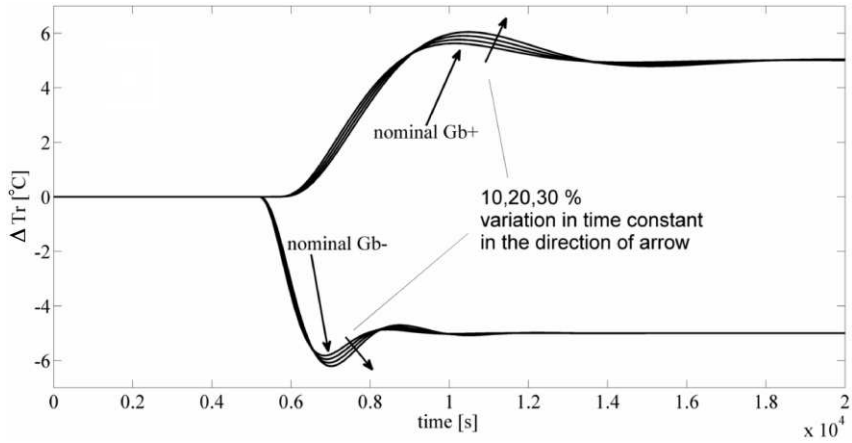


Figure 11

Robustness of Varela controller for up to 30% variation in system time constant for G_{B+} and G_{B-}

Conclusion

In this paper water temperature control for biomass-fired boiler with Varela immune controller was proposed. This process is quite difficult to control using conventional PID controller due to its excessive time-delay. Varela controller was shown to be able to stabilize the process and offered robust performance in terms of significant system parameter variations. Despite the fact that the controller was tested only in limited range of desired temperature changes, it could be seen that the optimized parameter values provided stable performance in spite of controller's nonlinearity. The effect of particular parameters on controller's performance is not as clear as in PID controllers and to achieve at least near-optimal solution, some (nongradient) optimization method might be needed. Even though there is no guarantee for obtaining a global optimum with used method, simulated annealing algorithm was capable of finding solution of reasonable quality. One of the most striking features of Varella controller performance for water temperature control that might be considered for improvement is the speed of its response. It will be of interest to find out a way in which its response could be made faster without compromising its robust and good anti-delay capabilities

whether by using additional technical or biological paradigms. Moreover, additional research is needed on the effect of other, nonoptimized parameters (initial antibody concentration and maturation and proliferation scaling parameters) on the controller's performance.

Acknowledgements

The research work is supported by the Project of the Structural Funds of the EU, Operational Programme Research and Development, Measure 2.2 Transfer of knowledge and technology from research and development into practice. Title of the project: Research and development of intelligent control systems for biomass based heat production and supply. ITMS code: 26220220030

References

[1] Van Loo, S., Koppejan, J.: Handbook of Biomass Combustion and Co-Firing, Earthscan, London, 2008, 442 p.

[2] Plaček, V., Šulc, B., Vrána, S., Hrdlička, J., Piteľ, J.: Investigation in Control of Small-scale Biomass Boilers, Proceedings of the 12th International Carpathian Control Conference (ICCC), Velké Karlovice, 25-28 May 2011, 308-311

[3] Mižáková, J., Mižák, J., Piteľ, J.: Monitoring of Operating Conditions of Biomass Combustion Process, Applied Mechanics and Materials, Vol. 308 (2013) 39-44

[4] Šulc, B., Oswald, C.: Enhanced PID Controllers in Combustion Control, Proceedings of the International Conference on Development, Energy, Environment and Economics (DEEE), Puerto De La Cruz, November 30 – December 2, 2010, 44-50

[5] Boržíková, J., Mižák, J.: Analýza možností inovácie systémov riadenia spaľovacieho procesu biomasy, Transfer inovácií, No. 16 (2010), 225-227

[6] Máša, V.: Mathematical Model of Biomass-Boiler for Control Purposes, doctoral thesis, Faculty of Mechanical Engineering, Technical university in Brno, 2010

[7] Boržíková, J., Mižák, J., Piteľ, J.: Riadenie spaľovacieho procesu biomasy s využitím techník umelej inteligencie, Strojárstvo Extra, No. 5 (2010), 45/1 – 45/4

[8] Piteľ, J., Boržíková, J., Mižák, J.: Biomass Combustion Process Control, Using Artificial Intelligence Techniques, Proceedings of the Instruments and Control Workshop (ASŘ), 30 April 2010, 317-321

[9] Abe, N., Yamanaka, K.: Smith Predictor Control and Internal Model Control, Proceedings of the SICE Annual Conference, Fukui, 4-6 August 2003, 1383-1387

[10] Mo, H.: Handbook of Research on Artificial Immune Systems and Natural Computing, Medical Information Science Reference, New York, 2009, 605 p.

[11] Fu, D.-M., Zheng, D.-L.: Designing and Analysis of a Immune Controller Based on the Improvement Varela Immune Network Model, Proceedings of the 4th International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005, 1121-1126

[12] Zhao, Y., Fu, D., Yin, Y., Wang J.: A Design Method of Immune Controller Based on Varela Artificial Immune Network Model, Proceedings of the Control and Decision Conference, Yantai, 2-4 July 2008, 3726-3731

[13] Varela, F. J., Countinho, A.: Second Generation Immune Networks, Immunology Today, Vol. 12, No. 5 (1991), 159-166

[14] Dasgupta, D., Nino, L. F.: Immunological Computation, CRC Press, Boca Raton, 2009, 277 p.

[15] Xu, L., Jiang, S.: Research of Dynamics of DC Speed Regulator of Flywheel Double Closed Loop Based on Immune Principle, Proceedings of the IEEE International Conference on Computer Science and Automation Engineering, Shanghai, 10-12 June 2011, 259-263

[16] Aarts, E., Korst, J.: Simulated Annealing and Boltzmann Machines – A Stochastic Approach to Combinatorial Optimization and Neural Computing, John Wiley & Sons, Chichester, 1989, 272 p.

[17] The MathWorks Inc., Global Optimization Toolbox User's Guide, The MathWorks Inc., Natick, 2011

Stress-Strain Interaction Model of Plasticity

Kalman Ziha

University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture
Department of Naval Architecture and Ocean Engineering
Ivana Lucica 5, 10000 Zagreb, Croatia, kziha@fsb.hr

Abstract: The article firstly investigates a discrete numerical model of finite interaction between successive microstructural bond failures and remaining intact internal bonds in materials. Secondly, it reveals the general linear finite continuous cause and effect interaction concept. The interaction model is examined numerically, experimentally and analytically on an illustrative case of a parallel system of bonds. The general concept is applied to the macroscopic stress-strain interaction model of material plasticity. Examples of metallic materials are elaborated on reported theoretical and experimental strain data.

Keywords: yielding; necking; plasticity; tensile test; metals; interaction

1 Introduction

This research is motivated by the stance that the curve fitting methods based on experimental data about plasticity in engineering of materials often do not have appropriate physical foundation and in some cases are not accurate enough for practical applications. The methods of thermodynamics, continuum mechanics and dislocation physics in materials sciences provide theoretical solutions for complicated problems in engineering plasticity. The article advocates that a comprehensive, more accurate and straightforward definition of non-linear material mechanical properties of plasticity might be of interest in practice, particularly for determination of the ultimate strength of engineering structures. The plasticity model in this article focuses on internal failures of microstructural bonds between discrete material particles rather than on crystallographic defects in shape regularities resulting in dislocation of particles. The article investigates the suitability of an empirical Cause-and-Effect Interaction concept (CEI) for definition of a Stress-and-Strain Interaction (SSI) macrostructural model of plasticity definable by propensity to and intensity of interaction. The applications of the numerical procedure of the analytic model of the SSI concept are illustrated by reported examples of experimental results of plasticity testings.

2 The Linear CEI Concept of Plasticity

The formulations of microstructural processes and applications of thermodynamics in material sciences, e.g. [1], and macroscopic continuum mechanics are in wide use for modelling of plasticity in engineering problems, e.g. [2] [3] [4] [5] [6] [7] [8]. Multilevel theories of irrecoverable deformations, where macro-strains are related to the processes occurring on the micro level of material, provide relatively simple stress-strain and strain-time formulae [9] [10] [11]. The rearrangements of the internal structures within which the particles are being collectively dislocated to new positions of internal equilibrium are frequently explicated in discrete dislocation physics as interactions, e.g. [12] [13] [14] [15] [16]. Simulation methods based on dislocation physics and using finite element analysis, e.g. [17] [18] [19], are important but time-consuming numerical tools. The linear CEI concept [20] [21] [22] in the article holds the internal failures of bonds among discrete particles in materials accountable for the defects on the microstructural level. The starting assumption in this model is that the macrostructural mechanical properties of materials under loading depend on great but finite total number C_R of intact internal elastic bonds intrinsic to the basic material physical microstructure (Fig. 1). The primary effect $E(C)$ induced by successive bond failures C is gradual reduction of strength until yielding of overloaded elastic bonds. The primary effect E (weakening, yielding, plasticity) under loading is linearly related to the cause C (elastic bond failures) as shown:

$$E(C) = p \cdot C \tag{1}$$

Simultaneously the remaining number of intact elastic bonds (C_R-C) preserves the residual load-carrying capacity (strength), which is the left-over resistance to deformation after C successive bond failures. From the initial assumption of linearity between the primal cause and effect (1), it follows that the durability $R(C)$ also has to be linearly related to the remaining number of intact bonds:

$$R(C) = r \cdot (C_R - C) \tag{2}$$

The hypothesis of the study is that the weakening $E(C)$ is not just a simple cause-and-effect relation C⇒E or E⇐C as in (1 and 2) with respect to the cause C but rather a more complex cause-and-effect interaction C⇔E. The weakening $E(C)$ with respect to the durability $R(C)$ is the consequence of the redistribution of internal loads between the numbers of failed C and intact (C_R-C) bonds (Table 1). The interaction rate expresses how the weakening $E(C)$ (1) reduces the remaining durability $R(C)$ (2). That in turn interactively accelerates the primary weakening $E(C)$ by the amount of $E(I)$ induced by a secondary cause $I(C)$ due to interaction with the cause C (e.g. Fig. 2). Hence, the interaction rate is simply in proportion i to the ratio of numbers of failed C and intact (C_R-C) bonds as shown next:

$$\Delta E\left[I(C)\right] = \frac{E(C)}{R(C)} = i \cdot \frac{C}{C_R - C} \tag{3}$$

The secondary weakening $E(I)$ (3) by each successive failure of elastic bonds results in redistribution of load to remaining intact bonds, that is, the weakening of the elastic system of bonds until fracture. However, the overall secondary weakening $E(I)$ by each successive failure of elastic bonds accumulates all the former effects induced by the interaction between the weakening $E(C)$ and the durability of material $R(C)$ that is expressed by the following summation:

$$E\left[I(C)\right]=\sum_{C=1}^{C_R}\Delta E\left[I(C)\right] \tag{4}$$

The overall plasticity may be viewed as the consequence of the overall weakening $E(C,I)=E(C)+E(I)$ resulting from the primary $E(C)$ (1) and secondary $E(I)$ (4) weakening (e.g. Table 1, Fig. 2).

The two parameters p and $i=p/r$ (1-4) represent the propensity to and the intensity of interaction between the weakening and the durability of material. The work done in weakening $E(I)$ (4) (Fig. 1) is equivalent to the accumulated energy of interaction $U[E(I)]$ attainable by integration of all successive secondary effects of failures of elastic bond commonly available from experiments, as shown below:

$$U\left[I(C)\right]=\sum_{C=1}^{C_R}E\left[I(C)\right] \tag{5}$$

The exposed CEI concept (1-5) is not in contradiction with the rules in mechanics.

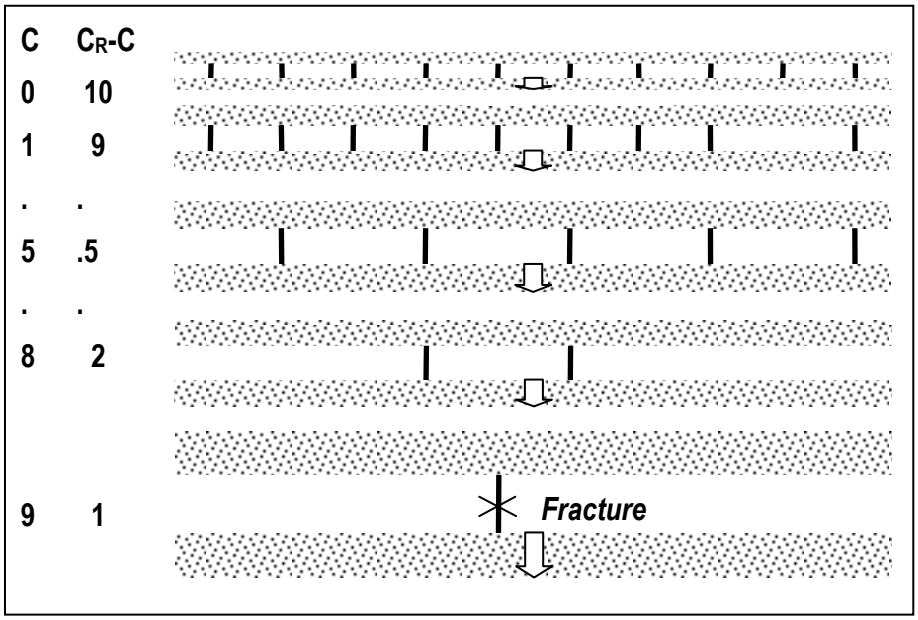


Figure 1

Internal elastic bond failures till fracture in parallel arrangement of bonds (C_R=10)

Table 1
CEI calculation for C_R=10 initial intact internal bonds in material

C	E(C)	C_R-C	Redistribution	C/(C_R-C)	E(I)	E(C,I)	U(I)
0	0	10	10/10=1+0/10	0,00	0,00	0.00	0.00
1	1	9	10/9=1+1/9	0,11	0,11	1.11	0.11
2	2	8	10/8=1+2/8	0,25	0,36	2.36	0.47
3	3	7	10/7=1+3/7	0,43	0,79	3.79	1.26
4	4	6	10/6=1+4/6	0,67	1,46	5.46	2.72
5	5	5	10/5=1+5/5	1,00	2,46	7.46	5.18
6	6	4	10/4=1+6/4	1,50	3,96	9.96	9.14
7	7	3	10/3=1+7/3	2,33	16,2	13.29	15.4
8	8	2	10/2=1+8/2	4,00	10,3	18.29	25.7
9	9	1	10/1=1+9/1	9,00	19,3	28.29	45.0
10	10	0	10/0=1+10/0	∞	∞	∞	∞

3 Mathematical Formulation of the CEI Concept

The direct application of infinitesimal calculus to the massive discrete systems of a great but finite number of micro-structural bonds decomposed into linear (1) and nonlinear (4) parts $E(C,I)=E(C)+E(I)$ provides the following analytical formulation of the general CEI concept (1-5) [20] [21] [22] (Fig. 2) of continuous finite systems on the macroscopic level as shown:

$$\frac{\mathrm{d}^2E(C,I)}{\mathrm{d}^2C} = i \cdot \frac{1}{(1-c)^2} \tag{6}$$

$$\frac{\mathrm{d}E(C,I)}{\mathrm{d}C} = p + i \cdot \frac{c}{1-c} \tag{7}$$

$$E(C) = p \cdot \int_0^c \mathrm{d}C = C_R \cdot p \cdot c \tag{8}$$

$$E(I) = i \cdot \int_0^c \frac{C}{C_R - C}\mathrm{d}C = i \cdot C_R \cdot \left[-c - \ln(1-c)\right] \tag{9}$$

$$U(I) = \int_0^c E(I)\mathrm{d}C = i \cdot C_R^2 \cdot \left\{-c^2/2 + \left[c + (1-c)\cdot\ln(1-c)\right]\right\} \tag{10}$$

The relation $C(E)$ can be obtained from (9) by integration of the inverse derivative of (7), that is, the rate of change of the cause C with respect to the effect E as:

$$\frac{\mathrm{d}C}{\mathrm{d}E(C,I)} = 1/\left[\mathrm{d}E(C,I)/\mathrm{d}C\right] = \frac{1-c}{p+c(i-p)} \tag{11}$$

The interaction intensity parameter is attainable from the equivalence of the observable work $W(I)=U(I)$ done on interactions and the interaction energy (10):

$$i = W(I)/C_R^2 \cdot \left\{-c^2/2 + \left[c + (1-c)\cdot\ln(1-c)\right]\right\} \tag{12}$$

In the mathematical formulation of the CEI concept (6-11) $c=C/C_R$ and $e=E/C_R$ are the dimensionless cause C and effect E relative to final value C_R.

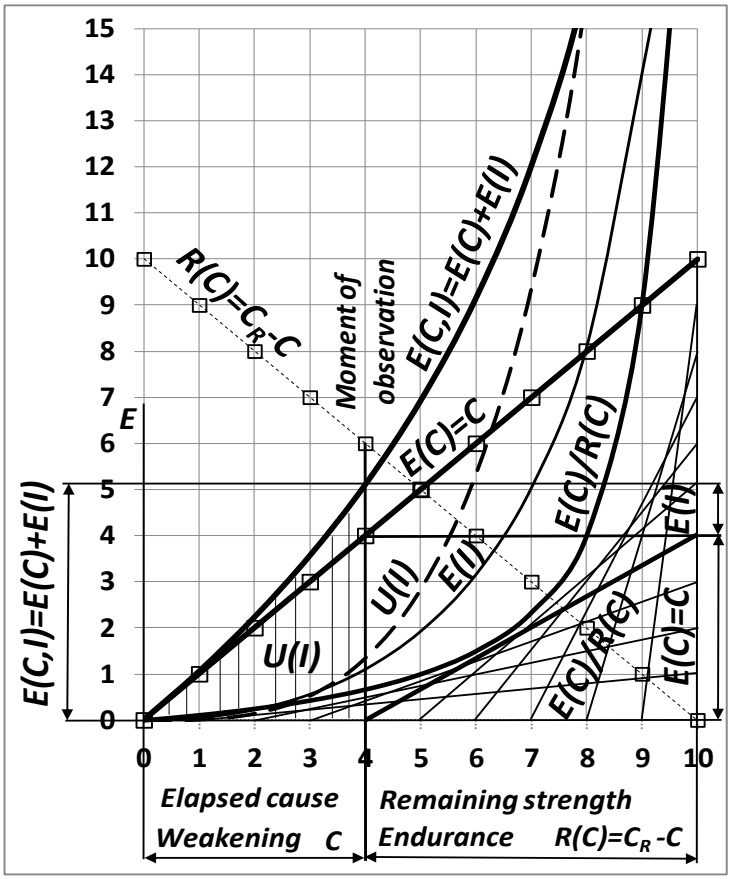


Figure 2
Numeric and analytical example of the CEI concept for C_R=10 bonds (Table 1)

In terms of the exposed CEI concept, the strain hardening in polycrystalline materials under excessive mechanical loadings may be viewed as the consequence of overloading of remaining intact bonds after some bonds have failed. The resulting load redistribution increases the internal stresses in microstructural grain boundaries that intensify the massive propagation of dislocations in material followed by observable macroscopic permanent deformations.

4 Experimental Investigation of the CEI Concept

The following experiments physically reproduce the CEI concept for C_R=10 bonds simulated by ten elastic rubber bands bonds in parallel arrangement under constant load (Fig. 1). Rubber band bonds of lengths ÿ=4 cm, ÿ=6 cm and ÿ=12 cm are investigated in five independent tests each (Fig. 3).

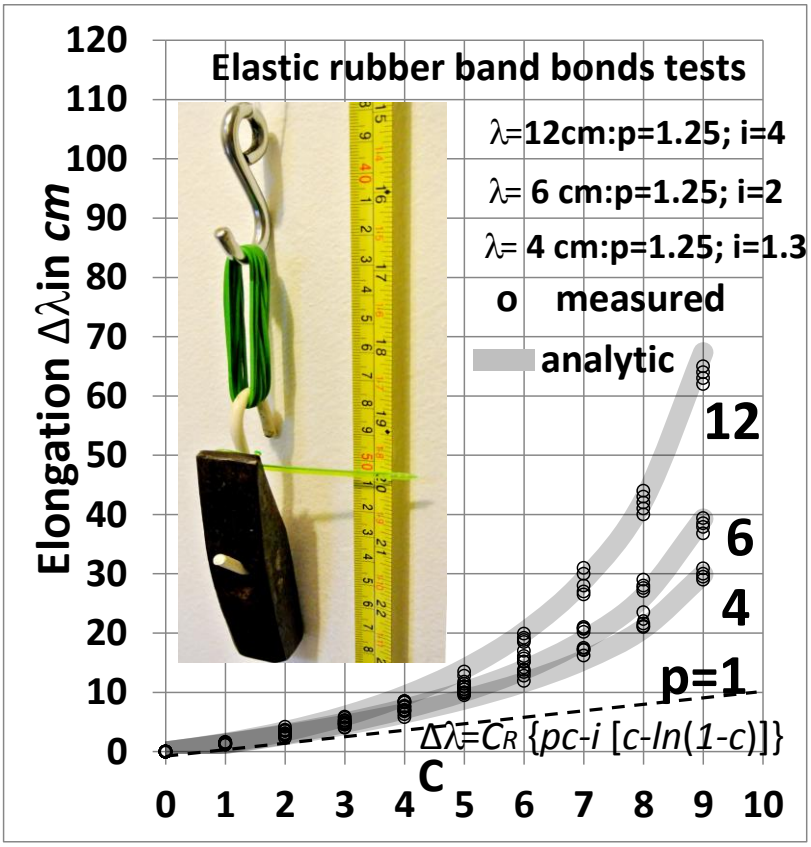


Figure 3
Tension experiments on C_R=10 elastic rubber band bonds and CEI analytical results

Each experiment consists of measurements of elongations Δÿ after random one-by-one removal of rubber bands. The elongations are used to find the propensity *p* and intensity *i* parameters of the CEI model (6-11). The propensity to interaction *p* is obtained by measurement of the elongation for a single band under the same load. The interaction intensity *i* is determined from the work done in stretching/extending of the remaining rubber bands obtained by numerical integration of the elongation curves by using the trapezium rule. The experiments confirm the relation between the CEI model and the measured values (Fig. 3).

5 The Appliance of the CEI Concept on the SSI Model

The power rule was proposed earlier for fitting of non-linear $\sigma-\varepsilon$ stress-strain curves: $\sigma=\sigma_o+K\cdot\varepsilon_p^n$ [23], $\varepsilon=\sigma/E+K\cdot(\sigma/E)^n$ [24], $\sigma=K\cdot\varepsilon_p^n$ [25] $\sigma=K\cdot(\varepsilon_o+\varepsilon)^n$ [26] [27]. The exponential rule was suggested as well $\sigma=\sigma_o+\sigma_{sat}\cdot(1-e^{-m\cdot\varepsilon})$ [28]. The study considers the Stress-Strain Interaction (SSI) model of plasticity $\varepsilon\Leftrightarrow\sigma$ apparent on the macrostructural level as the manifestation of the interactions between a massive number of failed and intact bonds on the microstructural level patterned after the CEI model (1-5).

The application of the general CEI concept (6-11) to the total plastic strain $\varepsilon_p(\sigma,\sigma_I)=\varepsilon(\sigma)+\varepsilon(\sigma_I)$ composed of the primary linear plastic strain $\varepsilon_p(\sigma)$ induced by stress σ and of the non-linear accumulation of secondary plastic strains $\varepsilon_p(\sigma_I)$ resulting from the interactions of secondary stresses σ_I and plastic strains $\varepsilon_p(\sigma)$ provides the analytical terms for continuous material plasticity as follows:

$$\varepsilon_p{}'' = \frac{\mathrm{d}^2\varepsilon_p(\sigma,\sigma_I)}{\mathrm{d}\sigma^2} = \sigma_R\cdot i\cdot\left(\frac{1}{1-s}\right)^2 \tag{13}$$

$$\varepsilon_p{}' = \frac{\mathrm{d}\varepsilon_p(\sigma,\sigma_I)}{\mathrm{d}\sigma} = \sigma_R\cdot\left(p+i\cdot\frac{s}{1-s}\right) \tag{14}$$

$$\varepsilon_p=\varepsilon_p(\sigma,\sigma_I)=\varepsilon_p(\sigma)+\varepsilon_p(\sigma_I)=\sigma_R\cdot\left\{p\cdot s-i\cdot\left[s+\ln(1-s)\right]\right\} \tag{15}$$

$$U(E,I)=U(E)+U(I)=\sigma_R^2\cdot\left\{p\cdot s^2/2+i\cdot\left\{-s^2/2+\left[s+(1-s)\cdot\ln(1-s)\right]\right\}\right\} \tag{16}$$

The interaction intensity can be obtained from the work $W(I)$ done in experimentally defined plastic deformations and the theoretical energy in (16) as it is shown below:

$$i = W(I) / \left\{ \sigma_R^2 \cdot \left\{ -s^2 / 2 + \left[s + (1-s) \cdot \ln(1-s) \right] \right\} \right\} \tag{17}$$

where $s=\sigma/\sigma_R$ in (13-17) is the stress σ relative to its reference value σ_R.

The parameters $p=1/P$ and $i=1/I$ in (13-17) represent the propensity to and the intensity of plasticity respectively, and can be derived from experimental data. Parameters P and I represent the propensity and intensity module of linear and non–linear plasticity induced by interaction between stresses and strains. The parameters can be obtained directly (17) or numerically using least squares or general nonlinear optimization methods.

The SSI assumption for necking is that the rate of the decrease of stresses induced by changes of the sectional geometry due to interaction between the progressing strains ε and the residual strain capacity ε_R-ε can be defined analogously to (14):

$$\sigma_n' = \frac{\mathrm{d}\sigma_n(\varepsilon_p, \varepsilon_{pI})}{\mathrm{d}\varepsilon} = M + N \cdot \frac{e}{1-e} \tag{18}$$

The application of the CEI concept $\sigma \Leftrightarrow \varepsilon$ to the decrease in stresses $\sigma_n(\varepsilon_p, \varepsilon_{pI}) = \sigma_n(\varepsilon_p) + \sigma_n(\varepsilon_{pI})$ due to necking consisting of a primary linear decrease $\sigma_n(\varepsilon_p)$ induced by strain ε_p and of non-linear accumulation of a secondary decrease $\sigma_n(\varepsilon_{pI})$ resulting from the changes in strains ε_{pI} due to interactions with $\sigma_n(\varepsilon_p)$ provides the expression for necking as follows:

$$\sigma_n = \sigma_n(\varepsilon_p, \varepsilon_{pI}) = \sigma_n(\varepsilon_p) + \sigma_n(\varepsilon_{pI}) = \varepsilon_R \cdot \left\{ M \cdot e - N \cdot \left[e + \ln(1-e) \right] \right\} \tag{19}$$

where $e=\varepsilon/\varepsilon_R$ in (18, 19) is the strain ε relative to its asymptotic reference value ε_R.

The parameters $m=1/M$ and $n = 1/N$ represent the propensity to and the intensity of necking and can be derived from experimental data. Parameters M and N represent the propensity and intensity module of necking, respectively.

6 Examples

The first example demonstrates the appropriateness of the SSI model with respect to tension test results of mild shipbuilding steel (Fig. 4). The propensity to yielding $p=1/P=1/4000=0.00025\ MPa^{-1}$ is obtained by numerical derivation at the beginning of the yielding. The work done in plasticization is obtained by

numerical integration of the experimental σ-ε curve using the trapezium rule and amounts to U=24.2 *MPa*. The plasticization intensity (17) is obtained from the interaction energy (16) as $i=1/I=1/4145=0.000241$ MPa^{-1}. The SSI expression for plasticity (15) (Fig. 4) in this example is:

$$\varepsilon_p(\sigma,\sigma_I) = 150 \cdot \left\{0,00025 \cdot s - 0,000267 \cdot \left[s + \ln(1-s)\right]\right\} \tag{19}$$

Ramberg-Osgood power law parameters obtained by the least squares method, K=11300 and n=2,36, do not match the stress-strain curves over the whole range of the σ-ε curve (Fig. 4). The propensity modulus to necking is M=0 *MPa*. The necking intensity modulus is N=760. The SSI expression for necking in this example (Fig. 4) is:

$$\sigma_n(\varepsilon_p,\varepsilon_{pI}) = 474 - 0,15 \cdot 760 \cdot \left[e + \ln(1-e)\right] \tag{20}$$

The example of mild shipbuilding steel tested in the Laboratory of Experimental Mechanics of the Faculty of Mechanical Engineering and Naval Architecture shows that the stress-strain curves obtained by the SSI model based on the CEI concept fit the whole range of tests results (Fig. 4).

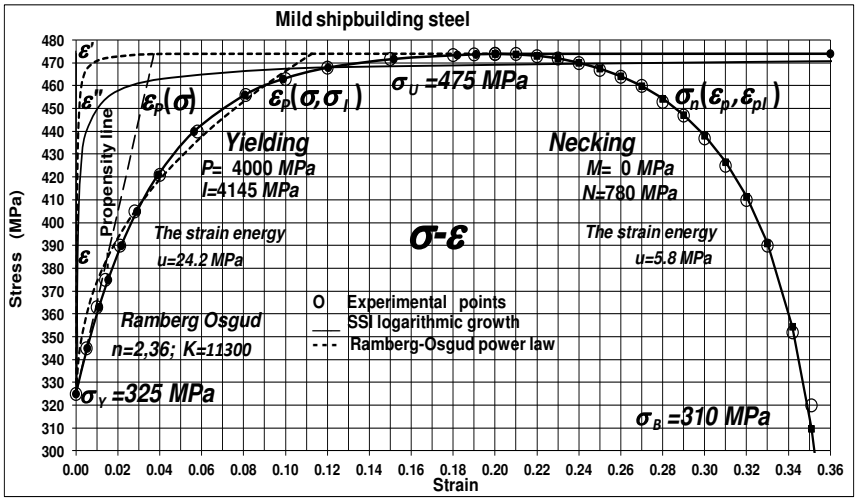


Figure 4
The SSI model of mild shipbuilding steel

The second example applies the SSI model on three types of unclassified cast irons (spheroidal, compacted and flake). The differences between the irons are in the influence of graphite morphology on stress-strain curves that harden with plastic deformation. The example confirms the smooth elastic-plastic transition typical for brittle materials [29] (Fig. 5).

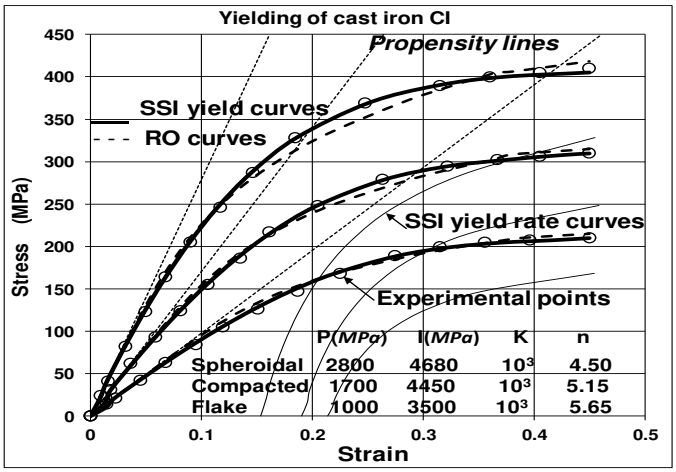


Figure 5
The SSI model of three types of cast irons

The third example compares the SSI model results with the Crystal Plasticity Finite Element (CPFE) [30] numerical study of the polyslip behaviour of single aluminium crystals of different initial crystallographic orientations (111, 112, 123, 100) under tensile loading and with experiments [31] and [32] (Fig. 6).

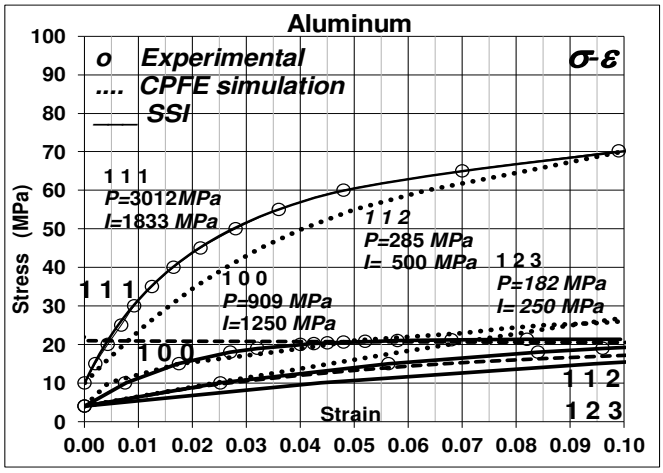


Figure 6
CPFE simulated and experimentally observed orientation-dependence of the stress-strain of single aluminium crystals during tensile loading

The fourth example compares the SSI model results with experimental results [33] and with the CPFE simulation [30] using the assumption of statistically stored dislocations (SSDs) and geometrically necessary dislocation (GND) density addition [34] for different grain diameters (14, 33 and 220 μm) (Fig. 7).

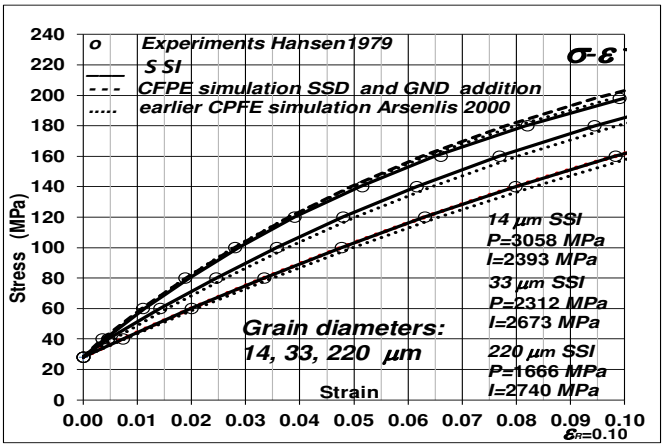


Figure 7
Stress-strain curves for average grain diameters of 14, 33 and 220 μm

The fifth example compares the SSI model results with experimental results obtained by laser extensometer type W-80 from Fiedler Optoelektronik of stress and strain measurements on aluminium specimens in time [35] (Fig. 8).

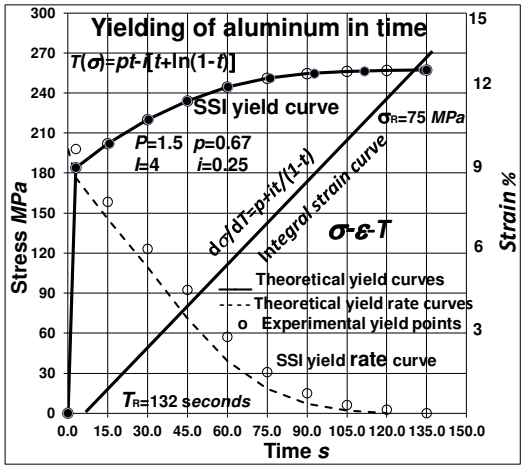


Figure 8
Stress and strain measurements on aluminium specimens in time

Conclusions

The application of the cause-and-effect interaction concept to the time independent stress-and-strain interaction model demonstrates in this paper how the material yielding and plasticity could be viewed as asymptotical growth processes analytically definable as logarithmic function over the whole range of plasticity rather than unlimited power growths in some segments of the stress-strain curves.

The rapid asymptotic growth of the sensitivity to failures may explain sudden and uncertain breaks in continuity of a material's behaviour under increasing loadings due to structural and environmental imperfections and defects in the material. The experience of this study indicates that some time-variant material mechanical properties such as creeping could be investigated in the future as asymptotically propagating processes following the cause-and-effect interaction concept.

The presented model is governed by two unique properties of a material, which are the propensity to and the intensity of interactions, both evident from experiments. The initial propensity represents the starting microstructural constellation of internal bonds between the constituent particles of a material and their consistency normally reflects the initial state of the strength of the material. The interaction intensity parameter stands for the average of massive progressions of internal bond failures relatable to the overall material durability on the macroscopic level. The two interaction parameters are straightforwardly available from standard tensile testing of material mechanical properties. For tensile tests performed in time the interaction parameters can be calibrated in the time scale.

The interaction model of material yielding elaborated in the article is not another curve-fitting method based on experimental data points but rather an implementation of a more general physical concept to investigate the mechanical properties of materials. This physical concept uses the equivalence of theoretical energy of micro-structural interactions between the failed and intact bonds to the experimentally observable stress-strain energy on macro-structural level. The general cause-and-effect interaction concept describes a part of trans-temporal continuum that relates the past and a future time separated by limitation of human's ability of perceptions beyond the instant of observation. The results in this study reveal how the empirical cause-and-effect interaction concept could be a rational approach to an alternative understanding of the non-linear material stress-and-strain interaction model, sufficiently simple and accurate for practical engineering problems of non-linear strains, strain hardening, yielding and plasticity.

Acknowledgement

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under grant No. 120-1201703-1702.

References

[1] Lupis C. H. P., Chemical Thermodynamics of Materials. Elsevier Science Publ. Co., North-Holland, 1983

[2] Van Vlack L. H., Elements of Materials Science and Engineering. Addison-Wesley, 1985

[3] Khan S. A., Huang S., Continuum Theory of Plasticity. Wiley-Interscience, 1995

[4] Hibbeler R. C., Mechanics of Materials, 5th Edition, Prentice Hall, 2002

[5] Davis J. R., Tensile testing. ASM International, 2004

[6] Hosford, W. F. J., Fleischer R. L., Backofen W. A., Tensile Deformation of Aluminum Single Crystals at Low Temperatures. Acta Metal. 1960, 8, 187-199

[7] Rees D., Basic Engineering Plasticity. Butterworth-Heinemann, 2006

[8] Wolff M., Boehm M., Helm D., Material Behavior of Steel - Modeling of Complex Phenomena and Thermodynamic Consistency. Int. J. of Plasticity, 2008, 24, 746-774

[9] Rusinko, A. and Rusinko, K., Synthetic Theory of Irreversible Deformation in the Context of Fundamental Bases of Plasticity. Mechanics of Materials, 2009, 41, 106-120

[10] Rusinko, A. and Rusinko, K., Plasticity and Creep of Metals. Springer Berlin Heidelberg, 2011

[11] Rusinko, A., Non-Classical Problems of Irreversible Deformation in Terms of the Synthetic Theory. Acta Polytechnica Hungarica, 2010, 7 (3), 25-62

[12] Sozinov L.V., Gavriljuk G., Estimation of Interaction Energies Me-(C, N) in f.c.c. Iron-based Alloys using Thermo-Calc Thermodynamic Database. Scripta Materialia, 1999, 41 (6) 679-683

[13] Orsini V. C., Zikry M. A., Void Growth and Interaction in Crystalline Materials. Int J Plasticity, 2001, 17 (10) 1393-1417

[14] Bieler T. R., Eisenlohr P., Roters F., Kumar D., Mason D. E., Crimp M. A., Raabe, D., The Role of Heterogeneous Deformation on Damage Nucleation at Grain Boundaries in Single Phase Metals. Int. J. of Plasticity, 2009, 25 (9) 1655-1638

[15] Mareau C., Favier V., Weber B., Galtier A., Berveiller M., Micromechanical Modeling of the Interactions between the Microstructure and the Dissipative Deformation Mechanisms in Steels under Cyclic Loading. Int J Plasticity, 2012, 32-33, 106-120

[16] Khraishi T. A., Yan L. C., Shen Y. L., Dynamic Simulations of the Interaction between Dislocations and Dilute Particle Concentrations in Metal-Matrix Composites (MMCs) Int. J. of Plasticity. 2004, 20 (6) 1039-1057

[17] Raabe D., Roters F., Using Texture Components in Crystal Plasticity Finite Element Simulations. 2004, 20 (3) 339-361

[18] Queyreau, S., Monnet G., Devincre B., Slip Systems Interactions in Alpha-Iron Determined by Dislocation Dynamics Simulations. Int. J. of Plasticity, 2007, 25 (2) 361-377

[19] Kanjarla A. K., Van Houtte P., Delannay L., Assessment of Plastic Heterogeneity in Grain Interaction Models using Crystal Plasticity Finite Element Method. Int. J. of Plasticity, 2009, 26 (8) 1220-1233

[20] Ziha K., Fatigue Yield. Int. J of Fatigue 2009, 31, 1211-1214

[21] Ziha K., Modeling of Worsening. Journal of systemics, cybernetics and informatics. 2012, 10 (4) 11-16

[22] Ziha, K., Cause-and-Effect Interactions in Natural Sciences. La Pensee, 2014, 76 (3) Part 3

[23] Ludwik P., *Elemente der Technologischen Mechanik*. Springer, Berlin, 1909, 32

[24] Ramberg W., Osgood W. R., Description of Stress-Strain Curves by Three Parameters. Technical Report 902, National Advisory Committee for Aeronautics, Washington DC, 1943

[25] Hollomon J. H., Tensile Deformation. Trans. AIME 1945, 162, 268-290

[26] Swift H. W., Plastic Instability under Oplane Stress. J. Mech. Phys. Solids 1952, 1, 1

[27] Fernandes J. V., Rodrigues D. M.,Menezes L. F., Vieira M. F., A Modified Swift Law for Prestrained Materials, Int. J. of Plasticity, 1998, 14(6) 537-550

[28] Voce E., The Relationship between Stress and Strain for Homogeneous Deformation. J. Inst. Met. 1948, 74, 537-562

[29] Tamarin Y., Atlas of Stress-Strain Curves. ASM International, 2002

[30] Arsenlis A., Parks D. M., Modeling the Evolution of Crystallographic Dislocation Density in Crystal Plasticity. J. Mech. Phys. Solids, 2002, 50 (9) 1979-2009

[31] Kocks U. F., Polyslip in Single Crystals of Face-centered Cubic Metals. PhD Thesis, Harvard University, Cambridge, MA, 1959

[32] Hosford W. F., Mechanical Behavior of Materials. Cambridge University Press, 2005

[33] Hansen N., The Effect of Grain Size and Strain on the Tensile How Stress of Copper at Room Temperature. In: Haasen, P, Gerold V, Kostorz G (Eds.), Proceedings of the 5th International Conference on the trength of Metals and Alloys. 2. Pergamon Press, Oxford, 1979, 849-854

[34] Evers L. P., Parks D. M., Brekelmans W. A. M., Geers M. G. D., Crystal Plasticity Model with Enhanced Hardening by Geometrically Necessary Dislocation Accumulation. J. Mech. Phys. Solids, 2002, 50 (11), 2403-2424

[35] http://www.fiedler-oe.de/en/applications/materials/alu/ Fiedler Optoelektronik GmbH, Lűtzen, 2007

Thread Forming Tools with Optimised Coatings

Péter Tállai, Sándor Csuka, Sándor Sipos

Óbuda University, Donát Bánki Faculty of Mechanical Engineering and Security Technology, Institute of Material Science and Technology

e-mail: tallai.peter@bgk.uni-obuda.hu, csuka.sandor@bgk.uni-obuda.hu, sipos.sandor@bgk.uni-obuda.hu

Abstract: Different tool geometries have been developed by leading companies in the tool industry to different aims of application. The fastest progress can be observed in the development of coatings, in spite of this it can be noticed that there are only few coating types, but they can be used in a wide range. For users it offers the great advantage to select appropriate tool easier. Coating types, developed specially to individual aim of application, offer much better solution, compared to the generally used coating types. Platit AG, dealing with the development of optimised coating types, has requested our specialised group to carry out tests on specialised coating types. Our present study is going to summarise the results of the tests, carried out with thread forming tools.

Keywords: thread forming; minimal quantity of lubrication; coating types; CrTiN; AlCrN; AlTiN

1 Characteristics of Thread Forming Operations

Thread forming operation is a technological alternative to the thread drilling. Due to the developed coating types and tool materials, high-performance tools have appeared on the market, being able to achieve greater productivity under certain conditions: although higher forming speed values can be applied, the tool life has a multiple of the value, gained by thread drills [1, 2].

Compared to the thread drilling, the thread forming operation has the below listed advantages:

- there is no chip development, causing several problems in case of thread drilling during the chip flow from the hole [3],
- the same geometry can be used to blind and through holes,
- due to the cold formation there is an increase in the strength of the thread (this fact has an increased importance especially in case of threads, having small diameter, see **Figure 1**)[4],

- compared to the thread drilling operation, it is possible to carry out the thread forming operation with double forming speed values and at the same time with increased tool life,
- the same clamping device can be used like in case of thread drilling (there are no special demands).

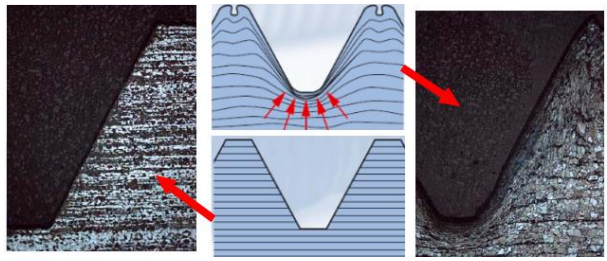


Figure 1
The difference between threads, machined with cutting and forming operation (on the left side of the picture: thread, machined with cutting operation)

As disadvantage we must mention that tool producing companies do not suggest to carry out forming operations of plastic materials or materials, having high fracture strength (>1400 N/mm^2) and low breaking extension (<5%). In case of steels, titanium and nickel alloys, having a fracture strength value higher than 1200 N/mm^2, it is recommended to apply emulsion lubrication of 10% instead of MQL (minimal quantity of lubrication). In case of thread forming it is necessary to make a core hole; however, the size of the necessary hole is different: for example, in case of a thread with size M6x1, the diameter of the core hole is 5 mm in case of thread drilling, while in case of thread forming it is Ø5.55 mm. Thread forming operation is more sensible to the diameter of the core hole and to the shape adherence (e.g. roundness error). If the core hole is smaller than the permissible size than there will be an increase in the torque demand and a decrease in the tool life. Knowing this fact we have to emphasise: a greater attention should be paid to the preparatory works in case of thread forming operation.

2 Aims and Conditions of the Examination Process

The examinations have been carried out on thread formers, having a size of M6x1 6HX (HSSE) – the tools have been deposited with 5 different coating types. The types and the most important data of the coating types, deposited on the tools, are the following:

- #1 original (deposited by the producing company) TiN-coating, with a thickness of 2-4 µm and hardness of 2300 HVM,
- #2 **CrTiN** (monolayer) coating with thickness of 2.5 µm (hardness: 38 GPA),

- #3 **$Al_{0,6}$ $Ti_{0,4}$N-coating** with thickness of 2.5 μm (hardness: 42 GPa),
- #4 **$Al_{0,5}Cr_{0,5}$N**-coating with thickness of 2.5 μm (hardness: 42 GPa),
- #5 **TiN** coating with thickness of 2.5 μm (hardness: 38 GPa).

The coating types, marked with #2 … #5, have been deposited by the subsidiary of Platit AG (Pivot s.r.o., Sumperk, Czech Republic) on bright, high-speed steel tools. Having studied the tool catalogue of the company, selling thread forming tools, this tool structure is recommended to constructional and carbon steels, alloyed and unalloyed aluminium and to unalloyed yellow brass. The allowable range of forming speed values is 12-25 m/min. The hole, to be formed, may be either through (> 1.5xd) or blind hole (> 2.5xd).

With the execution of the examinations we wished to find answer to the following questions:

- Where and what kind of wear types do develop on the forming tools? How is the deterioration of the tool affected by the coating type?

- How is the behaviour of the forming tools affected by different forming speed values?

- Which from the coating types will be the optimal solution under the test conditions, applied by us?

- What kind of behaviour can be observed in case of thread forming operation of materials, having high fracture strength and in case of MQL?

- In what way can the deterioration process of the forming tools be followed best?

- What kind of changes can be observed in the torque development and in the machined surface if flood type of lubrication is applied instead of minimal quantity of lubrication?

In order to evaluate the results in an objective way we have measured the torque demand of the tools by a rotating multicomponent dynamometer, produced by Kistler at every fifth thread and after every 5^{th}-10^{th} thread we have checked the wear process of the tools by a microscope. From the point of view of users, the behaviour of tools is not the only fact, being relevant, the quality of the machined threads has an important role as well; therefore the threads have been regularly checked by us with a gauge.

The workpieces, prepared to the examinations, have been made from a pre-tempered material grade 40CrMnMo7 (W. Nr. 1.2311), having a fracture strength of 945 N/mm^2 and breaking extension of 7.75%. In the industry this tool steel is a widely used base material in case of injection moulding and pressure casting die.

The thread forming operations have been carried out on through holes of 1.5xd. In case of a deeper hole it is recommended to use internal MQL, so the external lubrication, applied by us, will not be optimal anymore [5]. The MQL device

vaporizes vegetable based oil, type: TRIM TAP NC, with the help of compressed air, and it can be directly led to the forming tool through the nozzles. The desired oil quantity can be regulated with the valves of the device. According to the most literatures, the applied quantity is between 5...50 ml/h. (In case of application of the traditional, so-called flood type of lubrication the liquid flow may achieve or even exceed a value of 50 l/min.) On the applied machine (type: HCS250, produced by TKM (Germany)) the oil consumption can be measured with the help of an external electronic device, developed by us [3].

Every coating has been tested at three different forming speed values (14, 20 and 25 m/min) so that the circumstances of the MQL should remain as it is. During the trials, the internal (operating) pressure of the MQL device had a value of 3 bar, while the external pressure, causing the pulverizing, was 0.6 bar. The trials have been carried out with these values, the delivered oil quantity was 21±0.5 ml/hour, according to our measurements.

Table 1
Test's parameters

Forming speeds, m/min	14	20	25
Hole depth (trough hole), mm	1.5xd		
Pilot hole diameter, mm	Ø5.55		
Workpiece	40CrMnMo7 (W. Nr. 1.2311)		
MQL unit parameters (*TKM HCS250*)	internal pressure, bar	3	
	external pressure, bar	0.6	
Measuring system	KISTLER 9257A measuring unit+Dynoware software (Kistler AG)		
Machine tool	Mazak Nexus 410A Vertical Center		

3 Tests' Results in Case of Application of Minimal Quantity of Lubrication

The suitability differences between the coating types have been reflected well by the results of the tests, carried out by us. Based on our experiences, the most evaluable results have been gained at forming speed value of 20 m/min. At the lowest forming speed value (14 m/min) the tools have worked with „too" long tool life, therefore (due to time and material saving) we have discontinued the execution of tests after having achieved a certain number of machined holes. The highest forming speed value (25 m/min) has been found too high in case of certain coatings: due to the very low number of machined holes these types could not be used under industrial circumstances.

Figure 2 shows the phases of thread forming operations, furthermore, the deterioration process of tool teeth and the development of vibrations. The left side of the picture shows the temporal development of forming operation parallel with the curve of torque development, registered by the torque measuring device where,

- the *1st part* is the so-called running-in phase when the roughing part of the thread forming tool enters the hole (it means a three-thread upgrade),
- the *2nd part* is the phase, belonging to the gradual development of thread. A small increase can be noticed in the torque demand as an increasing portion of the forming tool surface will ream the wall of the hole,
- *3rd part* is the phase of reversion. Reversion means a change in the rotation direction after that the forming tools unscrews from the hole, already machined. The change in the rotation direction results in negative values of torque, the amount of this value reflects how great friction is caused by the thread forming tool in the thread during the reversion.

The development of torque can be seen on the right side of **Figure 2**, when the forming tool has already suffered a deterioration to a great extent and it has resulted in the development of vibrations.

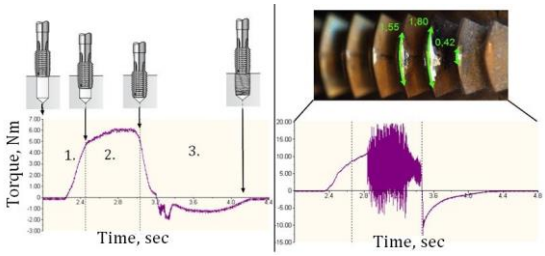


Figure 2

The phases of torque curve, the development of vibrations due to the deterioration of the tool

During the execution of the tests we have observed that every forming tool has started to work with a relatively great torque. The values, registered by us at the beginning, have decreased after having achieved 10-20 holes, after that – parallel with the increase of the tool deterioration – they have started to rise again. The drastic torque increase and the appearance of the vibrations have occured when the critical wear has developed on each forming row of the tools (**Figure 3**).

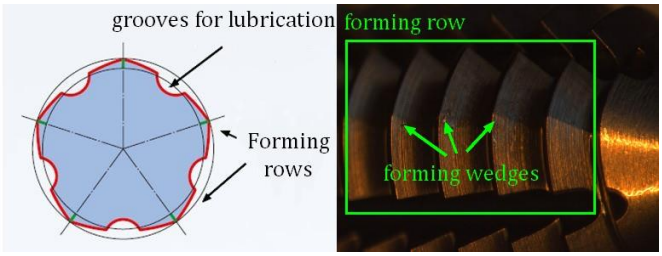


Figure 3

Basic construction of the forming tool. The left side of picture: interpretation of forming rows (section). The right side of picture: forming row and forming wedges (view from above)

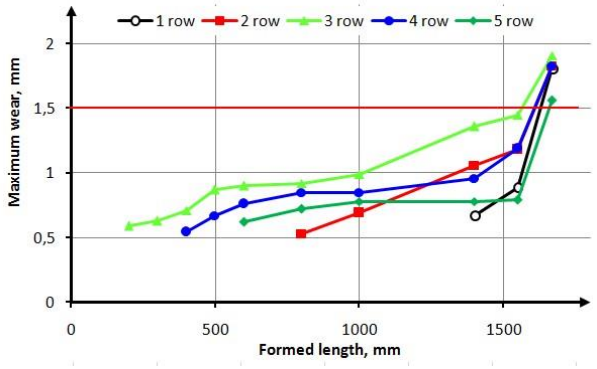


Figure 4

Maximum wear values, measured on the CrTiN-coated tool

On **Figure 4** the greatest wear values are shown, measured on the forming rows of CrTiN-coated tool, in function of the machined hole length at forming speed value of 20 m/min. In the majority of cases the wear has developed on every roughing forming wedge (i.e. on the first three pitches of the tool). Our experiences show that in case of M6 tool it is enough to pay attention to the greatest measurable wear value as it reflects the condition of the tool under test circumstances with the necessary accuracy. The thick line, marked on the diagram, shows the wear limit (1.5 mm): if this value is exceeded by the wear values, measured on every forming row, then the tool most likely will deteriorate during the formation of the next few holes.

On **Figure 5** the maximum torque values have been depicted in function of the summarised maximum wear values. The continuous graph has been drawn based on the measured values and it shows a correlation of 99 percent. Based on it we rightly suppose that there should be a tight and clear correlation between the torque demand and wear development, under such circumstances. This announcement can not be made universal for tools, having different construction and size: it would be necessary to carry out several check inspections to do this statement. As it can be seen well, there are no measured results between the two

last torque measurement values (in the zone, marked with a circle), it means we do not have information about the „behaviour" of this zone. In order to draw well-founded conclusions it would be necessary to carry out further examinations on this zone.

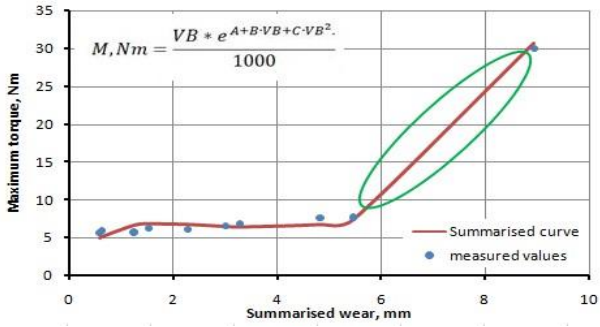


Figure 5

The development of the maximum torque values, in function of the summarised wear

The main aim of the examination process was to determine differences between the performance of different coating types. The number of machined holes is the most suitable character to do it. The maximum measured wear values have been shown on **Figure 6**, in function of the number of machined holes. The thick line, marked on the diagram, shows the maximum torque limit of 8.5 Nm: if this value is exceeded during the tests, then the tool most likely will deteriorate during the formation of the next few holes. Analysing the figure it seems to be clear that there can be even a fourfold difference between the number of machined holes (due to the coating layer type).

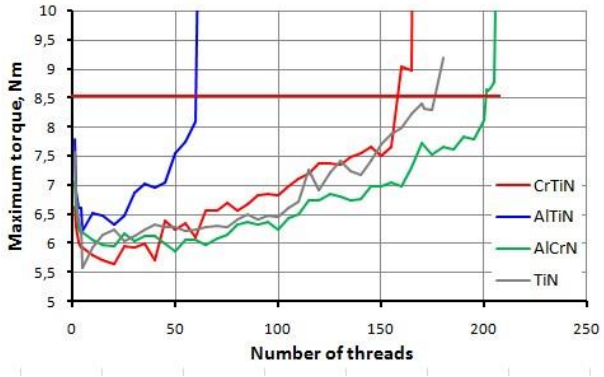


Figure 6

Maximum torque values at a forming speed value of 20 m/min

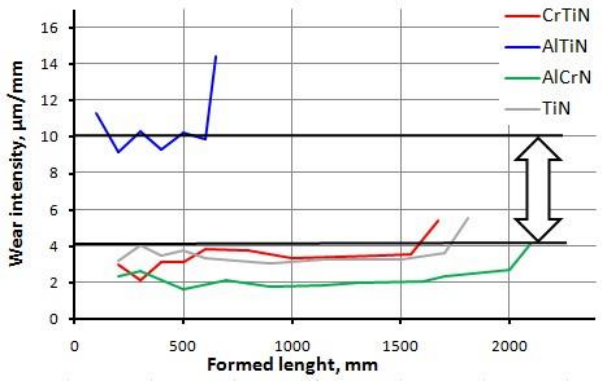


Figure 7
Wear intensities at a forming speed value of 20 m/min

The differences between the coating types can be defined numerically with the below formula:

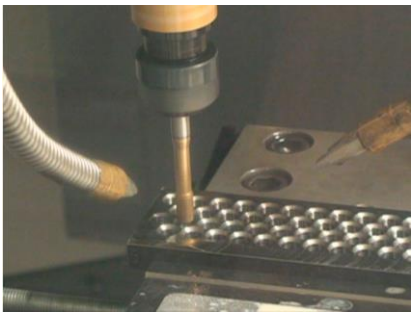
$$\text{Wear intensity}, \frac{\mu m}{mm} = \frac{\text{Cumulated wear size of rows, } \mu m}{\text{summarised machined length, mm}} \quad (1)$$

The wear intensity has been indicated on **Figure 7**, in function of the formed length. The AlTiN-coated tool excels with its considerable and oscillating wear intensity. Other coating types have worked with (approx. 60 percent) lower and almost steady wear intensity. This is an important information from the point of view of predictable deterioration process and condition monitoring of tools.

4 Comparison of Cases with MQL and Flood Type of Lubricaton

The flood type and the minimal quantity lubrication can be seen in **Figure 8**.

Tests were made at a forming speed value of 18 m/min. The core holes were made with different diameters (Ø7.3; Ø7.4; Ø7.5; Ø7.6 mm), because we wanted to get a clear picture about the effect of the change in the core hole.



a.) minimal quantity lubrication
(15-20 ml/hour)

b.) flood lubrication
(35 l/min)

Figure 8
Two types of cooling and lubrication

10 holes have been made in case of both types of lubricating methods. Tool wear has not been measured during the trials, the same tool was used for the tests in case of both lubrication methods. The **Table 2** shows the values recommended by the manufacturers.

Table 2
The measured and the real size of the core holes

Size of pilot drill, mm	Ø7.3	Ø7.4	Ø7.5	Ø7.6
Measured size (average), mm	7.36±0.01	7.41±0.01	7.58±0.01	7.63±0.01
Recommended hole diameters, mm: 7.41-7.48 mm [6]				

The registered and measured torque values can be seen in **Figure 9.** Based on the measured results, it can be observed that minimal quantity lubrication required less torque in every case. If we increase the core hole diameter, the difference between the two lubrication methods will be smaller and smaller. This phenomenon can be explained by the fact that by the increase of the core hole diameter, the forming part of the tool touches the workpiece material on a smaller and smaller surface. On this surface the adhesion reducing effect of the lubricating oil film is lower.

It can be seen, that the difference between the two types of lubrication methods is 19 percent in case of the recommended core hole size.

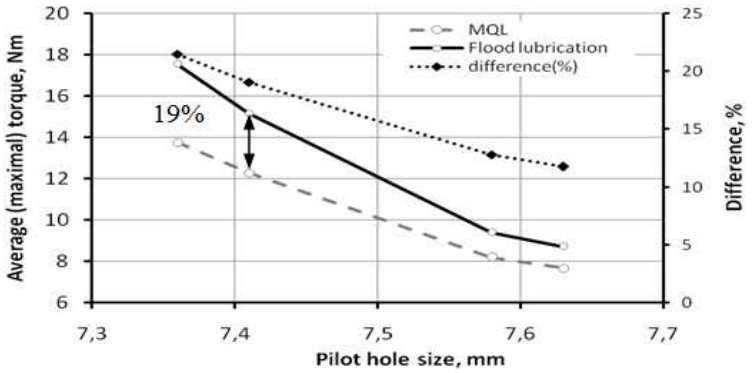
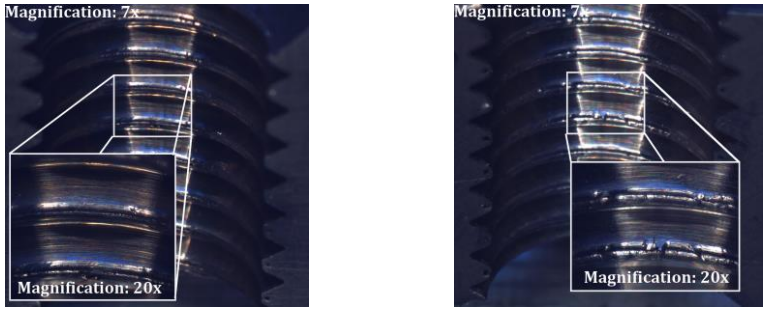


Figure 9

Average torque versus pilot hole size in case of different types of lubrication

From these results the conclusion can be drawn that the friction between the tool and the workpiece is significantly reduced by minimal quantity lubrication under the circumstances, applied by us – and parallel with it, less energy is needed for the forming operation. Our experiences and the negligible costs of the MQL unit should be considered when calculating operating costs. The difference between the two lubrication methods point to the fact that minimal quantity lubrication is less „sensitive" to the changing of core hole size. If there is a decrease in the core hole diameter (due to the deterioration of the drill) or, on the contrary, there is an increase therein, then minimal quantity lubrication can be used with greater certainty. Of course, several tests are necessary to verify this statement. We will have the opportunity to carry them out.

The machined surface can be seen in **Figure 10**. A more favourable surface finishing effect of minimal quantity lubrication can be observed on the images. (The photos were made by stereomicroscope, with different magnification.) In case of cooling and lubrication by an emulsion, the machined surface is more „ragged", cavities and droplets can be noticed.



minimal quantity lubrication cooling and lubrication by emulsion

Figure 10

Pictures of the thread – pilot hole 7.41 (7x, 20x magnification)

Summary

Some questions have been put by us during the description of the aims of the present examination process. The answers, given below, have been found:

- From the analysis of the places, where wear has developed, it has become clear that primarily the roughing-forming wedges have suffered wear, it does not depend on the coating type, tested by us. Based on the measurements, carried out by us, it has been proved that in case of forming tools, having size of M6, the maximum wear can be determined with the length of wear arch, it is 1.5 mm: if this value is exceeded by the wear values, measured on every forming row, then the tool will achieve its performance limit soon.

- It has become clear that the performance of the original tool, coated by the producing company, has been overmatched by the tools, coated by Platit (it has become clearly visible at the lowest forming speed value).

- Under the present test circumstances, the increase of the forming speed values drastically decreases the number of machinable holes, therefore the forming speed value of 25 m/min is too high.

- The wear condition of the forming tools can be followed by the torque measurement very well. From the examination process, carried out by us, it has become clear that the upper torque limit is 8.5 Nm – if this value is exceeded, the tool deterioration is expected.

- Our tests have clearly confirmed the following fact: with the use of an appropriate selected coating (i.e. optimised to the task) results can be achieved, being unapproachable by any universal („industrial") coating. This conclusion, drawn by us, can be confirmed by the following fact: at a forming speed value of 14 m/min the so-called „industrial" coating has „produced" only 15 holes, the AlTiN- and AlCrN-coating types more than 500 holes.

- The value of the forming torque may be by 19 percent lower in case of MQL (with mostly recommended pilot hole), compared to the flood type of lubrication. According to the photos, taken by the electron microscope, the surface intactness and smoothness of the thread profile is much favourable if MQL is applied, compared to the flood type of lubrication. In this way the machined thread may be better not only from point of view of strength, but – probably - from the aspect of the corrosion resistance as well.

Acknowledgement

Authors wish to thank to Mr. Tibor Cselle, CEO of Platit AG, for his versatile support and for the permission to publish the test results.

The project was realised through the assistance of the European Union, with the co-financing of the European Social Fund, namely: TÁMOP-4.2.1.B-11/2/KMR - 2011 - 0001: Researches on Critical Infrastructure Protection.

Literature

[1] T. Cselle: Dedication-Integration-Open Source - New Rules in the Coating Industry (and in progressive economics), Werkzeug Technik, 15 July 2011, Nr. 120b, pp. 22

[2] J. Destefani: Don't Cut Threads--Form'Em In the Right Application, Thread Forming can Boost Quality and Throughput, Manufacturing Engineering April 2004 Vol. 132, No. 4, pp. 59-66

[3] Halász G., Pálinkás T., dr. Sipos S.: Korszerű menetfúrók környezetbarát alkalmazása Gyártóeszközök, szerszámok, szerszámgépek (Műszaki Kiadványok, XVI. évf.), 2011, pp. 3-9

[4] H. Sağlam, R. Kuş: Performance of Internal Thread Rolling Head and The Mechanical Properties of Rolled Thread, 6th International Advanced Technologies Symposium (IATS' 11), 16-18 May 2011, Elaziğ

[5] Walter Product Handbook (Drilling&Threading) The Perfect Thread, p. 62

[6] Emuge Threading Technology: InnoForm Cold-forming Taps pp 22., EMUGE-Werk Richard Glimpel GmbH & Co. KG Fabrik für Präzisionswerkzeuge, 2011

Multidimensional Scalling Analysis of the World Economy During the Period 1972-2012

J. A. Tenreiro Machado [1], Maria Eugénia Mata [2]

[1] Coordinator Professor, Department of Electrical Engineering, Institute of Engineering, Polytechnic of Porto, Porto, Portugal, jtm@isep.ipp.pt

[2] Associate Professor, Nova SBE, Universidade Nova de Lisboa, Faculdade de Economia, Campus de Campolide, Lisbon, Portugal, memata@fe.unl.pt

Abstract: The last 40 years of the world economy are analyzed by means of computer visualization methods. Multidimensional scaling and the hierarchical clustering tree techniques are used. The current Western downturn in favor of Asian partners may still be reversed in the coming decades

Keywords: Economic development; Multi-dimensional scaling; global hegemony

1 Introduction

The Western Golden Age of economic growth that lasted from the end of the Second World War to the early 1970s consecrated the global hegemony of the United States [21]. Following two victories in the two World Wars without suffering any destruction within the nation's territory, the United States' role as a model of modern technologies and venture capital corporations became quite clear on the international scene [23]. Superior management abilities drove capitalistic professionalism, technological superiority, considerable growth rates, and a global upper hand [29, 26].

The macroeconomic environment of the early 1970s was not considered to be especially favorable to investment, according to venture capital experts [22]. A less careful approach may consider the 1974 oil price increase as the historical turning-point for a long-run downward trend. In spite of the sizable amount of America's domestic oil production, challenges to the industry in the non-Arab world may be cited. However, the OPEC cartelization undertaken to increase oil prices was justified as a defense against the dollar depreciation, which eroded the value of oil exports from the Arab producers. The dollar depreciation was the direct consequence of the end of the Bretton Woods monetary system, bringing to an end the 1944 commitment to fixed exchange rates [18]. The extraordinary

military expenditures associated with the Vietnam War had led to a dollar glut and the need for the inconvertibility that was consecrated in the Bretton Woods system.

From then until today Europe and America have faced several recoveries and recessions, along with other important global changes that occurred in the 1980s, owing to the breakup of the Soviet Union and the end of certain communist regimes around the world, including China [14]. The expansion of corporate enterprises into those new areas brought positive capitalism and business opportunities [17]. The unification of Germany fueled hopes for a stronger European Union. A common currency, the euro, intensified European cohesion at the turn of the millennium, which was also a hopeful aspect for Western economic growth in the 1990s.

The new millennium ushered in fierce international rivalry among the largest international partners [13]. Capitalist Russia tends to control neighboring areas, threatening European tranquility. India became a talented front-runner in IT and service sectors, China welcomed large global corporations and began a rampant industrialization that floods all world markets with inexpensive convenience goods. Other Asian partners became small tigers with modern and fashionable competitive economies (Taiwan, South Korea, Indonesia, and Singapore) [7]. They all have been able to cultivate the ranks of engineers, statisticians, computer experts, chemists, and other qualified professionals that are required in modern economies and societies [27]. Thanks to generous educational programs and government policies, Asia is favoring tertiary enrollment. Nothing comparable is occuring in Africa [2].

In the 2007 crisis, the Nobel laureate Robert Fogel made the forecast that in 2040 the existence of democracy will not depend on the usual world democratic bastions, the EU, the U.S., and Japan, but will depend rather on the new global hegemony of Asian partners [15]. The aim of this paper is not to discuss democracy, an issue that is better left to political science and political economy, but to observe the convergence process, using Multi-Dimensional Scaling (MDS) methodologies, to chart the path of these partners throughout the last four decades of Western decline (1972-2012).

Four decisive World Bank social-economic indicators were selected to examine the comparative evolution of a dozen countries' processes, and describe them until 2012 (as no data are yet available for 2013). Data and methodology are discussed in Section 2. We divide the period into two twenty-year groups, and also look at the whole forty-year process. Section 3 discusses the results obtained on economic convergence, and includes many more countries [25, 20]. Section 4 summarizes the main conclusions.

By increasing "the ultimate potential of the economy", convergence may lead to global prosperity (Current World Bank calculations say that "the number of people living below the $1.25 a day line plunges from 1.2bn people in the

developing world to fewer than 600m") [16]. However, it may also bring "potential warfare economic strategies", because war may compensate victors with greater territory and more resources, thereby increasing their production possibilities frontiers [23].

Our conclusions may not be altogether surprising, but they demonstrate that the world economic system is following a path of convergence on which the Asian partners still have a way to run, while Russia is looking very much more like the most developed countries.

2 Methodology and Data

Economic growth means new and better conditions of life at per capita levels, diversified consumption that is reflected in foreign trade openness (increasing imports and exports), as well as increasing population longevity, translated into higher life expectancy, while youth education becomes a driver for equal social opportunities. As education endogenously provides the human capital that is required in more and more sophisticated technological systems, this is a decisive long-run indicator for the ultimate potential of a national economy (for life expectancy see [1]). A stock variable such as schooling years for citizens above 25 (or 15) years of age may be the most representative, as it is more revealing about labor markets and production efficiency, but flow variables on education are longer and more complete.

Data were collected from the World Bank national development indicators covering the period from 1972 to the present (2012), sourced in [3]. The series are homogeneous, and the tertiary education level is the variable that most relates human capital with technological achievement throughout the period analyzed. To discuss the similarities among the 12 countries, the potential economic growth, and global hegemony of the main world powers, the selected indicators are GDP per capita, the weight of international trade in GDP, life expectancy at birth, and tertiary enrollment. Moreover, the human development index concept, based on indicators for human development, also recommends these same indicators [24].

The 12 selected countries (representing all continents with the exception of Africa) are {BRA, CHN, DEU, FRA, GBR, HUN, IND, ITA, JPN, PRT, RUS, USA} ≡ {Brazil, China, Germany, France, UK, Hungary, India, Italy, Japan, Portugal, Russian Federation, US}.

Here is the complete description of the four indicators:

- GDP per capita, comes from NY.GNP.PCAP.KD. It is the GNI per capita (constant 2000 USD$). The preference for per capita values has to do with the need to observe countries with very different economic

dimensions. Partners such as China, Russia, and India are much larger economies in which larger populations contribute to the GDP. The per capita indicator makes comparisons more plausible, in spite of the many differences that international comparisons always involve (e.g. climate, culture, natural resources, politics). GNI per capita is gross national income divided by midyear population. GNI (formerly GNP) is the sum of value added by all resident producers plus any product taxes (less subsidies) not included in the valuation of output plus net receipts of primary income (compensation of employees and property income) from abroad. Data are in constant 2000 U.S. dollars.

- Annual exports of goods and services (percent of GDP), from NE.EXP.GNFS.ZS. Exports of goods and services represent the value of all goods and other market services provided to the rest of the world. They include the value of merchandise, freight, insurance, transport, travel, royalties, license fees, and other services, such as communication, construction, financial, information, business, personal, and government services. They exclude compensation of employees and investment income (formerly called factor services) and transfer payments. It is a weighted average. Data are expressed as a percentage of GDP.
- Life expectancy comes from SP.DYN.LE00.IN. Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
- School enrollment, tertiary (gross percent) comes from SE.TER.ENRR. Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education shown. Tertiary education, whether or not to an advanced research qualification, normally requires, as a minimum condition of admission, the successful completion of education at the secondary level. Data are expressed as a percentage [4].

The time series for tertiary enrollment is the least dense, but missing observations occur in all available education time series. Barro & Lee's (2012) database for 1950-2010 is an excellent source for education, but the information is only given for every five years [12]. The UNESCO education database has yearly information, but covers only 1970-1997, with missing points.

MDS is a statistical, computational visualization methodology that examines similarity between objects through distance estimations, by means of graphical relatedness representation [10]. Similarity means the "likeness" between the objects in a mathematical degree [19].

Any MDS representation of n objects departs from estimating a $n \times n$ symmetrical matrix **R** for cross correlation measures between all objects. In the MDS map

objects are points and they represent the data vectors that describe the selected variables of the exercise. If the correlation between two objects is one, this means that the distance between them is zero. A zero correlation means an infinite distance, and the corresponding MDS points are graphically very far apart. If two points almost coincide there is a large correlation, while if they are located far away in the MDS plot, the correlation between the corresponding data vectors is small [8].

In order to estimate the coordinates of the points in an m-dimensional plot, the MDS numerically optimizes the estimation of the distance between all pairs of objects, according to matrix **R** of correlations [9]. The use of $m = 2$ or $m = 3$ is very frequent, because it represents a direct graphical visualization of the MDS map in 2 or 3 dimensions, respectively.

In this paper the country economies are objects characterized by 4 variables performing during 40 years. The main diagonal of the **R** symmetric matrix $\mathbf{R} = [r_{ij}]$ is composed of ones/zeros for correlation/distances, while the other elements of the matrix must be described as $0 \le r_{ij} \le 1$, $r_{ij} \ge 0$, $i,j = 1,..., n$. MDS maps are not sensitive to translations or rotations because the method uses relative measurements. Axes have only the meaning and units (if any) of the measuring index. Objects are rearranged so that the map obtained can best approximate the similarities that exist. For example, study [28] applies it to genomic datasets. The accuracy of the MDS solution is measured using the raw stress. The smaller its value, the more accurate is the fit. If stress is plotted versus the m dimensions of the MDS map, a monotonic decreasing chart is obtained. Users may choose the best dimension as a trade-off between lower stress levels and the dimensions for the map representation.

Some literature refers to MDS as a statistical tool, while other authors mention it as a computer representation scheme. The important thing is that it can provide a means of visualizing items without *a priori* restrictions or any additional constraints. Of course different indices produce different maps, because MDS maps reflect the measures for similarity. The great value of MDS plots is the direct visualization of results they provide.

Time dynamics analysis may require the explicit division of the total time period into several sub-periods of width h to be considered by MDS as independent objects. For a total time period T, $p = T/h$ samples are produced and the number of MDS points increases proportionally in the plots. The time samples to be adopted are a compromise between the possibility of understanding fast dynamics (with small values of h, meaning many sub-periods), and the advantage of plotting a limited number of MDS points (which requires large values of h). As time series cover a period $T = 40$ years (1972-2012), two cases are developed, namely a comparison based on the whole period of time, that is $h = 40$ ($p = 1$) and a division into two sub-periods of 20 years each $h = 20$ ($p = 2$) for 1972-1992 and 1992-2012. We obtain $n = n_c \times p$ objects to be analyzed in the MDS of h years length

each. Therefore, the two cases consist of $p = 2$, $h = 20$, $n = 40$, and $p = 1$, $h = 20$, $n = 20$. Points labeled as "USA1" mean USA during the first twenty-year period (1972-1992), and JPN2 means Japan during the second twenty-year period (1992-2012). Of course $l = 4$ economic variables {GDP per capita, openness, life expectancy, tertiary education enrollment} are adopted for each country economy, having identical weights.

The data set of the 4 economic variables had missing values. Some countries had some years without data values, which had to be estimated by means of linear interpolation between adjacent years (in particular the Russian Federation, for all variables). Germany had many missing values for the tertiary education enrollment. An iterative procedure was adopted to avoid anomalous values: (*i*) all variables' time series were plotted, for the whole set of countries (*ii*) a non-linear individual trend line was estimated for each country, (*iii*) the values emerging from a given trend line were compared with the succeeding values (in order to test the consistency with the real values) and with the values of the remaining countries (to test the relative positions), and (*iv*) the trend line was accepted or replaced by another one, depending on whether the visual comparison did or did not confirm. This was a time-consuming procedure, but it prevented the acceptance of misleading values. The practical implementation of this careful approach proved to be faster than expected because of the smooth path of the variables, which tend to evolve similarly. Different trend lines have produced values relatively close and, therefore, the estimation error is not significant.

Two methods can be used to construct the matrix $\mathbf{R} = [r_{ij}]$, the cosine correlation and the Euclidean distance, defined as:

$$r_{ij} = \frac{\sum_{t=1}^{h}\sum_{k=1}^{l} x_i(k,t)x_j(k,t)}{\sqrt{\sum_{t=1}^{h}\sum_{k=1}^{l} x_i^2(k,t)\cdot\sum_{t=1}^{h}\sum_{k=1}^{l} x_j^2(k,t)}}, \quad i,j = 1,\cdots,n \tag{1}$$

$$r_{ij} = \sum_{t=1}^{h}\sum_{k=1}^{l}\left[x_i(k,t) - x_j(k,t)\right]^2, \quad i,j = 1,\cdots,n \tag{2}$$

where x_i and x_j denote economic variables for the i-th and j-th objects, t and k are two dummy indices for time and type of economic variable, h is the sampling period, and n is the total number of objects. Equation (1), which measures the angle between two vectors, is often called the cosine coefficient because it denotes an angular metric, and is a normalized inner product [11]. Equation (2) describes the Euclidean distance over the time period. Note that expression (1) is not sensitive to the amplitude of the vectors, while expression (2) captures differences between vectors, on both amplitude and direction.

The two-decade approach for describing the relative similarities among countries seems to be successful. The use of shorter periods would allow for more time detail, but plots would contain many more points, and reading would become increasingly difficult. The twenty-year visualization also may have a special economic meaning in capturing any Kondratief business-cycle influence for the behavior of the countries' economies. Expressions (1) and (2) have an implicit embedded description of the time evolution. As a result, the next plots of cases a) and b) differ according to equations (1) and (2), because they reflect time dynamics more or less explicitly.

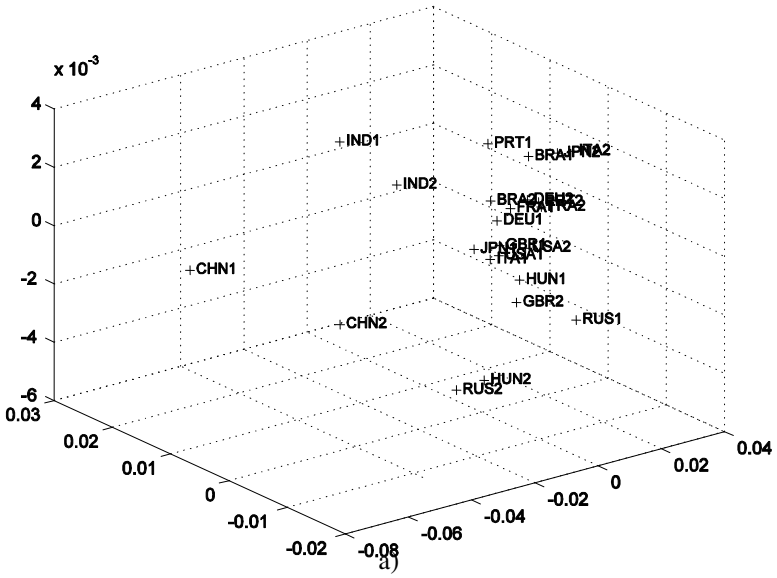
3 Discussing the Estimations

Looking at the cosine correlation maps (Fig. 1a), one can follow countries' evolution along distance vectors, revealing the path of the comparison among the 12 sampled countries during the two twenty-year periods (1972-2012). China, India, and Russia occupy different positions according to the four socio-economic indicators selected, in a three-dimensional view. In spite of their common policies of promoting high economic growth rates, they still remain far from the most developed countries of the world economic system, which are so similar that a real cloud is formed by the Western World set of European countries, and the USA [5].

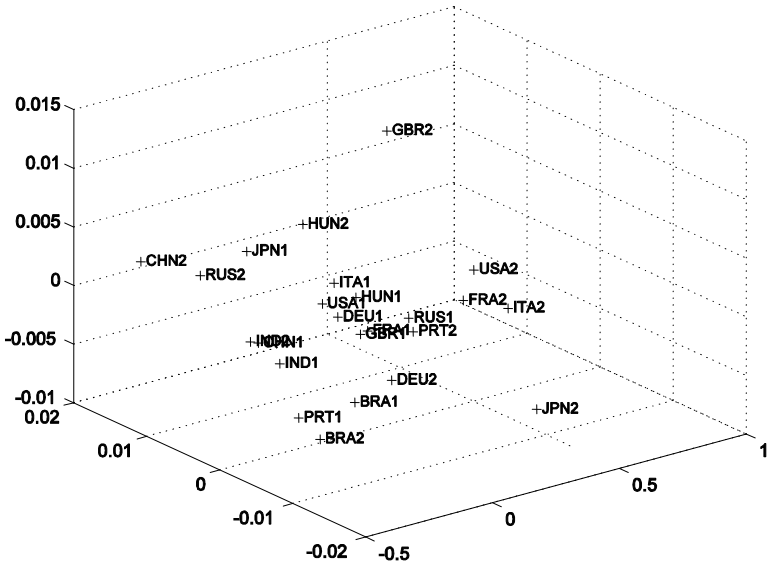
The Euclidian-distance plot (Fig. 1b) shows how China and Russia could perform well in becoming more similar (closer) to Japan and the most developed partners. India also converged, in spite of the social turmoil resulting from the prevailing caste system, "which divides the population into a hereditary hierarchy that determines economic and social opportunities" [15]. This handicap has relaxed somewhat as a result of government policies: "The government has sought to offset the discrimination against lower castes with educational subsidies" [15]. Nevertheless, the caste system remains rigid in rural areas.

The MDS method also allows for a projection in two dimensions (Fig. 2), which confirms that Asian partners still have a long way to go in to become "similar" to the Western partners. Russia and Brazil are the two partners that moved to positions closer to the large cluster of developed countries in the last 20 years. Brazil shows the success of the government policies implemented, as well as the discovery of oil in the Brazilian offshore Atlantic. Russia also followed a move from its 2002 position until now (2012 data), reflecting its openness, as well as oil and gas exports. How to qualify this move?

Using the Euclidean distance instead of the cosine correlation pattern for distance measuring, the same messages can be seen in this different representation. In becoming more similar to the large cluster of developed countries, Russia also converged to a position closer to the Japanese (and Chinese) profile.



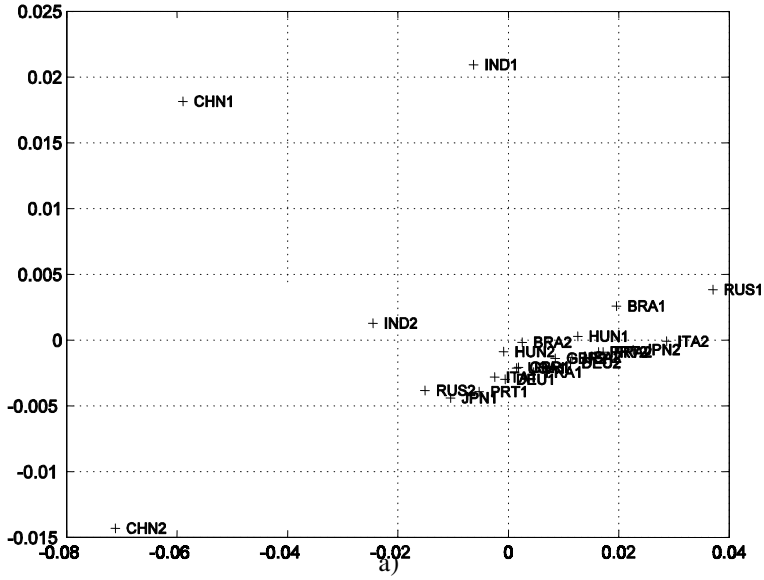
a)



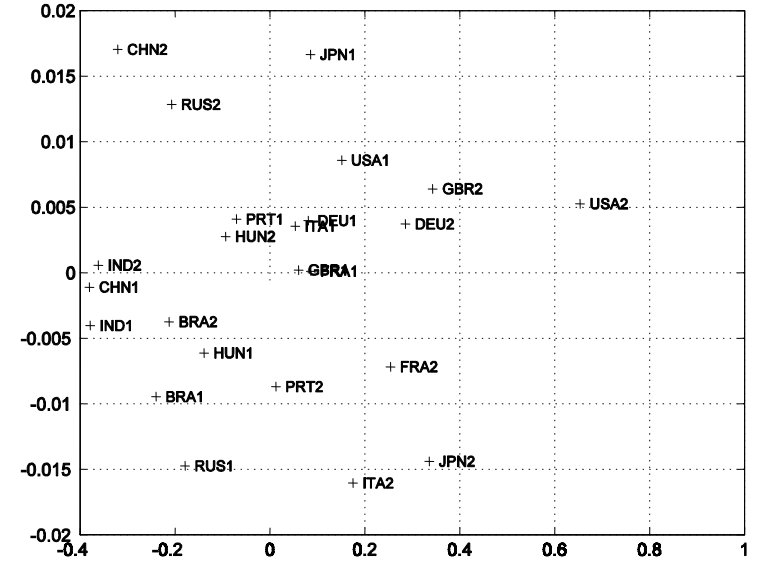
b)

Figure 1

Three-dimensional MDS representation with $m = 20$ for the 12 countries and two 20-year periods, based on the 4 selected variables). Using: a): Cosine correlation (1), b) Euclidean distance (2)



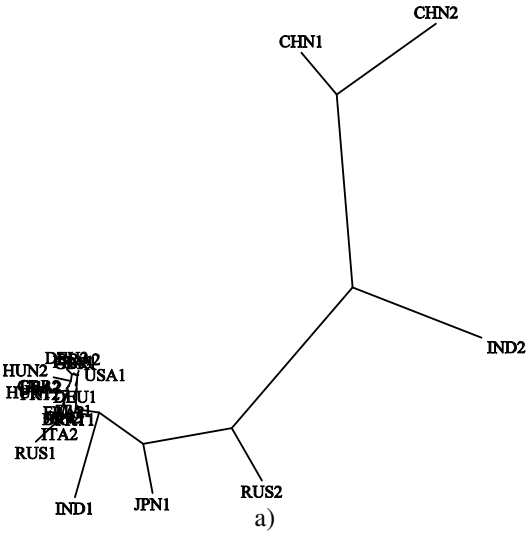
a)



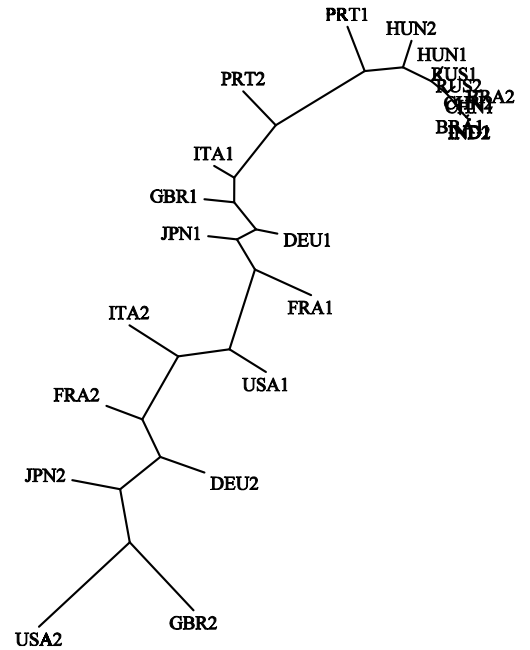
b)

Figure 2

Two-dimensional MDS representation with $m = 20$ for the 12 countries and two 20-year periods, based on the selected variables). Using a) Cosine-correlation (1), b) Euclidean distance (2)



a)



b)

Figure 3

Hierarchical clustering tree representation with $m = 20$ for the 12 countries and two 20-year periods, based on the selected variables). Using a) Cosine correlation (1), b) Euclidian distance (2)

The robustness of the MDS representations and related conclusions were always checked through Sheppard tests and stress tests, which indicated good results.

Alternative ways of plotting the countries' relative positions are given now, using hierarchical clustering tree representations. Hierarchical clustering is a common statistical technique for data analysis. The goal is to build a hierarchy of clusters, in such a way that objects in the same cluster are, in some sense, similar to each other. Clusters are combined or, alternatively, split, based on a measure of their dissimilarity. This is achieved by adopting an appropriate metric, quantifying the distance between pairs of objects, and a linkage criterion, defining the dissimilarity between clusters. They are, perhaps, the most suggestive representations for visualizing results.

In the tree of Fig. 3 the United States and the European partners in general remain in top positions on the world scene. The Asian dissimilarity is confirmed, and Fogel's forecasts may have been too pessimistic.

For a better sum-up of the 40-year process (1972-2012), the following 3-dimensional MDS maps (Fig. 4) help to visualize the distances that separate the partners in the sample, in considering the four selected socio-economic indicators.

The use of a 2-dimensional plot (Fig. 5), or hierarchical clustering tree representations (Fig. 6), may be also useful.

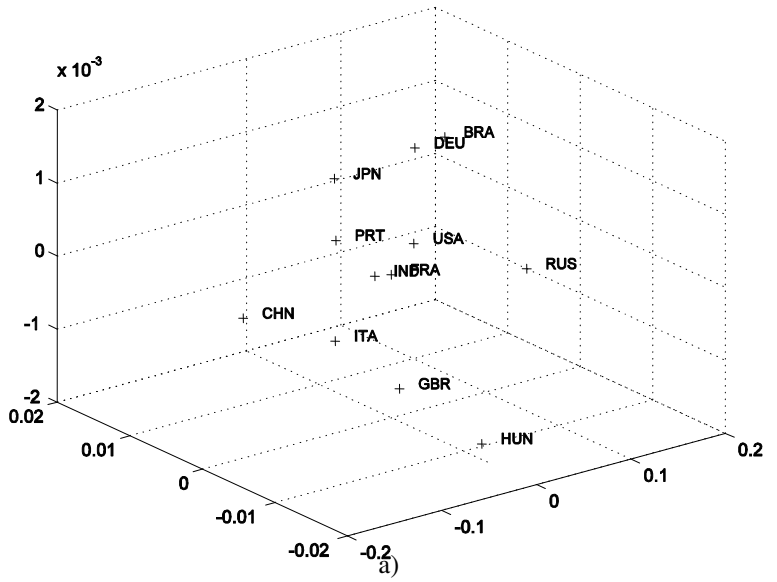
The most notable similarities occur among the Europeans, the two American countries (USA and Brazil), and Russia. The Asian partners will need some time to reach the current degree of similarity seen in the most developed world. Using Euclidean distances the proximity of Russia to small European partners stands out. If some worries exist in terms of possible military aggression, invasion, or hegemony, these relative positions tell us something about the current peaceful global balance.

The same robustness tests were carried out to assure confidence and trust in the results obtained. All tests give credibility and robustness to the plots and their respective conclusions.

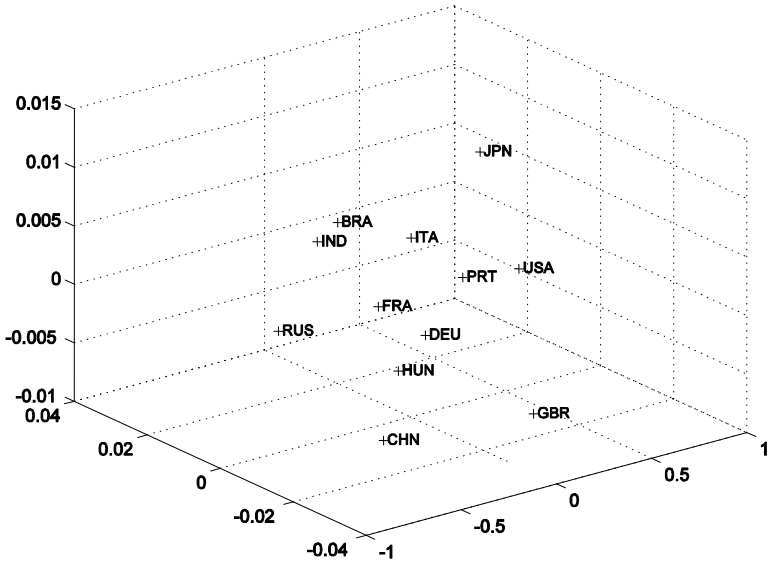
Conclusions

There is considerable convergence in progress among the world's leading countries, and data for 12 partners including European, American, and Asian nations, leads to estimations for comparative views among them, using the MDS methodology.

For economists the problem seems to be more difficult for Europe, where unemployment undermines social stability and reinforces the stagnant local demography. As it is mainly affecting the European youth, it postpones marriage and hinders fertility and birth rates, leading to a precarious social stability. This crisis may be a cyclical problem, as in long-run MDS views trends go favorable to European countries' economic growth [6].



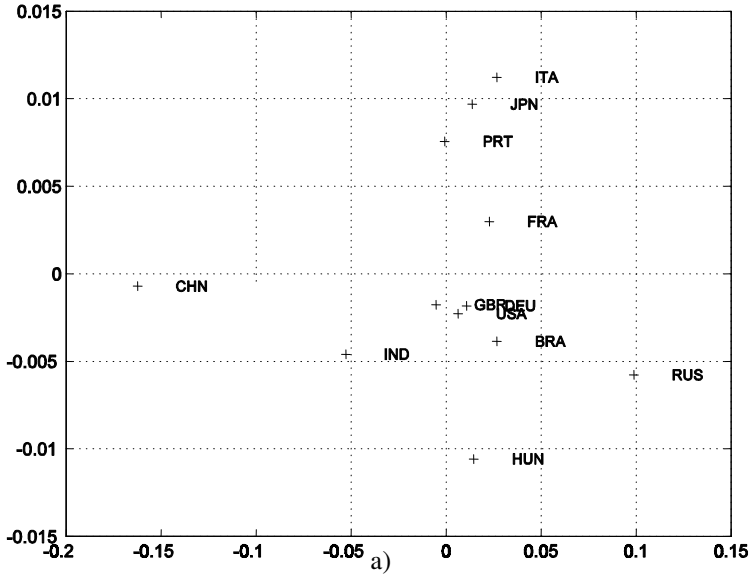
a)



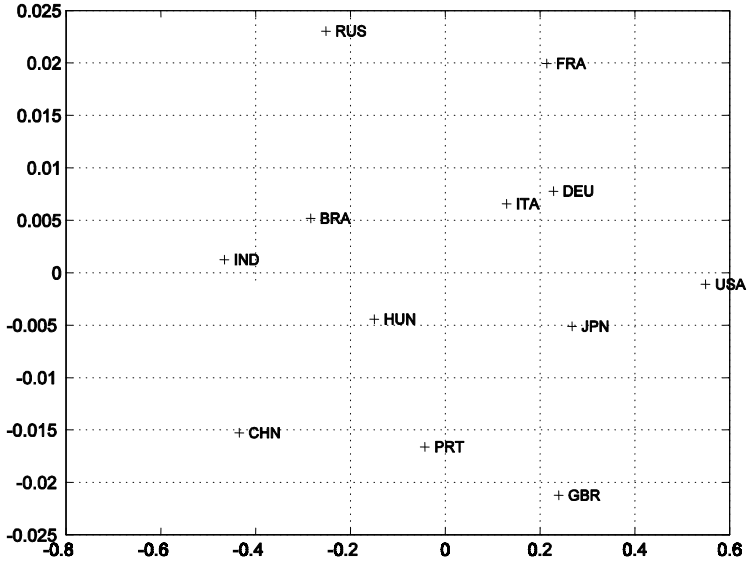
b)

Figure 4

Three-dimensional MDS map with $m = 40$ for the 12 countries and one 40-year period, based on the selected variables). Using a) Cosine correlation (1), b) Euclidean distance (2)



a)



b)

Figure 5

Two-dimensional MDS map with $m = 40$ for the 12 countries and one 40-year period, based on the selected variables). Using a) Cosine correlation (1), b) Euclidean distance (2)

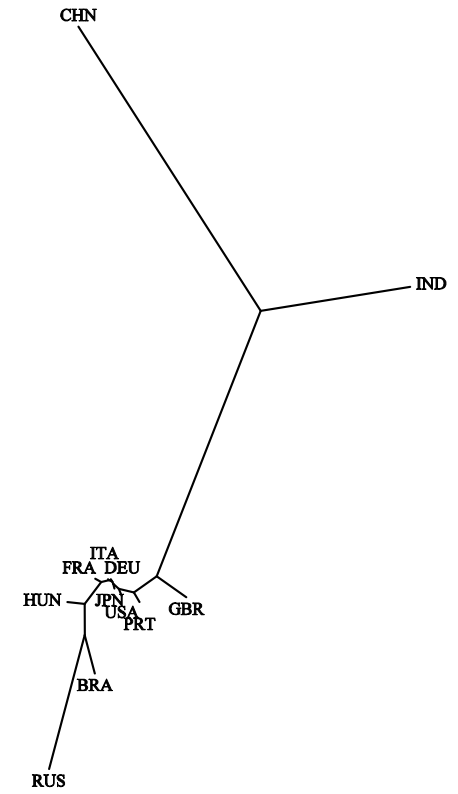


Figure 6

Hierarchical clustering tree representation with $m = 40$ for the 12 countries and one 40-year period, based on the 4 selected variables. Using a) Cosine correlation (1), b) Euclidean distance (2)

Based on data for four socio-economic indicators throughout the last 40 years, it is possible to conclude that Asian partners still need some more decades before similarities can be comparable to the those that exist among the most developed

partners. Russia, however, presents a convergence path, particularly in the last 20 years, that deserves to be emphasized. Strategists, military experts, and politicians can learn a lot from the observation of these estimations. European difficulties, even resulting from a business-cycle context, may witness a noticeably negative influence, if economic warfare takes place, because they may reduce the Milward's ultimate potential of the European partners' economies.

References

[1] D. Acemoglu, and S. Johnson, Disease and Development: The Effects of Life Expectancy on Economic Growth, Journal of Political Economy, Vol. 115, n. 6, 2007, pp. 925-985

[2] D. Acemoglu, and J. A. Robinson, Why is Africa Poor, Economic History of Developing Regions, Vol. 25, n. 1, 2010, pp. 21-50

[3] World Bank, World Bank national development indicators, 2013, http://data.worldbank.org/data-catalog/world-development-indicators

[4] R. Barro, and Jong-Wha Lee, A New Data Set of Educational Attainment in the World, 1950-2010, NBER Working Paper No. 15902, April 2010, (revised 2012)

[5] J. Belak, B. Milfelner, Enterprise Culture as One of the Enterprise's Key Success Factors (Integral Management Approach): Does the Internal and External Cultural Orientation Matter?, Acta Polytechnica Hungarica, Vol. 9, No. 3, 2012, pp. 27-44

[6] P. Benedek, Compliance Management – a New Response to Legal and Business Challenges, Acta Polytechnica Hungarica, Vol. 9, No. 3, 2012, pp. 135-148

[7] M. Bordo, The Global Financial Crisis of 2007-08: Is it Unprecedented? (with John S. Landon-Lane). NBER Working Paper No. 16589, December 2010

[8] I. Borg, Modern Multidimensional Scaling-Theory and Applications, 2nd edition, Springer-Verlag, New York, 2005

[9] S-H Cha, Taxonomy of Nominal Type Histogram Distance Measures, American Conference on Applied Mathematics (MATH'08), Harvard, Massachusetts, USA, March 2008, pp. 24-26

[10] T. Cox, and M. Cox, Multidimensional Scaling, 2nd edition, Chapman & Hall/CRC, 2001

[11] E. Deza, and M. M. Deza, Dictionary of Distances, Elsevier, 2006

[12] Education database for 1950-2010, http://www.barrolee.com/data

[13] B. Eichengreen, Financial Crises and What to Do About Them, Oxford University Press, 2002

[14] B. Eichengreen, Global Imbalances and the Lessons of Bretton Woods. The MIT Press, 2006

[15] R. Fogel, Capitalism and Democracy in 2040: Forecasts and Speculations, NBER Working Paper 13184, June 2007

[16] R. Harding, US economy grows 4.6%, The World Blog, Financial Times, May, 10, 2014 http://www.ft.com/intl/world/us

[17] G. Jones, (ed.), The Making of Global Enterprise, London: Frank Cass, 1994

[18] C. Kindleberger, A Financial History of the Western World, Cambridge, Cambridge University Press, 1993

[19] J. Kruskal, Multidimensional Scaling, Newbury Park, CA, Sage Publications, Inc., 1978

[20] J. Machado, M. Mata, A multidimensional scaling perspective of Rostow's forecasts with the track-record (1960s-2011) of pioneers and latecomers, in J. Awrejcewicz, M. Kazmierczak, P. Olejnik, J. Mrozowski (editors) Dynamical Systems, Theory, Łódź, Poland, 2013, pp. 361-378

[21] A. Maddison, The World Economy, A New Millenium Perspective, Development Centre Studies, 2001

[22] M. Martorelli, Professionalizing Venture Capital, Financial History, Museum of American Finance, Vol. 109, Winter, 2014, pp. 34-37

[23] A. Milward, War, Economy, and Society, Berkeley, Los Angeles, University of California Los Angeles, 1977

[24] L. Prados, Capitalism and Human Development 1870-2007, University Carlos III, Madrid, mimeo. Accessed on 10 January 2012

[25] W. Rostow, Stages of Economic Growth, A Non-Communist Manifesto, Cambridge University Press, 1960

[26] Z. Szabo, E. Herman, Productive Entrepreneurship in the EU and Its Barriers in Transition Economies: A Cluster Analysis, Acta Polytechnica Hungarica, Vol. 11, No. 6, 2014, pp 73-94

[27] P. Tóth, Learning Strategies and Styles in Vocational Education, Acta Polytechnica Hungarica, Vol. 9, No. 3, 2012, pp. 195-216

[28] J. Tzeng, H. Lu, W. Li, Multidimensional Scaling for Large Genomic Data Sets, BMC Bioinformatics, Vol. 9, 179, 2008, pp. 1-17

[29] H. Van der Wee, Prosperity and Upheaval, World Economy, 1945-1980, California, University of California Press, Penguin Books, 1986

Application Security through Sandbox Virtualization

Liberios Vokorokos, Anton Baláž, Branislav Madoš

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovakia
liberios.vokorokos@tuke.sk, anton.balaz@tuke.sk, branislav.mados@tuke.sk

Abstract: This article is aimed at the creation of a secure, virtualized system sandbox environment at the level of the respective applications. The proposed sandbox model allows us to generate a secure environment for various untrusted applications and resolve potential security incidents, such as zero day vulnerabilities. The resulting work is a functional sandbox within the MS Windows operating system, which protects the system against potentially hazardous applications. The sandbox has a minimal impact on the semantics and the time of the executed program and provides an efficient sandbox configuration interface.

Keywords: system sandbox; operating system security; virtualization; security policy

1 Introduction

Security threats are closely related to the overall level of software security in computer systems. When writing applications, various errors occur; these are then eliminated by testing and real-life usage. Insufficiently tested applications cause a number of problems, especially operating system infections, caused by various kinds of malware. These errors are often found only post-mortem. The attacker managed to abuse the fault. Subsequently, the system vendor creates a patch and updates the application. However, the attack has already happened and the system may have remained infected. Another problem is that the end user must react immediately to the various security measures, since (s)he has a direct interest in the operating system security policy or the application itself. There is a significant amount of malware on the Internet – claiming to be useful and necessary, but containing adware, viruses or other malicious code. Antivirus solutions are not capable of detecting these programs on time and reliably at all times.

The method of using system sandboxes is based on the idea of preferring prevention instead of detection in case of security. In general, there are multiple categories of sandboxes and system containers – each of these has its own specific

use. The most widely used are virtual systems isolating the individual applications, including the operating system itself. This type of – full – virtualization degrades the system in terms of performance and it has a significant administration overhead too. In this article, we describe a lightweight application container, which would resolve the above security threats and have minimal impact on the system.

2 Sandboxes and Virtualization

The system sandbox comprises a number of programs. The specific definition depends on the solution type. Hoopes defined the sandbox universally as software providing a monitored and controlled environment, in which no unknown program may cause any damage to the system it is running in [1].

A more specific definition states that a sandbox allows applications to be executed so that these applications are not allowed to read or write the data beyond the specified path, i.e. beyond the sandbox. In a broader sense, one has to add the control and allocation of operating system resources to this definition, such as network services, hardware management, low-level access, etc.

The concept of controlled execution was first time introduced by R. Wahbe et al. in the context of software bug protection [2]. The term coined for this technique – sandboxing – stands for a method of isolating untrusted modules executed in the same address space as the trusted modules, with a minimal impact on execution time.

Security in general, system security and the field of system sandboxes are all closely related to virtualization. Virtualization is a set of procedures and techniques, which allow the separation of the available system resources into multiple environments. The virtualized environment may be tailored to the needs of the users, it is easier to use or to hide the details (such as the physical location of the hardware resources) from the users. Virtualization is performed by a virtual machine (VM). A virtual machine represents an independent environment and/or a software implementation of a machine executing programs, similarly to a physical computer [3].

Virtualization is possible at various levels: from the whole computer down to its individual hardware components, or even a certain specific software environment. The various virtualization types may differ in the level of isolation, the resource requirements, the execution costs, scalability and flexibility. In general, the “closer” the virtualization level is to hardware, the more the virtual machines are isolated and separated from the host computer; however, those require also more resources and are less flexible. The "farther" the virtualization level is from the hardware, the more powerful and scalable these virtual machines are.

When used correctly, virtualization has a number of advantages – Hoopes summarized these as follows: *consolidation, reliability* and *security* [1]. Consolidation includes a more efficient use of computers, simplifies migration to newer versions of various systems, significantly shortens time needed for development and testing and allows a single physical platform to host various operating systems. Reliability has become a priority of many IT companies, more than ever. A system error in the virtual machine does not affect the other parts of the system on the same hardware platform – this ensures the reliability and the consistency of the system as a whole. Technology providing protection against application faults provides isolation from security faults. If the security of a specific part of the virtual machine is compromised, it may be terminated at any time.

In accordance with the characteristics of the respective virtualization methods, the system sandbox is a specific virtual machine. Similarly to virtualization excluding consolidation, reliability and security are the basic motives of the creation of system sandboxes. A sandbox and/or its virtualization layer may be created at various levels. Isolation of multiple virtual machines or the virtual machine and the host environment makes virtual machines an efficient platform to execute potentially error-prone and untrusted applications. While using virtual machine technology, it is a common requirement to execute potentially dangerous operations in the created VM, which is an instance of the operating system (OS) host environment. This can be achieved by using a virtual machine at hardware level. However – as it was stated above – this is not too efficient, because virtual machines are fully isolated from each other and each of these has a separate OS running in it. The initialization of such a machine has a high overhead in terms of resources and has a start-up delay too. Considering the distance of the virtualization layer from hardware, in most cases software designated as a sandbox is implemented as a layer between the operating system and the individual processes. In this solution, we focus on the isolated environments implemented at operating system level, depending on the OS virtualization model, in accordance with the scheme depicted in Figure 1.

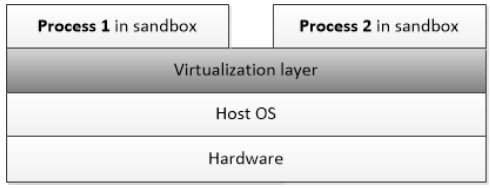


Figure 1
The position of the sandbox virtualization layer within the OS

3 Implementing the Security Policy using a System Sandbox

Access to the individual system resources is implemented by means of a unified interface provided by the operating system. Therefore, the system sandbox must provide access control of the individual OS components within this security implementation. The most important OS components are:

- files
- the registry
- network interfaces
- the CPU
- I/O
- locks
- processes
- other

Seen from the aspect of the proposed solution of secure sandboxes in MS Windows, determined by the architecture of the OS itself, the key system resources are the following:

- Files: files contain all user and operating system data. Changes in the file system without the knowledge of the user could be fatal. Removed or modified system files would directly compromise OS security. Therefore, controlling file access is a must in case of any application container.
- The registry: the registry contains the configuration data of the operating system and the respective applications. Malware should never access or alter the important data stored in the registry. In spite of the registry being stored in files, access to it occurs by means of a separate interface. Therefore the registry, as a specific component of the OS must be adequately secured.
- Network interface: computer worms often contact remote servers, which coordinate their activities within the system. They try to steal confidential data and infect further computers. Controlling network access is therefore an equally significant factor in these cases.
- System resources: The ability to strictly limit the number of active processes, the usage of operating memory or the processor load may fully eliminate malware execution, or, eventually, minimize the damages and prevent the OS from being inaccessible. Controlled access to these system resources shall thus increase the isolation of processes in the sandbox.

3.1 File Access

The architecture used as a starting point of the proposed solution is the sandbox core created as a layer between the MS Windows operating system and the respective applications. Our concept of supporting virtualization at OS level is based on file virtualization and/or redirecting and constraining file access requests from the virtual machine to the local, root partition of the host OS. E.g. if a process in the V1 virtual machine tries to access a file with the path C:\foo\bar, the virtualization level may redirect the request to the file C:\V1\foo\bar. If a process in another environment (e.g. V2) wants to access C:\foo\bar, the path may be diverted to the file C:\V2\foo\bar, different from file \foo\bar in V1. This mapping is done transparently in the virtualization layer. The process accessing the file C:\foo\bar cannot tell it is in fact accessing a different file.

The targets of the file system access requests may be mapped to the directory containing the file system root, which will contain an equivalent tree structure, or a file container. This container will appear to the host system as a single file. A sandbox knowing the specification of the container may access all elements within the container and control the file access requests for the virtualized process transparently. As to the implementation, the file container may be a simple archive (e.g. a ZIP file), or it may contain a full and separate file system and make use of the existing file formats as VHD, VMDK, etc. specifications [17].

A relevant aspect of file virtualization is whether the application in the virtual machine "sees" the other files and disk volumes in the host environment. There are two different access types in this aspect:

- the process does not know the content of the host file system;
- the process may read arbitrary files in the host system, as part of the specified security policy.

In the first case the root directory changes fully, similarly to the *chroot* tool of Unix systems. This approach has advantages: absolute file isolation and a decreased performance overhead. On the other hand, all libraries and files required for the problem-free execution of the applications in the sandbox must be present in the virtual root directory. This may be a problem in the MS Windows systems, because it is not easy to find out and resolve all DLL dependencies. A further problem is loading duplicates of DLL libraries into the operating memory and wasting disk space. The other solution allows the virtualized process to read arbitrary processes in the host system. The redirected requests result in the following: each file accessed by the process is copied into the redirected directory. However, resource duplication has a cost in terms of performance, it requires space in the primary memory and also increases the overhead related to the initialization and removal of the sandbox. Therefore, the virtual machine shares the majority of its resources with the host and creates private copies of files in the virtualized directory only if it is necessary – upon file write or modification

operations. This approach is known as the copy-on-write mechanism [18]. When using the copy-on-write approach, we have to pay attention to an important fact. Files will be located at two places: in the host environment and in the sandbox environment. When a process running in the VM requests to read the content of a specific directory, the resulting data should contain the unification of the two locations, omitting the duplicate entries in the host directory.

Removing or renaming files initiated by the process running in the sandbox should not affect the host environment. Thus, if the sandboxed application tries to remove/rename a file located beyond the sandbox, the virtualization layer shall not remove the file. The process running in the sandbox shall not know that the file was not really removed. Upon a request of the same process of the specific sandbox to access the removed file, the virtualization layer shall report to the process that the file doesn't exist.

In the proposed solution we copy the file upon opening it with the write flag set. Unlike the copy-on-write approach, the process accessing the existing file is not diverted to a private directory until the first request to open a file with the write flag set appears. This approach has advantages: the virtualization layer overhead is low, which means also a simpler implementation. The disadvantage is wasting storage space and time required to perform a copy when opening the file. The request target does not use a file container but rather a separate directory, the content of which shall mirror the tree structure of the file system.

3.2 Registry Access

Malware, except for modifying files, usually accesses the Windows system registry too [19]. The registry is a hierarchic database containing system information and program settings. In addition to other things, the registry contains the configuration data of the operating system, including the keys required for automatic program execution [20]. Most programs write to the registry at least during installation. Due to all of these reasons, the registry is a critical part of the system, accesses to it must be controlled. Therefore, the sandbox catches the system calls to the registry and modifies the requested keys and values. The principle is the same as with file access, slightly altered, with a lower overhead, when the whole key and value tree is moved to the sandbox [9].

In the proposed solution, registry filtering is based on copying keys on write operations and/or when altering the values. It is analogous to the copy-on-write method. The key or value is created in the sandbox only upon creating the keys and their values or upon their change. The scheme of accessing files is depicted in Figure 2.

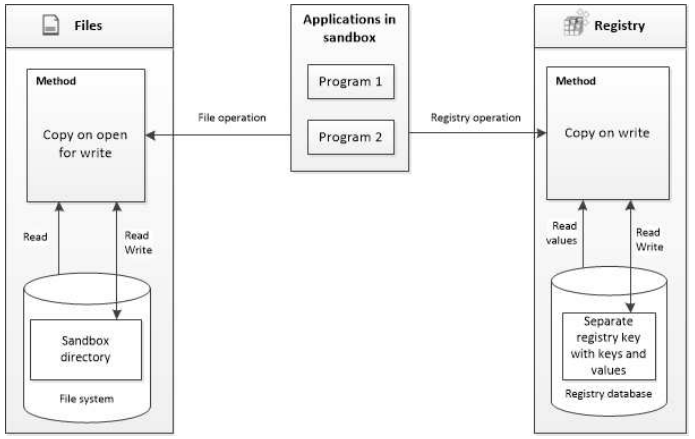


Figure 2
File and registry access

3.3 Network Isolation

Network isolation – in a broader sense – is the creation of an independent environment for an arbitrary process communicating on the local network and the Internet [21]. Any such process gets a unique virtual IP address within the system. The support for this kind of isolation is not widely used in system sandboxes. Most often it appears in solutions providing full virtualization implementations at OS-level, where it is not desirable to have the virtual machines appear under the same IP address and interfere with the network operation of each other. Another approach, emulating a separate network subsystem in the operating system for each process is a very good way to analyze malware behavior. The implementation of such a mechanism is a complex and tedious job, therefore it appears primarily in hardware virtualization, when a separate operating system is running in the virtual machine.

Network isolation – in the more specific sense – means controlling and limiting access during communication on the local network and the Internet. The sandbox should block Internet access and prevent sending out confidential information, especially when the program has access to the host files. The proposed solution blocks network access by means of the firewall integrated into the Windows operating system.

3.4 Controlling System Resources

The control of system resources provides a means of protection from overloading and losing access to the system resources. The sandbox may control the use of the file system, the registry, the operating memory and the CPU load. Depending on the settings it then applies the limitations.

Moreover, in addition to resource management and as part of the privilege system, the sandbox may also contain a policy to modify the system settings (icons, fonts, screen settings, power management, etc.), logging out, system restart and shutdown. The sandbox contains an implementation of controlling resources and system settings by means of Windows Job objects. This technology allows the individual processes to be assigned to the respective containers and apply the requested properties. Job objects allow the processes to be handled as uniform units. The set of constraints may be specified within a single object, which forces the use of the constraint on each process in the process file. An overview of the methods used in network isolation and system resource and setting control is depicted in Figure 3.

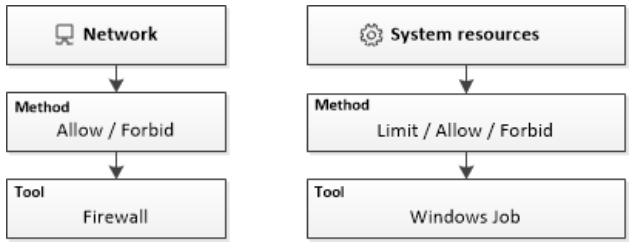


Figure 3
Network isolation, system resource and setting control

4 Sandbox Architecture

The proposed sandbox consists of three independent parts. The core of the sandbox is executed at the OS kernel level. The main module consists of a software driver. It filters file, registry and process operations. A Windows service loads the driver into the system and implements network isolation. The communication with the user occurs using a third, separate program. The user controls the sandbox by means of a graphical interface, which communicates with the driver and the service. The basic architecture of the sandbox is depicted in Figure 4.

The article contains a detailed description of the sandbox driver architecture. This component is the core of the sandbox.

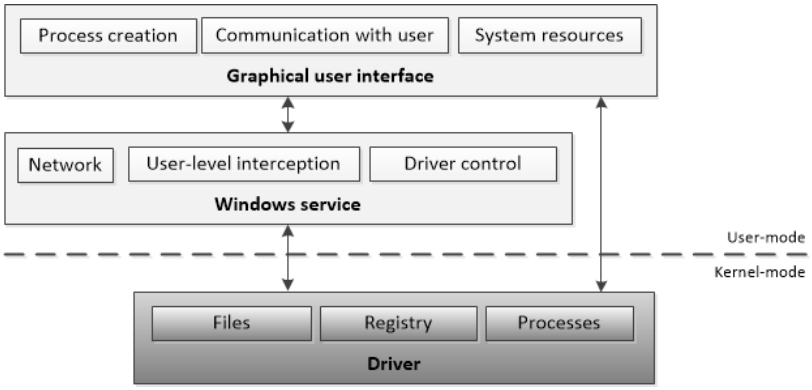


Figure 4
Basic sandbox architecture

4.1 Minifilters

With the advent of Windows XP SP2, Microsoft created a model to ease the creation of file system filters as an alternative to the previously used legacy filter model. In this the drivers – minifilters – are managed by the filter manager (FM). Most system calls issued by the applications, sent to the drivers are in I/O request packet (IRP) format. In the minifilters, the IRP requests are first processed by the FM. The minifilter registers the callbacks of the requested operations, which are then called by the FM. The FM encapsulates the IRP into a structure called CallbackData and sends the structure gradually to the individual minifilter instances, just as it was in the case of the legacy filters. A minifilter instance is an abstraction of the minifilter at the specific disk partition.

An important part of the filter architecture is context support. The context is a structure defined by the minifilter to store arbitrary data associated with the filter manager objects. The context may be associated to minifilter instances, files, file objects, open file streams, disk partitions and transactions.

4.2 Files

The elementary mechanism providing the file name modification is based on setting the reparse status flag (STATUS_REPARSE) of the corresponding IRP operation in the pre-operation. This indicates that the I/O manager discards the current IRP and initiates a new IRP with the file object containing the file location in the sandbox domain. In addition to altering the path and the state of the IRP operation the kernel must finish the current operation to make sure that the IRP structure does not access the file system. This is achieved by returning the appropriate state from the pre-operation to the filter manager. An important problem arises with this redirection operation: How does the filter know, which

request has been redirected and which has not. With the advent of Windows Vista, each creation operation may contain a reference to an ECP (Extra Create Parameter) context. This context is retained even after finishing the IRP, when the I/O manager adds a reference to the context to the new IRP structure. The kernel uses the ECP context to mark the redirected operation in the pre-operation phase and to mark the creation operation initiated by the sandbox in the redirection between disk partitions in the whole driver. The ECP context is only available in the creation pre-operation and post-operation phases. For a simple identification of the request and further required data, the ECP context is transformed to a context associated with an open file stream (SHC) in the creation post-operation phase, where the file object already exists (unlike the pre-operation phase), therefore this context may be associated with it. The kernel associates the SHC only with the file object representing a file in the sandbox and it is available only to our minifilter, unlike the ECP context available for all drivers taking part in the creation operation.

4.2.1 Opening Files

When opening/creating a file, each application tells the OS how to behave if the file exists/does not exist. Windows knows six different dispositions, as stated in Table 1.

Table 1
File opening dispositions

Disposition	**The file exists**	**The file doesn't exist**
CREATE	failure	created
OPEN_IF	open	created
OPEN	open	failure
OVERWRITE	open and overwritten	failure
SUPERSEDE	replaced	created
OVERWRITE_IF	open and overwritten	created

The program also specifies the requested file access: read, write, erasable, etc.. The disposition and access specifications are the main elements in selecting when the kernel should redirect the request, copy the file to the sandbox domain or ignore the request. The situations, which may arise in conjunction with the specified disposition and the selected flag are depicted in Figure 5.

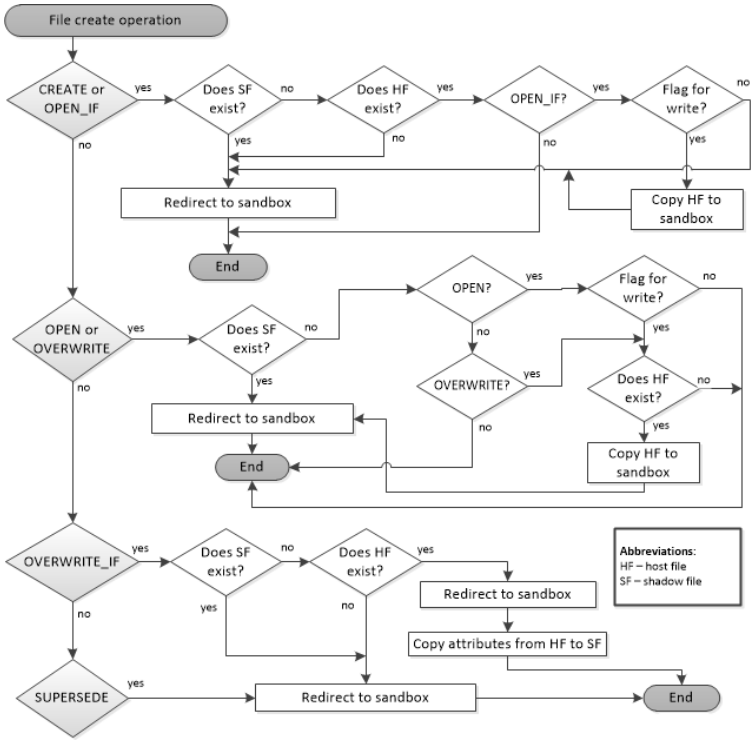


Figure 5
Request flow upon opening a file

4.2.2 Removed Files

The sandbox core registers the information setting operation (IRP_MJ_SET_INFORMATION) and selects the disposition structure. Then, the whole path to the file is stored in the table of removed files. If the file comes from the sandbox, further execution is left over to the I/O manager. If the file comes from the host environment, the core initiates its own information setting operation with the file removal disposition flag unset. Depending on the result of the initiated operation, the kernel will apply the final state and finish the original information setting operation.

The kernel manages a separate table with the list of removed files for each logical sandbox. To keep the state of the sandbox even after operating system shutdown and restart, the table contents are stored in a file and read into the memory upon the creation of the logical sandbox.

The table implementation influenced the decision process in the redirection specified above. If the file is being opened, the first step is to check the existence of the file path in the table. If the file is in the table, the request is denied, the operation is finished. Otherwise, the above process is executed.

4.2.3 Renamed Files and Hard Links

The sandbox controls file renaming and hard link creation. The kernel is primarily interested in the first three phases of these operations. Opening a file to rename; opening the target file with the new name; and catching the IRP called IRP_MJ_SET_INFORMATION associated with the file object of the first operation. The first phase is executed in the sandbox kernel in accordance with the file opening algorithm. The second phase is a little different from the standard opening algorithm, because the target file does not exist. Actually, the request contains the flag indicating that the calling process tries to open the target of the renaming operation or to create a hard link. The sandbox kernel receives the flag and forwards the target to the sandbox domain without further analyzing the request. The third phase is different, depending on whether the operation is aimed at renaming a file or creating a hard link.

When renaming a file, the filter retrieves source file information in the third phase. If the source file comes from the sandbox domain, the kernel is sure that the host cannot be modified and the rest of the operation is left over the I/O manager. If the source file is with the host, the kernel retrieves information of the target file, copies the source file from the host to the sandbox domain under a new name, adds the source file to the table of removed files, stops further execution and indicates success. Renaming within disk partitions is not possible due to the principle of file system independence; such operations are performed internally as copy and deletion operations.

4.2.4 Directory Enumeration

The result of the enumeration is a set of files and directories from the queried directory. Since the proposed sandbox uses the copy-on-write technique, the files may be at two locations. If the isolated process requests information on the items residing in a specific directory, the result must contain a unification of both locations, omitting the duplicate entries and the entries listed in the table of deleted files. Duplicate entries represent files and directories located both in the sandbox domain and the host domain. The result shall not contain the entries from the host domain in these cases.

4.3 Registry

The registry filter is a driver filtering the registry calls. The configuration manager (CM) implementing the registry allows filtering any process call related to the registers. Similarly to files, the register filter receives pre-notifications and post-notifications. The driver catches all notifications. If the given notification is irrelevant to the manager, it must at least return the execution to the configuration manager. If the operation is filtered during the pre-notification phase, the requested processing may be done and one of the three status types is returned:

- *success (STATUS_SUCCESS) – the CM processes the registry operation and then initiates a post-notification containing the required data and the resulting status of the operation;*
- *failure – the operation is rejected and the process error status is returned immediately, without calling the post notification and the assistance of the CM;*
- *bypass (STATUS_CALLBACK_BYPASS) – all processing must be done by the driver and the registry operation terminates with success and without initiating post-notification and CM assistance.*

As with minifilters, registry filters have contexts too. The context may be associated with registry filters, registry notifications and registry objects representing open keys. Figure 6 shows the principle of filtering pre-notifications in the registry filter, including specific notification types.

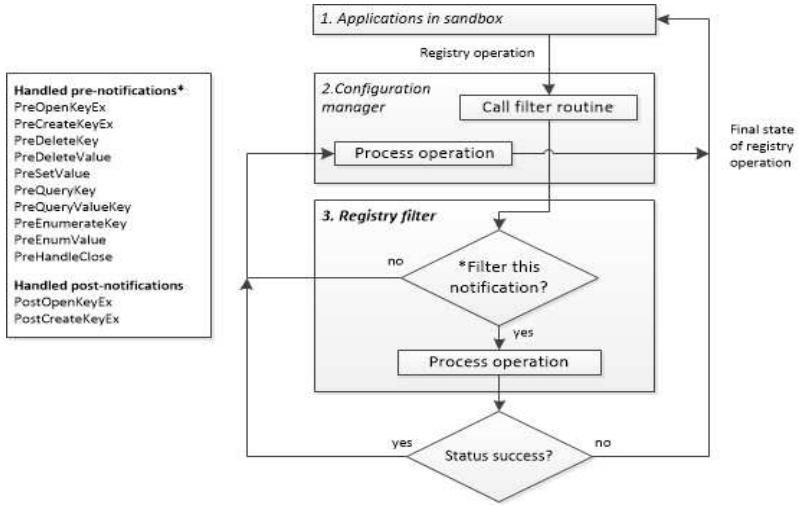


Figure 6
The principle of filtering pre-notifications of registers in the sandbox

4.4 Processes

The driver, the minifilter, the registry filter or other components catch all system calls by default. The aim of the solution is to filter specific programs. The sandbox kernel must determine the context of the currently executed thread for all callbacks and associate them with the processes. Each process in the operating system has a unique numerical identifier, the PID. Windows allows each driver to register a callback initiated upon process creation and termination. The client part of process execution sends the PID to the sandbox, which stores it in the global process table. The callback contains not only the PID of the process being created,

but also the PID of the parent process, which created the process. If the isolated process creates another process, the sandbox core identifies the new process using the parent PID and adds it to the process table.

5 Evaluation of the Solution

To evaluate the solution, we have performed a functionality test of the proposed secure sandbox. For creating snapshots of the file system and the registry database we used System Explorer and OS Forensics software. The sandbox was tested in a virtual machine with an updated 32-bit Windows 7 OS without antivirus software. The tests could be performed on the MS Windows XP and MS Windows Vista too. We have purposefully downloaded and executed 11 malicious applications in the proposed sandbox using Internet Explorer. We have selected malware according to its current popularity. According to the statistics of the company G Data Software AG, the most widespread are various forms of adware. Trojans, spyware and worms are the largest threats, therefore we have included them in the test, in spite of the fact that these are less widespread globally.

We have tested the following threats (the names come from the virus database of the company ESET):

1. ZeroAccess (called Trojan-Dropper.Win32.Smiscer.hf in the Kaspersky database) is one of the most widespread and most dangerous rootkits. This malware is part of various key generators and cracks. When executed in a 32-bit OS, ZeroAccess overwrites the existing driver and loads itself into kernel space. The rootkit usually hides in a hidden partition. In a 64-bit OS, ZeroAccess does not contain code executed in kernel space. In this case it hides in the Global Assembly Cache (GAC) of the .NET framework.

Result: The rootkit could not be executed in the proposed sandbox. The system was not infected.

2. Win32/Injector.AAKO – carrier of various kinds of malware, trojans catching passwords and other confidential information from the infected computer. This specific variant named AAKO does also copy the payload to the C:\Windows\InstallDir directory and creates a value in the registry database, in the HKLM\Software\Microsoft \Windows\CurrentVersion\Run\HKLM key.

Result: The program was executed in the sandbox. The registry and file sandbox caught all executed changes. After terminating the sandbox, all malicious processes were terminated. The system was not infected.

3. Win32/TrojanDropper.VB.OJG – changes the Internet connection settings in the registry and attempts to switch off the firewall. Further, it also copies an executable with a random name to the TEMP directory, executes it and writes malicious data in the GDIPFONTCACHEV1.DAT file, normally used as a font cache. In addition to this, it copies an svchost.exe executable file into the C:\ProgramData directory, pretending to be a service of the operating system.

Result: The sandbox successfully caught all executed changes and processes, without infecting the OS.

4. MSIL/Injector.CUXtrojan – behaves similarly to the previous program, but pretends to be a Java update in the registry.

Result: The sandbox successfully caught the executed changes and processes.

5. Adware.Relevant.CC – currently the most widespread adware currently, part of many freeware programs. It analyzes user activity on the computer (including Internet surfing), it adds an exception to the firewall and from time to time it opens a window to fill in the user data.

Result: unsuccessful installation in the sandbox.

6. Win32/Spy.Zbot.YW – a trojan stealing passwords and other confidential information. It also serves as a backdoor, it may be controlled remotely.

Result: This trojan was not execute in the sandbox. Upon start, the program terminated with an error, so the system was not infected.

7. Win32/Kryptik.BFCO – Ransomware, encrypting the file system and offering unencryption for money. In addition to this, it switches off Windows Firewall, Windows updates, Windows Defender and other antivirus software. It is automatically started by means of registry values and it was copies itself to various places.

Result: This malware was successfully executed and started encrypting the files. However, the original files remained untouched; all changes remained limited in the sandbox. The sandbox contained a number of encrypted files. The host system (files and the registry) remained uninfected and the OS settings remained unchanged. During the test it was not possible to execute the task manager and terminate the processes manually. After terminating the sandbox, the file encryption processes were terminated and the task manager could be switched on again.

8. Win32/Spy.Hesperbot – the known banking trojan spreading by email as an attachment named zasilka.pdf.exe, most active in Turkey and the Czech Republic. Win32/Spy.Hesperbot records keystrokes, creates screenshots, records video using the web camera of the computer and creates a remote proxy.

Result: This malware started successfully, it stored system information in an encrypted file and copied its code into the newly created attrib.exe and explorer.exe files, then terminated. The active network isolation prevented it from communicating with the server. The host system was not modified. After terminating the sandbox, both infected processes were terminated.

9. Win32/Dorkbot.B – a very widespread computer worm spreading through Skype, Facebook Chat, Twitter messages, etc. This worm steals passwords and contains a backdoor.

Result: None of the tested variants of this worm could be started in the sandbox.

10. Win32/LockScreen.ALY – another example of ransomware.

Result: Without administrator privileges it was not possible to execute the malware in the sandbox. With administrator privileges and active limitations the sandbox caught the changes.

The chart in Figure 7 shows the results. The chart shows a total of 11 experiments, since the behavior of malware no. 10 was significantly dependent on the method of testing.

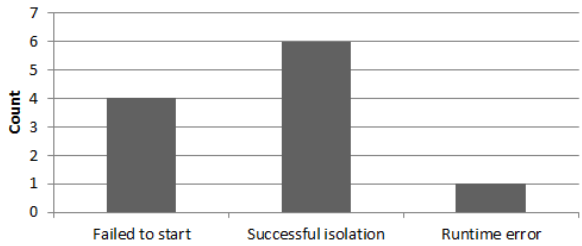


Figure 7
Security test results

Malicious code did not manage to infect the host system in any of the cases. Some code did not even start in the sandbox. If the malware started, in some cases it was capable of performing its malicious activities; however, after terminating the sandbox, it was terminated successfully. Network access constraints and mainly the restrictions in the number of active processes and system settings play a significant role in sandbox efficiency.

To evaluate the additional computing costs of the solution, we have measured both the initial and repeated program start-ups. We have tested the start-up of the Firefox 28 Internet browser, WinRAR 5 compression software, Foxit Reader 6 PDF reader and Emule 0.5 file transfer software. When measuring the initial start-up, we have restarted the OS after each attempt. The measurement results are in Figure 8. The average initial start-up overhead was 8%, subsequent start-up overhead amounted to 17%. The smaller percentage of the initial startup is related to the increased time of loading the libraries from the hard drive.

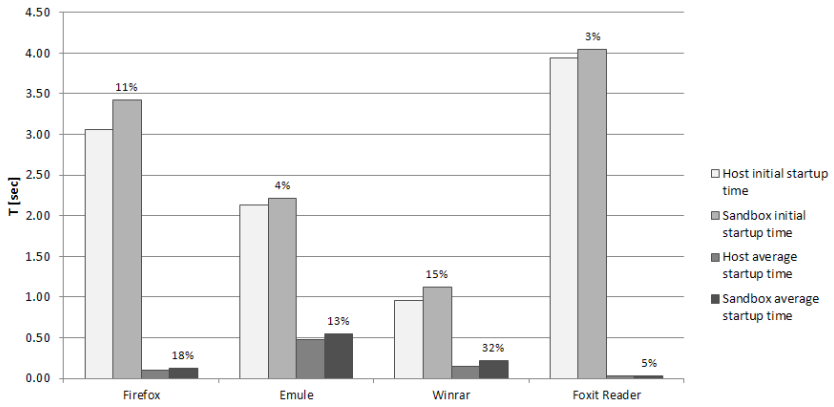


Figure 8

Initial and subsequent startup times of the programs in the sandbox

Standard usage of programs degrades the execution time of programs minimally. Increased time costs are presented in case of programs which are heavy on I/O operations. All other system calls are executed natively, which is an advantage in comparison with hardware-level virtual machines. The whole performance penalty depends on the application in use; it may range from 5% to 20%.

Conclusion

The goal of this work was to design a system sandbox in the MS Windows system, providing security in case of the executed applications. The proposed architecture was implemented using standard methods offered by the Windows operating system. The behavior of the sandbox was primarily based on the copy-on-write model, which creates a redundant isolated OS environment. This article presents the design of the sandbox and the main parts of the implementation in the MS Windows OS. The execution of the respective applications is not functionally limited in the sandbox. The applied copy-on-write approach at the level of OS virtualization has a smaller performance penalty than full virtualization; therefore the applications are executed without greater latency in the sandbox. The performed security test showed the functionality of the proposed sandbox model in solving current security problems, where the user is forced to react to various security issues. A drawback of the solution is the possible waste of storage space. In future, the virtualization of inter-process communication should be improved and the network isolation with deeper integration.

Acknowledgements

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-0008-10 and project KEGA 008TUKE-4/2013: Microlearning environment for education of information security specialists.

References

[1] Hoopes John: Virtualization for Security including Sandboxing, Disaster Recovery, High Availability, Forensic Analysis, and Honeypotting, Burlington 2009

[2] Wahbe Robert et al: Efficient Software-based Fault Isolation, ACM SIGOPS Operating Systems Review, Vol. 27, No. 5, New York 1993, pp. 203-216

[3] Smith James Edward, Nair Ravi: Virtual Machines: Versatile Platforms for Systems and Processes (The Morgan Kaufmann Series in Computer Architecture and Design), San Francisco 2005

[4] Barrett Diane, Kipper Gregory: Virtualization and Forensics: A Digital Forensic Investigator's Guide to Virtual Environments, New York 2010

[5] Moya del Barrio Victor: Study of the Techniques for Emulation Programming [online], Barcelona School of Informatics, Barcelona 2001, Available on: http://personals.ac.upc.edu/vmoya/docs/emuprog.pdf [quoted on April 12, 2013]

[6] Calzolari Federico: High Availability using Virtualization: Doctor's thesis, University of Pisa, Pisa 2009, p. 83

[7] Nakajima Jun, Mallick Asit: Hybrid-Virtualization – Enhanced Virtualization for Linux, Proceedings of the Ottawa Linux Symposium, Vol. 2, Ottawa 2007, pp. 87-96

[8] Chaudhary Vipin et al: A Comparison of Virtualization Technologies for HPC, Proceedings of 22nd International Conference on Advanced Information Networking and Applications, IEEE Computer Society, Buffalo 2008, pp. 861-868

[9] Yu Yang: Os-Level Virtualization and its Applications: Doctor's thesis, State University of New York at Stony Brook, Stony Brook 2007, p. 120

[10] Shan Zhiyong et al: Facilitating Inter-Application Interactions for OS-Level Virtualization, ACM SIGPLAN Notices – VEE, Vol. 47, No. 7, New York 2012, pp. 75-86

[11] Laadan Oren, Nieh Jason: Operating System Virtualization: Practice and Experience, ACM Proceedings of the 3rd Annual Haifa Experimental Systems Conference, New York 2010, pp. 1-12

[12] Mihályi Daniel, Novitzká Valerie: Towards the Knowledge in Coalgebraic Model of IDS, Computing and Informatics, Vol. 33, No. 1, Bratislava 2014, pp. 61-78

[13] Kampert Paulus: Taxonomy of Virtualization Technologies: Thesis, Delft University of Technology, Faculty of Systems Engineering Policy Analysis & Management, Delft 2010, p. 93

[14] Jakubčo Peter, Šimoňák Slavomír, Ádám Norbert: Communication Model of emuStudio Emulation Platform, Acta Universitatis Sapientiae: Informatica, Vol. 2, No. 2, Cluj-Napoca 2010, pp. 117-134

[15] Kamp Poul-Henning, Watson Robert: Jails: Confining the Omnipotent Root, Proceedings of the 2nd International SANE Conference, Maastricht 2000, pp. 116-127

[16] Ma Kun, Yang Bo, Abraham Ajith: A Template-based Model Transformation Approach for deriving multi-tenant SaaS applications, Acta Polytechnica Hungarica, Vol. 9, No. 2, Budapest 2012, pp. 25-41

[17] Singh Jasmet. et al: An Application Sandbox Model based on Partial Virtualization of Hard-Disk and a Possible Windows Implementation, International Journal of Computer Applications, Vol. 7, No. 57, New York 2012, pp. 16-21

[18] Kasampalis Sakis: Copy On Write Based File Systems Performance Analysis And Implementation: Thesis, Technical University of Denmark, Department of Informatics, Lyngby 2010, p. 83

[19] Apap Frank et al: Detecting Malicious Software by Monitoring Anomalous Windows Registry Accesses, Recent Advances in Intrusion Detection: Lecture Notes in Computer Science, Berlin 2002, pp. 36-53

[20] Honeycutt Jerry: Microsoft Windows Registry Guide, Redmond 2005, p. 1327

[21] Vokorokos Liberios, Pekár Adrián, Ádám Norbert, Darányi, Peter: Yet Another Attempt in User Authentication, Acta Polytechnica Hungarica, Vol. 10, No. 3, Budapest 2013, pp. 37-50

[22] Vokorokos Liberios, Baláž Anton, Trelová Jana: Distributed Intrusion Detection System using Self Organizing Map, Proceedings of 16th IEEE International Conference on Intelligent Engineering Systems, Lisbon 2012, pp. 131-134

A Content-based Image Retrieval System Based On Convex Hull Geometry

Santhosh P. Mathew[1], Valentina E. Balas[2], Zachariah K. P.[1]

[1]Department of Computer Science, Saintgits College of Engineering, Kerala, India; E-mail: santhosh.mathew@saintgits.org, zacharia.kp@saintgits.org

[2]Department of Automatics and Applied Software, Aurel Vlaicu University of Arad, Romania; E-mail: valentina.balas@uav.ro

Abstract: Developments in data storage technologies and image acquisition methods have led to the assemblage of large data banks. Management of these large chunks of data in an efficient manner is a challenge. Content-based Image Retrieval (CBIR) has emerged as a solution to tackle this problem. CBIR extracts images that match the query image from large image databases, based on the content. In this paper, a novel approach of comparing the convex hull geometry of the query image to that of the database image in terms of a relative metric which is denoted as the Convex Hull Area Ratio (CHAR) is used. The metric CHAR is the ratio of the area of the intersection of the two convex hulls to the area of their union. Convex hull shape polygon is extracted from the database images and the coordinate values are stored in the feature library. When a query image is given, the convex hull values are extracted in the same fashion. Ratio of the intersected area to union area of the two convex hulls (CHAR) are found and stored in an array. Subsequently, similarity measurement is performed and the maximum value of the CHAR indicates the closest match. Thus, the database images that are relevant to the given query image are retrieved. Scale and translational invariance have been preserved by a suitable co-ordinate transformation. The proposed CBIR technique is evaluated by querying different images and the retrieval efficiency is evaluated by determining precision-recall values for the retrieval results.

Keywords: Image Retrieval; Shape Signature; Image Segmentation; Edge detection; Convex hull; Area ratio

1 Introduction

Content-based Image Retrieval system is an information processing system that makes use of the image content in the retrieval process. Design and development of an automated retrieval system is quite difficult, considering the complex objects and information present in the image. An image retrieval system can be defined as a computer system for browsing, searching and retrieving images from a

large database of digital images. Some of the most important characteristics that are used to extract information from the images are Color, Shape and Texture.

Color histograms are commonly used in Content-based image retrieval. Though Color and texture contain important information, it is possible that two images with similar color histograms may actually represent very different images. Hence, shape describing features play a vital role in effectively retrieving images by making use of a CBIR system. Though much research is going on in shape based representation and retrieval techniques, there is no direct solution as to which kind of shape features are to be considered for best performance [18].

In this paper, a novel approach of comparing the two convex hulls of the prominent edges in the query image to that of the database image by computing the Convex Hull Area Ratio (CHAR) is made use of. The proposed CBIR system effectively retrieves the images relevant to the query image when compared to the other CBIR system. The remaining part of the paper is organized as follows: Section 2 discusses some of the related works. Section 3 briefs the Content-based image retrieval with proposed shape signature extraction, feature calculation and image retrieval process. Results and analysis of the proposed technique are discussed in Section 4. Finally, concluding remarks are provided at the end.

2 Background of the Research Work

Tools that effectively browse, search and retrieve images are needed by users from different domains. Remote sensing, Fashion, Crime prevention, Publishing, Medicine, Architecture, etc. are some of them. Image retrieval has been in the thick of research for the past many decades. Database management and Computer vision are the two major research groups who have been contributing to this field. These groups study image retrieval from different perspectives. Text based approach is adopted by the former, while the latter follows content-based approach. High level features like keywords and text descriptors are normally used by humans to interpret images and measure their similarity. But the low level features automatically extracted using computer vision methods are normally of low level in nature [1] [2]. Early techniques of image retrieval were based on the manual textual annotation, a difficult and laborious task. Texts alone are not sufficient because the interpretation of what we see is hard to characterize by mere words. This led to the contents in an image, like color, shape, and texture, gaining more importance and in turn in the birth of CBIR [3].

Shape is considered to be one of the most important low level image features in content-based image retrieval. Region-based and Contour-based are the two common approaches used by shape-based systems [4]. Region-based systems normally use Geometrical moments, Zernike moments, Legendre moments and such moment descriptors [5] [6]. Contour-based systems use the boundary of the

objects. We usually get better results for images that are distinguishable by their contours. Literature discusses various shape representation and retrieval techniques. FD (Fourier Descriptors), PA (Polygonal Approximation), IM (Invariant Moments), CSS (Curvature Scale Space) etc. are some of them [7]. Shape is a major component used for describing digital image along with other features such as color and texture. Contour-based shape descriptors, such as Fourier descriptors, are not appropriate for describing shapes consisting of several disjoint regions. Region based shape descriptors, such as moments, are used when shapes have complex boundaries [8].

Experiments on CBIR systems show that, on many occasions, low level contents fail to describe the high level semantic concepts. Ying Liu *et al*. [9] provided a comprehensive survey of the recent technical achievements in high-level semantic based image retrieval. Their research covered different aspects, including low level image feature extraction, similarity measurement, and deriving high level semantic features.

Manual textual annotation of images was made use for the earlier image retrieval systems. This was a very difficult and time consuming task. It is very difficult to describe what we see and interpret purely using text. This led to the contents in an image like color, shape, and texture gaining more importance. Amit Jain *et al*. [3] proposed an algorithm for retrieving images with respect to a database consisting of engineering/computer aided design (CAD) models. Retrieval of objects has been done using a similarity measure that combines shape and the depth information. They combined the shape obtained from the contour with the 3D embedding of the depth information at each point on the contour, to identify a feature set. Trademark Image Retrieval (TIR) was developed to check and prevent the repetition of the large number of trade mark images that are stored in the trademark registration system. Chia -Hung Wei *et al*. [6] derived a Content-based trademark retrieval system with a set of feature descriptors that could represent global shapes and local features of the trademarks.

Many retrieval systems use norm based distances on the extracted feature set as a similarity function. H. B. Kekre et al. [10] suggested that complex Walsh transform sectors of the images could be used to generate the feature vectors for the purpose of image search and retrieval. Xiang-Yang Wang *et al*. [7] proposed a color image retrieval scheme by combining color, texture and shape information. Robustness with respect to the image scale, Illumination and Noise are some of the important aspects to be considered in developing image matching systems. Shao-Hu Peng *et al*. [11] proposed a visual shape descriptor based on the sectors and shape context of contour lines.

From the discussion so far, it is quite clear that shape is one of the primary visual features in CBIR. Our work focuses on developing a CBIR system based on shape. Shape descriptors are normally classified as contour based and region based. Contour-based shape descriptors primarily use the boundary information

and neglect the local contents within the shape. But region-based shape descriptors make use of the contents within the shape [20]. Works on CBIR systems based on moment descriptors are in progress [12]. Fuzzy techniques are also being employed in retrieval and recognition [21].

The Precision and Recall (PR) levels of the existing CBIR systems based on Shape have not been very good, when minimal feature sets were used. Extensive feature sets have had a negative impact on the retrieval time. Our work aims at identifying a minimal feature set with high PR levels. A computationally efficient algorithm for getting a first level match is of utmost importance. Computationally expensive refinement can then be used, if required, on this trimmed dataset to get better matches. Considering this situation, we have first used the color and texture properties to extract the objects from the image through segmentation. The shape is extracted from the edges identified using Canny algorithm [13] and from the convex hull containing the prominent edges in the image. A novel similarity measure based on convex hull geometry (Convex Hull Area Ratio – CHAR) is proposed, to retrieve the exact match.

3 Content-based Image Retrieval Based on Convex Hull Geometry

Though Convex hull is a good shape representation and description technique, it has not been effectively utilized in CBIR. There has only been very few applications of Convex hull in image retrieval. Lancaster et al [16], used the Convex hull of the brain image in a medical application [17]. The Convex hull derived from the brain's surface was identified as the basis for automating and standardizing global spatial normalization. Their idea was to match the position, orientation and dimensions of the brain to that of a standard atlas brain.

The proposed Content-based Image Retrieval system comprises of the pre-processing steps of image segmentation, smoothing, feature extraction and image retrieval based on the query image [19]. The proposed CBIR is based on the shape signature extracted by employing color based segmentation, edge detection and extraction of the convex hull. When an image is queried, the system determines the shape signature for the image and then computes the similarity measure between the signatures of the query image and the existing database based on the convex hull area ratio (CHAR) comparison and retrieves a specified number of the best matches. The proposed method is detailed below.

In our work, we make use of the K-means clustering algorithm for image segmentation for further processing with k = 4. K-means clustering identifies partitions in such a way that objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. The number of

clusters to be partitioned and a distance metric for quantifying the distance between two objects are necessary for K-means clustering. Prior to the application of K-means clustering, the image, which is in the form of a 2D matrix, is rescaled to a 1D vector. Subsequently, the K- means algorithm is applied, to cluster the image, with k set as 4. After the K- Means Algorithm Is Applied, The 1D Vector Is converted back to 2D matrix and then the Canny algorithm is used for the detection of edges present in all the clustered sets of the image. The Canny edge detection operator detects a good range of edges in images. The canny algorithm consists of five steps, which are smoothing, finding gradients, non-maximum suppression, double thresholding and edge tracking by hysteresis [13] [14].

Canny edge detection process gives the different edges that are present in the image and then the indices of the shaped content are extracted by employing convex hull method. The convex hull of a set of points is the smallest bounding convex polygon that will contain the set. The convex hull is used for extracting the shape of the image. The approach is that the shape is represented by a single convex hull. The convex hull H of a region is its smallest convex region including it. In order to decrease the effect of noise, we first smooth a boundary prior to partitioning. The representation of the shape is obtained by storing the coordinate values of the convex hull. In a similar manner, the features of the rest of the images in the database are also extracted and stored in a feature library. For preserving scale invariance the convex hull extracted is normalized to unit area by performing a scaling of the co-ordinate system before storing for further analysis. Similarly for preserving translational invariance the co-ordinate system origin is fixed at the centroid of the extracted convex hull.

3.1 Retrieval Process

The Image retrieval process takes place after the Image Indexing process. The query image is matched to the images in the DB (database) for image retrieval. Let Q be the query image. We perform the shape extraction for the query image Q and extract the convex hull of the detected shapes. The Q polygon and DB Polygon (of convex hull) are compared by finding ratio of intersected area to union area (CHAR) as explained in Fig. 1. The CHAR values are computed for each image in the DB and stored in an array. The maximum value of this ratio is the closest match. Accordingly, a specified number of images with best match are retrieved from the database. The retrieval process is further explained in detail using Figure 1.

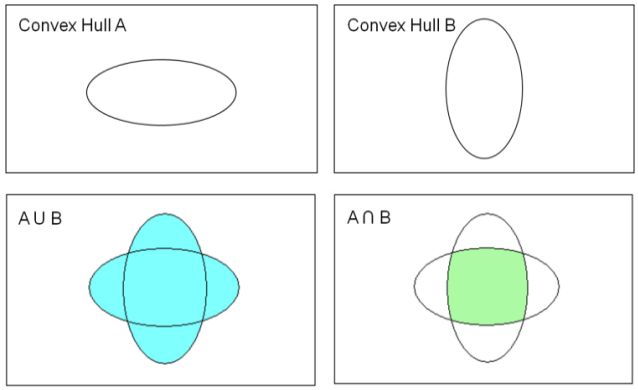


Figure 1
CHAR Computation (i) Convex hull A (ii) Convex Hull B
(iii) Area of union (blue) (iv) Area of intersection (green)

As shown in Figure 1, CHAR is maximum at a value of 1 when the query and database convex hulls are exactly identical. When there is no overlap between the two convex hulls the CHAR value reduces to a minimum value of 0. The CHAR is a measure of the overlap among the query and retrieved convex hulls.

The convex hull can be obtained by HMT (Hit or Miss Transform) with four structuring elements. Consider A to be a binary image set. Let B^i, i = 1, 2, 3, 4, represent four structuring elements as given below in Figure 2.

$$B^1 = \begin{bmatrix} 1 & \times & \times \\ 1 & 0 & \times \\ 1 & \times & \times \end{bmatrix}, \quad B^2 = \begin{bmatrix} 1 & 1 & 1 \\ \times & 0 & \times \\ \times & \times & \times \end{bmatrix}, \quad B^3 = \begin{bmatrix} \times & \times & 1 \\ \times & 0 & 1 \\ \times & \times & 1 \end{bmatrix}, \quad B^4 = \begin{bmatrix} \times & \times & \times \\ \times & 0 & \times \\ 1 & 1 & 1 \end{bmatrix}$$

Figure 2
Structuring elements

where the symbol $\times$ denotes don't care. B^i is a structuring element where i vary from 1 to 4. These structuring elements are called B^1, B2, B^3 and B^4.

Let the initial step be $X^i_0 = A$. The iterative process is conducted as in the kth step

$$X^i_k = (X \circledast B^i) \cup A, \; i = 1,2,3,4 \text{ and } k = 1,2,3,\ldots \quad (1)$$

If $X^i_k = X^i_{k-1}$, then it converges. That is, in two subsequent iterations, if the output does not change, the algorithm converges.

Let $D^i = X^i_{conv.}$

Finally, the convex hull of *A* is

$$C(A) = \bigcup_{i=1}^{4} D^i \tag{2}$$

Since there are four different structuring elements, four different point sets are obtained at the end of convergence. The union of these four different point sets gives the convex hull of the given image A.

Similarly, we find the convex hull of the image B. C(A)∪C(B) and C(A)∩C(B) are further calculated.

The metric CHAR is given by

$$C(A) \cap C(B) \,/\, C(A) \cup C(B) \tag{3}$$

4 Results and Discussion

This CBIR system has been implemented using MATLAB (version 7.10), has been tested with various query images and appropriate matching images have been recovered from the image database. Test images were taken from the database generated by Wang containing many images stored in the JPEG format [15]. When an image is queried, the system establishes a shape feature for the image and then computes the similarity measure between the shape feature of the query image and the shape features of all images existing in the database, so that, a specified number of database images similar to the query image are retrieved.

To retrieve the image from the image database, the query image is preprocessed to normalize the intensity levels of the input image. The output objects extracted from the input query image after the preprocessing stage is given in Row 2 of Figure 3. To extract the shape feature from the image, initially, the image in RGB color space is converted to gray scale image. Since, the mean filter can act on only one color channel. The mean filter is especially useful for reducing speckle noise and salt and pepper noise. The image is converted into gray scale using Craig's formula. After converting the RGB image into gray scale image, K-means clustering algorithm (k = 4) is used for the image segmentation for the further process. Before applying the K-means clustering, the image, which is in the form of 2D vector, is rescaled to a 1D. In this K means clustering process the image pixels are grouped under the color. The output image of the clustering process is shown in the Row 3 of Figure 3. After the clustering process using K-Means algorithm, the 1D image is again converted into 2D image format. Then canny algorithm is used for the detection of different edges present in all the clustered

sets of the image. The canny algorithm comprises of five steps, they are smoothing, finding gradients, non-maximum suppression, double thresholding and edge tracking by hysteresis. The image after detecting the edge is shown in the Row 4 of Figure 3.

After applying canny algorithm to determine the edges of the segmented image, the edges are tracked and then smoothed to remove the number of connected components. Then we get the prominent shapes that are present in the image and then the x, y values of the extracted convex hull polygon of the detected edges are stored.

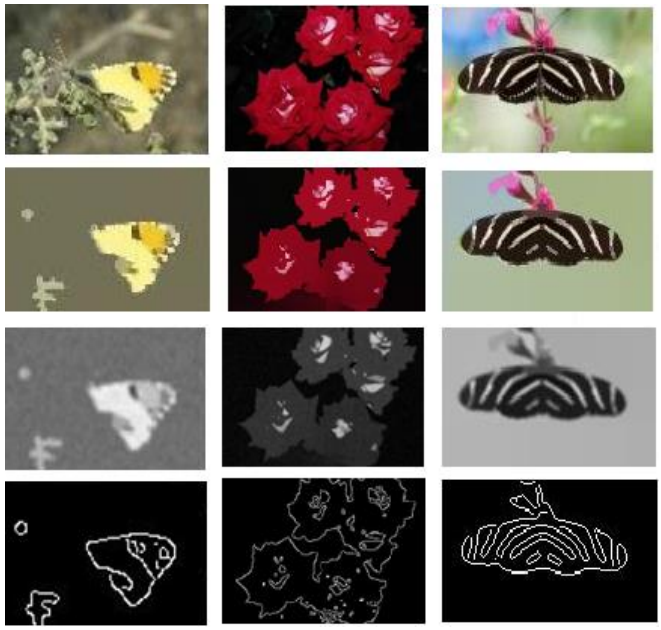


Figure 3
Row 1 – Original images, Row 2 – Objects extracted from the original images,
Row 3 – Clustered gray scale images, Row 4 – Edges detected by Canny algorithm

Retrieval is performed based on the proposed Convex Hull Area Ratio (CHAR) method, as explained by Figure 1 and Figure 4.

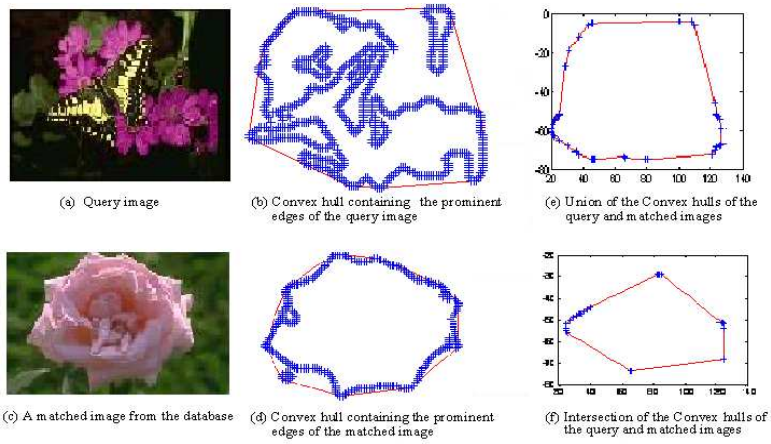


Figure 4
Union and Intersection areas of the query image and a matched image. Ratio of Intersection to Union will be close to 1 for a good match and low for dissimilar ones.

During the computation of the convex hull, the scale invariance is preserved by normalizing the area of each convex hull to unity. This assures that only the relative spread of the shape features are compared and similar shapes give a better CHAR. Translational invariance is preserved by setting the origin to the centroid of the convex hull polygon.

Retrieved images that correspond to the input query image are shown in Figure 5 and Figure 6.

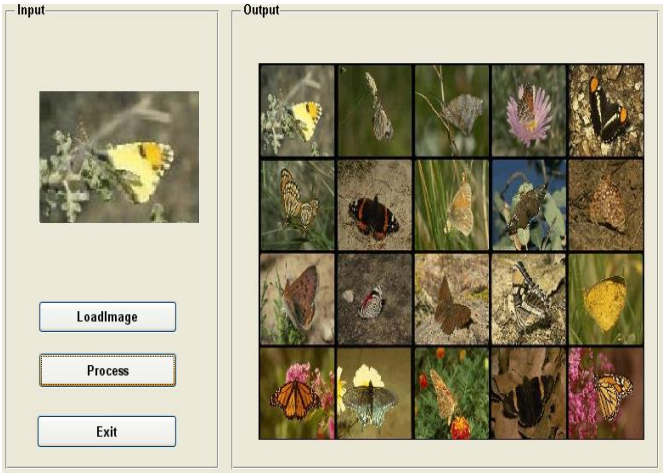


Figure 5
Sample Output – Query Image and Retrieved images

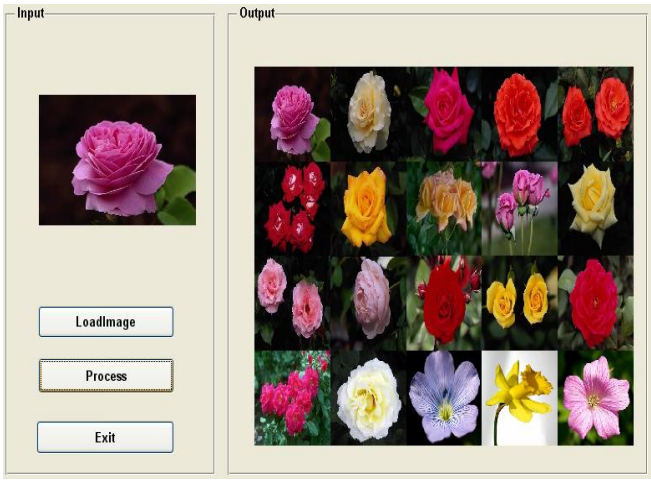


Figure 6
Sample Output - Query Image and Retrieved images

4.1 Comparative Analysis

The performance can be identified by using precision and recall. Precision is the fraction of retrieved images that are relevant to the query image, while recall is the fraction of relevant images that are retrieved from the database. Precision indicates the accuracy and Recall indicates the relevance of retrieval.

$$precision = \frac{\text{Count of retrieved images relevant to the query image}}{\text{Total count of images retrieved}} \tag{4}$$

$$recall = \frac{\text{Count of retrieved images relevant to the query image}}{\text{Total count of relevant images in the database}} \tag{5}$$

Using Equation (4) and Equation (5), the precision and recall values for the query image are calculated for the proposed method. The values obtained from the calculation are given in Table 1. The precision values are also compared with some of the existing methods and plotted in Figure 7.

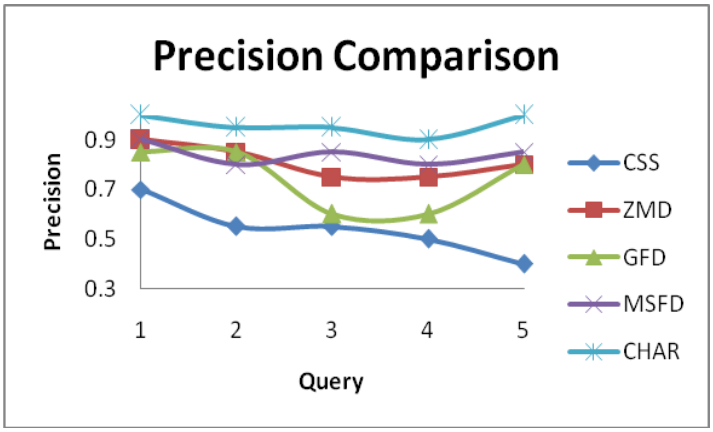


Figure 7
Comparison of Precision

Table 1

Precision (P) and Recall (R) Statistics for the proposed CHAR (Convex Hull Area Ratio) Method on the best 15, 20 and 40 retrieved images

Query	N = 15		N = 20		N = 40	
	P	R	P	R	P	R
1	1	0.30	1	0.40	0.9	0.72
2	1	0.42	1	0.58	0.65	0.74
3	1	0.30	0.95	0.38	0.98	0.78
4	1	0.30	1	0.40	0.9	0.72
5	1	0.30	1	0.40	0.9	0.72
6	0.93	0.28	1	0.40	0.93	0.74
7	0.93	0.28	0.9	0.36	0.88	0.70
8	1	0.30	0.95	0.38	0.88	0.70
9	0.93	0.28	0.9	0.36	0.8	0.64
10	0.86	0.26	0.9	0.36	0.83	0.66

Figure 7 shows the graph for the comparative analysis between the proposed Convex Hull Area Ratio (CHAR) method and existing methods like CSS (Curvature Scale Space), ZMD (Zernike Moment Descriptor), GFD (Geometric Fourier Descriptor) and MSFD (Multiple Shape Feature Descriptor) on 5 queries for a set of best 20 images retrieved. The data for other methods is from [4] [7].

Based on Figure 7, the proposed method has effectively retrieved the images with high precision, in comparison with other existing methods. The retrieval is also quite fast since only the convex hull points are used in the matching process. The Precision - Recall table and the Precision comparison establish the superiority of the proposed CHAR-based Content-based Image Retrieval System. This work can further be improved by preserving rotation invariance. Feature set can also be extended to include inner shape features and texture dimensions by grouping the edge sets based on their color and texture properties. However this will increase the complexity of the CBIR system and can be justified only when the additional computation is mandated for specific applications to delineate color as well as texture.

Conclusions

In this paper, we have proposed a novel and efficient CBIR system based on Shape Signature to retrieve relevant images from image database for a given query image. In this method, when an image is queried, the system establishes shape feature for the image and then the ratio of the intersected area to union area of the convex hull polygons of the query and database images are found and stored in an array. Subsequently, similarity measure is performed and the maximum value of the ratio indicates the closest match. Here we first pre-process the image by using K-means clustering algorithm for image segmentation based on color. Edges are extracted using the Canny algorithm. We have proposed and implemented the CHAR method for efficient matching and retrieval. The implementation results illustrate that this novel image retrieval process effectively retrieves the images that are close to the query image from the database.

References

[1] Ricardo da S. Torresa, Alexander X. Falcaoa, Marcos A. Gonçalvesb, Joao P. Papaa, Baoping Zhang, Weiguo Fanc and Edward A. Foxc: A Genetic Programming Framework for Content-based Image Retrieval, Journal of Pattern Recognition, Vol. 42, No. 2, 2009, pp. 283-292

[2] Ying Liu, Dengsheng Zhang, Guojun Lu and Wei-Ying Ma: A Survey of Content-based Image Retrieval with High-level Semantics, Journal of Pattern Recognition, Vol. 40, No. 1, 2007, pp. 262-282

[3] Amit Jain, Ramanathan Muthuganapathy and Karthik Ramani: Content-based Image Retrieval Using Shape and Depth from an Engineering Database, In Proceedings of the Third International Conference on Advances in Visual Computing, 2007, pp. 255-264

[4] Dengsheng Zhang and Guojun Lu: A Comparative Study of Curvature Scale Space and Fourier Descriptors for Shape-based Image Retrieval, Journal of Visual Communication and Image Representation, Vol. 14, No. 1, 2003, pp. 39-57

[5] Dengsheng Zhang and Guojun Lu: Shape-based Image Retrieval using Generic Fourier Descriptor, Journal of Signal Processing: Image Communication, Vol. 17, No. 10, 2002, pp. 825-848

[6] Chia-Hung Wei, Yue Li, Wing Yin Chau and Chang-Tsun Li: Trademark Image Retrieval Using Synthetic Features for Describing Global Shape and Interior Structure, Journal of Pattern Recognition, Vol. 42, No. 3, 2009, pp. 386-394

[7] Xiang-Yang Wang, Yong-Jian Yu, and Hong-Ying Yang: An Effective Image Retrieval Scheme using Color, Texture and Shape Features, Journal of Computer Standards & Interfaces, Vol. 33, 2010, pp. 59-68

[8] Chuan-Cheng Wang and Ling-Hwei Chen: Content-based Color Trademark Retrieval System Using Hit Statistic, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 16, No. 5, 2002, pp. 603-619

[9] Ying Liu, Dengsheng Zhang, Guojun Lu and Wei-Ying Ma: A Survey of Content-based Image Retrieval with High-level Semantics, Journal of Pattern Recognition, Vol. 40, No. 1, 2007, pp. 262-282

[10] H. B. Kekre and Dhirendra Mishra: Four Walsh Transform Sectors Feature Vectors for Image Retrieval from Image Databases, International Journal of Computer Science and Information Technologies, Vol. 1, No. 2, 2010, pp 33-37

[11] Shao-Hu Peng, Deok-Hwan Kim, Seok-Lyong Lee and Chin-Wan Chung: A Visual Shape Descriptor Using Sectors and Shape Context of Contour Lines, Journal of Information Sciences, Vol. 180, No. 16, 2010, pp. 2925-2939

[12] Srinivasa Rao, S. Srinivas Kumar and B.Chandra Mohan: Content-based Image Retrieval Using Exact Lengendre Moments And Support Vector Machine, The International Journal of Multimedia & Its Applications, Vol. 2, No. 2, 2010, pp. 69- 79

[13] Ravichandran and Ananthi: Color Skin Segmentation Using K-Means Cluster, International Journal of Computational and Applied Mathematics, Vol. 4, No. 2, 2009, pp. 153-157

[14] Krishnan N, Justin Varghese, S. Saudia, Santhosh. P.Mathew et al: A New Adaptive Class of Filter Operators for Salt and Pepper Impulse Corrupted Images, International Journal of Imaging Science and engineering (IJISE), Vol. 1, No. 2, 2007, pp. 44-51

[15] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang: Image Retrieval: Ideas, Influences, and Trends of the New Age, ACM Computing Surveys, Vol. 40, No. 2, 2008, Article 5

[16] Lancaster, JL, Fox, PT, Downs, H, Nickerson, DS, Hander, TA, Mallah, ME, Kochunov, PV and Zamarripa, F: Global Spatial Normalization of Human Brain Using Convex Hulls, The Journal of Nuclear Medicine, Vol. 40, No. 6, 1999, pp. 942-955

[17] Szabolcs Sergyán: A New Approach of Face Detection-based Classification of Image Databases, Acta Polytechnica Hungarica, Journal of Applied Sciences, Vol. 6, No. 1, 2009, pp. 175-184

[18] S. P. Mathew and P. Samuel: A Novel Image Retrieval System using an Effective Region-based Shape Representation Technique, International Journal of Image Processing (IJIP), Vol. 4, No. 5, Dec. 2010, pp. 509-517

[19] Santhosh P. Mathew, Valentina E. Balas, Zachariah K. P, Philip Samuel: A Content-based Image Retrieval System Based on Polar Raster Edge Sampling Signature, Acta Polytechnica Hungarica, Journal of Applied Sciences, Vol. 11, No. 3, 2014, pp. 25-36

[20] Loris Nanni, Alessandra Lumini, Sheryl Brahnam: Ensemble of Shape Descriptors for Shape Retrieval and Classification, International Journal of Advanced Intelligence Paradigms (IJAIP), Vol. 6, No. 2, 2014, pp. 136-156

[21] S. Bharathi, R. Sudhakar, Valentina E. Balas : Biometric Recognition using Fuzzy Score Level Fusion, International Journal of Advanced Intelligence Paradigms (IJAIP), Vol. 6, No. 2, 2014, pp. 81-94

Global Dynamic Slicing for the C Language

Árpád Beszédes

University of Szeged, Department of Software Engineering
Árpád tér 2, H-6720 Szeged, Hungary
beszedes@inf.u-szeged.hu

Abstract: In dynamic program slicing, program subsets are computed that represent the set of dependences that occur for specific program executions and can be associated with a program point of interest called the slicing criterion. Traditionally, dynamic dependence graphs are used as a preprocessing step before the actual slices are computed, but this approach is not scalable. We follow the approach of processing the execution trace and, using local definition-use information, follow the dependence chains "on the fly" without actually building the dynamic dependence graph, but we retain specialized data structures. Here, we present in detail the practical modifications of our global dynamic slicing algorithm, which are needed to apply it to programs written in the C language.

Keywords: program slicing; dynamic slicing; program analysis; program dependence; C

1 Introduction

Program slicing [1, 2] is a program analysis technique that is used to help solve various software engineering problems. A slice of a program is the program's subset which consists of only those statements that directly or indirectly affect the value of a variable occurrence (known as the slicing criterion). This form of slicing is referred to as backward slicing. In contrast, forward slicing involves looking for forward dependences; those statements that may be affected by a specific program point. If the dependence set is determined in such a way that it reflects the dependences for all possible executions, we call it a static slice. However, if only a specific program execution is investigated, it is called a dynamic slice. In our study we will focus on backward dynamic slicing.

Dynamic program slicing [3] has a certain advantage in some applications; namely, the dynamic slices are significantly smaller than their static counterparts. For instance, when debugging we seek the possible cause(s) of an error that was observed at a specific program point and for a specific run. The more precisely this set of causes is defined, the more effective the debugging should be.

A common approach to dynamic slicing is based on computing the *dynamic dependences* among the program elements. The method by Agrawal and Horgan [3] uses a graph representation called the Dynamic Dependence Graph (DDG), which includes a distinct vertex for each occurrence of a statement in the execution history (the list of statements executed), and the edges correspond to the dynamically occurring dependences among them. Based on this graph, the computation of a dynamic slice means finding all the reachable vertices starting from the slicing criterion. The DDG-based method can be used to compute dynamic slices in a general way, since it performs a full preprocessing [4] before the actual slicing. When building the graph in advance, the user has the possibility of computing different slices starting from different program points (the criteria) and going in different directions (forwards or backwards). Since the computation and storage for such a graph is expensive more specialized approaches that take into account the desired slicing scenario should be considered.

In previous studies [5, 6, 7], we devised new efficient dynamic slicing methods that were based on dynamic dependences, but did not require full preprocessing and the building of huge representations like the DDG graph. We also process the execution history and some elements of the complexity of our approach are related to the length of the execution as well; but other, more specialized data structures and algorithms are applied in order to improve the overall efficiency. One of the results is a backward slicing algorithm [6, 7] that computes all the possible dynamic slices globally, with only one pass through the execution history. This method significantly differs from the previously published slicing algorithms, and it is believed to be applicable for real-size programs and executions. We presented some details of the algorithm in different contexts; for C [6, 8] and for Java [9] programs, and for different applications [7, 10].

In this paper, we provide details on the implementation of the global algorithm for backward slicing for C programs. This makes its implementation possible in virtually any context and platform for C.

2 Previous Results and Related Work

Program slicing has a large literature and many different approaches have been devised. Surveys can be found at various places, e.g. [2, 11]. While the practical static slicing methods are mostly based on the PDG-based algorithm by Horwitz *et al.* [12], there are several, quite different approaches to dynamic slicing. One usual categorization of the dynamic slicing methods comes from asking whether the program subset produced (the slice) is an executable program or not. Executable slices are needed for certain applications, but they are less accurate.

Actual implementations of dynamic slicing algorithms were mentioned in very few publications, such as in [13, 14]. However, these implementations did not prove to be suitable for real-life applications. In a study, Venkatesh experimented with different algorithms and provided experimental data [14]. In this experiment, four kinds of slicing algorithms were implemented for the C language, including the dependence-based approach by Agrawal and Horgan [3] and the executable slicing method by Korel and Laski [15]. Unfortunately, no details were given on the design and functionality or the special features of the implementation for C.

In the DDG-based method by Agrawal and Horgan [3], the size of the DDGs may be huge. In fact, it is not bounded by the program dimensions, but it correlates with the execution length. In their study, Agrawal and Horgan therefore proposed a reduced DDG method, where the size of the reduced graphs was bounded by the number of different dynamic slices. Alas, even this reduced DDG may be very large for some programs. In [16], Zhang *et al.* elaborate on the problems of existing dynamic slicing algorithms concerning their computation and space complexity. They claim that most accurate (they use the term "precise") algorithms are significantly less efficient than the approximate methods, which in turn produce inaccurate dynamic slices. This inefficiency may be attributed to two factors: either the execution trace is completely processed before the actual slicing algorithm is performed (referred to as full preprocessing) or the slicing algorithm is invoked on demand, processing the trace from the start for each slicing request (referred to as no preprocessing). Our global and demand-driven algorithms [5] correspond to the first and second cases, respectively. In both approaches the authors gave their own implementations based on dependence graphs. To reduce the overheads of each approach they proposed a combined algorithm called limited preprocessing, where the execution trace is augmented with summary information to allow a faster traversal when the slice is computed.

2.1 Our Global Dynamic Slicing Algorithm

In a previous study [5], we investigated practical ways of computing the dynamic slices based on dynamic dependences, but without requiring costly global preprocessing. We proposed alternative methods based on the same dynamic dependences, but instead of DDG graphs specific data structures were used for each algorithm. These structures are different depending on the slicing scenario, and – having specific applications – some of the algorithms are more efficient in terms of storage, while others have improved runtime efficiency. The different slicing scenarios that we investigated are global vs. demand-driven slicing and computing backward vs. forward slices. One favorable property of these algorithms is that they are able to compute the same dynamic slices as the original DDG-based method. It turns out that the slices can be produced by traversing the execution history either in a forward or in a backward way, and that some processing directions fit better in one slicing scenario than in another. This gives

eight possibilities, some of which give useful algorithms, while others prove unfeasible. In our paper, we elaborate the topic by providing details on handling real C language constructs for the global algorithm.

Our approach for computing dynamic slices differs significantly from the previous methods. We designed the slicing algorithms so that they can be effectively implemented and used in practice. Hence, we tried to minimize the amount of information that must be computed and stored during the computations [5, 6, 7]. In our algorithms, we track the data and control dependences among the program instructions that arise dynamically during execution. The algorithm works on the trace of the execution, which is produced using a statically instrumented version of the program. The trace includes all the necessary information about the runtime behaviour of the program.[1] For producing the required slicing results, the algorithm relies on statically computed information from the code as well. The global algorithm (also referred to as the *forward* algorithm) starts at the beginning of the trace with the first executed instruction and propagates the dynamic dependences in parallel with the execution, and eventually provides the required slices for all the possible dynamic criteria (for all the variables). Evidently, this approach has its benefits and drawbacks, but the other algorithms presented in [5] provide feasible alternatives. For details of the conceptual algorithm with examples, please see the articles cited above.

3 Global Dynamic Slicing for C Programs

In other studies [5, 6, 7], we presented conceptual algorithms for dynamic slicing. The concepts were introduced for programs in which only scalar variables were used and without interprocedurality. The application of the conceptual algorithm to C programs gives rise to several problems. In our study, the handling of various language constructs were addressed in the following way:

1) All computations are performed on *memory locations* instead of handling scalar variables, pointers and other more complex objects differently. This approach enables an easy and uniform handling of pointers, pointer dereferences, arrays and structures by transforming them to the actual memory locations. In our approach, the dependences for a pointer dereference will include the dependences of the pointer itself and the dereferenced memory location as well. Also, accesses to union members and C bitfields are treated as dependences for the whole data structure (struct and array members are handled individually). Slicing on memory

[1] Execution tracing is used in other areas of software engineering as well; it has recently been proposed to extend software product quality frameworks [18].

locations is a feasible approach since all the dynamic information on the actual storage of objects is available.

2) Since each C statement (and expression with side-effects) may imply the definition of more than one object, a *definition-use list* is defined for each executable instruction, rather than a single definition-use pair as we have with the conceptual algorithm [5, 6, 7] (a *definition-use pair* or *def-use pair* consists of a variable name that is defined at the instruction and a set of variables that is used in the instruction for computing the value of the defined variable). This list is essentially a sequence of def-use pairs that all occur in an instruction (see below for a complete definition).

3) All slicing criteria and slicing results are given for line numbers in the original source file. However, since the computations are made on memory locations and for (possibly multiple) objects defined for one statement, the necessary mappings must be made.

4) Interprocedural dependences that arise across function calls can be handled relatively easily by adopting the memory slicing approach, since each memory address can be viewed as a "global variable." The execution history will contain each realized function call, and the order of the instructions executed will also be known. We only have to handle the actual arguments as special local variables and the return value as a special variable defined at the call site.

5) Local variables are also handled by using their addresses on the actual call stack frame. We only need to track the block scopes dynamically for lookup purposes. The handling of globals is also simple due to using their addresses for computation (which are fixed for the whole execution of the program).

6) The unstructured control transfers (*goto* and other jump statements) are handled by adding all the possible control dependences to the def-use representation (for a block-based language as in our conceptual description, the control dependences are determined by the syntax). As this way some statements may be dependent on multiple predicates, the handling of predicate variables in the presence of jumps needs to be slightly extended (the details are given below). Currently, C "long jump" constructs are not handled, but they could be treated in the same way.

7) The conceptual algorithm uses the concept of execution history to record the instruction numbers executed. To be able to slice a C program, however, some other information is also needed that is generated upon executing the program, and which is used by the slicing algorithms. This includes the addresses of variables, function calls and block scope information. We will call this extended execution history the *trace*.

8) Declaration lines will be added to the slices whenever the definition of the declared variable is added to the slices. Also, the eventual initializations will be added to the def-use representation.

9) Since programs generally rely on standard library code as well, we must handle interprocedural dependences arising from the parameters, side effects and return values of calls to library code. Since the source code of library functions is often unavailable, we will rely on the semantics of such functions and prepare, in advance, a def-use representation of each standard library function based on the specifications.

10) Real programs usually consist of multiple source files composed of header files (`.h`) and implementation files (`.c`), which produce translation units after preprocessing. Our slicing algorithm works on preprocessed units, which makes it possible to compute slices for the whole program. What is needed to achieve this is a global numbering of statements over all the source files of the program, and solving name identification for definitions coming from common header files and placed into multiple translation units, as we do with a linker.

Based on these considerations, the implementation of our dynamic slicing approach consists of four phases. During a static analysis, the def-use representation of the program is produced and stored on the disk, and the source code is instrumented.[2] Next, the instrumented code is built to produce an executable program, which is executed in the next phase. During this operation a trace of the program is produced with the help of the instrumentation code. Lastly, the slicing algorithm is executed, where the trace is used to drive the propagation of the dependences, in the global algorithm starting from the beginning of the trace. The slicing algorithm relies on the def-use representation produced in the first phase. Below, we will describe these phases and specific features of the implementation for C.

3.1 Static Analysis

Static analysis has two goals: to produce the def-use representation (Section 3.2) and to instrument the code (Section 3.3). Another task here is to create a mapping between the physical source code lines and the internal identifiers given to program elements by the analyzer. Our static analysis front end works on the preprocessed code, and it performs lexical and syntactic analysis, producing an annotated Abstract Syntax Tree as the result for each unit. The AST contains

[2] In this study, we used source code-level instrumentation, but other ways exist as well such as binary-level and virtual machine-level solutions. It should be added that source code-level instrumentation has the highest risk of changing program behavior, but when experimenting with our prototype we did not encounter any such problems.

sufficient information to compute the def-use representation and perform code instrumentation.

3.2 Def-Use Representation for C

In our implementation for the C language, an extended def-use representation is created and stored in a file, which will be used later by the algorithms. In the conceptual algorithm [5, 6, 7], the def-use representation was defined as *i. d: U* for each program instruction number *i*. For real C programs, this representation (also called the *D/U representation* below) will be extended so that it contains a sequence of *d: U* items for each instruction *i* in the program: $i.\ \langle (d_1 :\ U_1), (d_2 :\ U_2), \ldots, (d_{m_i} :\ U_{m_i}) \rangle$. We will use the notation $DU^C(i)$ for the D/U sequence of the *i*-th instruction.

This extension is needed because in a C instruction (i.e. an executable expression with side-effects), several l-values may be assigned new values. Note that the sequence order is important, since the *d* values of a previous D/U item can be used by the subsequent *U* sets. This sequence order is determined by the "execution-order" (evaluation) of the corresponding subexpressions. The order of the evaluation of subexpressions in C is not always defined by the language, hence there might be complications arising from the use of different compilers and compilation options. In our current implementation, we will rely on the parsing sequence determined by the context-free grammar of C, which proved to be sufficient in our prototype. In a production tool, however, care should be taken to handle the various possibilities.

The other modification needed for the D/U representation for C is that the variables (including artificially created ones) in it are not only simple scalar or predicate variables, but they can also take several different meanings as follows:

1) *Scalar variables.* These are the "regular" global or local variables (with static storage, they have a constant address for the actual call stack frame). The formal parameters of functions are also represented as if they were local variables in the function's scope. Note that dynamic variables used with dynamically allocated memory on the heap do not need special treatment as they will be treated as pointers and the corresponding allocator functions as library code (see above).

2) *Predicate variables.* Denoted by p_n, where *n* is the serial number of the predicate instruction, the predicate variables are artificial variables with the same semantics as those described in the conceptual algorithm. In the case of the C language, all iteration and selection statements will induce predicate variables. An additional, special form of predicate variables will be introduced, one for each function and will be denoted by *entry(f)*, to generalize the representation of control dependences. Such an "entry-

predicate" is defined upon entering the function *f* and is used by all statements outside any other predicates in the function.

3) *Output variables*. Denoted by o_n, the output variables are artificial variables that are generated at the places where a set *U* is used, but no other variable takes any value from *U*. These include function calls with their return values ignored, single expression-statements with no side-effects, jump statements, and some output statements in C such as `printf`.

4) *Dereference variables*. The notion of the (artificial) dereference variables is employed where a memory address is used in any possible way or where it gets a value through a pointer (or an array or structure member). They are denoted by d_n, where *n* is a global counter for each dereference occurrence. Dereference variables will be created for the following code constructs: `*expr`, `object.member`, `ptr->member` and `array[index]`. Note that in an implementation, some of these could be handled uniformly as a base pointer+offset, but source code instrumentation requires a different treatment. Dereference variables will be used in such a way that their dependences will be noted in the D/U representation only symbolically, while the actual dependences will be computed for the associated addresses written to the trace. Note that the order in which the dereference variables are stored in the use sets must be the same as the order in which they will be evaluated.

5) *Function call argument variables*. These are artificial variables denoted by *arg(f,n)*, where *f* is a function name and *n* is the function argument (parameter) number. An argument variable is defined at the function call site and used at the entry point of the function (by defining the formal parameter).

6) *Function call return variables*. Denoted by *ret(f)*, where *f* is a function name, the artificial return variables are defined at the exit point of the function and used at the function caller after returning.

In the extended D/U representation, regardless of the type of variable, all dependences are treated equally. For instance, a pointer dereference may be dependent on a predicate variable if the dereference subexpression is control-dependent on a predicate. This uniform handling allows a very concise capturing of the interdependences of the program, and a straightforward implementation of the algorithms. In the following, we will describe how the dependences in the D/U representation are built up and relate to special features of the C language.

Computation of the data dependences. Generally speaking, the structure of the D/U representation is such that it captures the definition-use relationships *locally* for each statement. This means that we do not need to deal with the classical problems of computing data dependences in the static case, as is required with the dependence graphs [12]; in our case only the names of the dependent variables

(and not the corresponding definition) need be stored. Thus, our representation for C can be constructed in a simple syntax-directed manner following the semantics of each C expression construct.

Function calls. Function calls and parameter passing are handled in the D/U representation using the artificial variables *arg* and *ret* (see above). Whenever a function call expression is found in a C instruction, a corresponding D/U item is created with the *arg* variable as the defined one and the appropriate use set. Next, in each function a D/U item is constructed for all its formal parameters in which the parameter is the defined variable (the parameter is later treated as a local variable) and the corresponding *arg* variable constitutes the use set. Furthermore, for each return statement in the functions a D/U item is created with the *ret* variable defined and the corresponding use sets. Lastly, at the call site these *ret* variables are used in the corresponding use sets for the expressions containing the function call. The order of elements in the D/U lists is important as this is required for the synchronization with the trace.

Structured control dependences. The predicate variables will be used in the D/U representation to capture the control dependences among the program instructions. In the case of structured control transfers (the *if* selection and the three types of C loops), for each predicate corresponding to the respective decision statement a predicate variable will be created and the dependences will be based on the nesting structure of the program; the directly nested statements of *if* branches or a loop will be dependent on the corresponding container predicate. To make the algorithm more general, for each function an additional predicate called the *entry-predicate* will be defined as well. The instructions that are not nested within another predicate statement will be dependent on the entry-predicate. (The entry-predicates are implicitly defined at the function beginning and their use sets are empty.) Note that shortcut logical expressions do not influence this operation.

Handling of goto and other unstructured jumps. While the direct control dependences can be readily determined for structured programs, *goto*-s and other arbitrary control transfers (*switch*, *continue* and *break*) must be handled in a more elaborate way. We will compute control dependences in the static analysis phase based on the traditional approach using postdominance relations [17], and then build the extended D/U representation based on this information. Namely, if an instruction i is found to be control dependent on some other instruction (which is then a predicate), we extend the use set of i with the corresponding predicate variable. Since in a program with arbitrary control flow an instruction may be control dependent on more than one instruction, our use sets may also contain several predicate variables. In one specific execution only one of them will be responsible for the actually realized control dependence, which we will call the *active predicate*. When propagating the dependences through the current instruction's use set, we must select just one predicate variable to continue with. If there are more predicate variables in the use set, our approach is to choose the one that has been defined most recently. In other words, for i^j. *d: U*, we will choose

predicate p for which $LD(p) = \max\{LD(r) \mid r \in U$ and r is a predicate variable$\}$, where $LD(v)$ is the last definition of variable v, i.e. the execution step at which v was defined just before the j-th step where i was executed. (In the following, we will refer to execution history elements as *actions* with the notation i^j, where i is the serial number of the instruction executed at the j-th step or position.)

Complex l-values. A side effect of certain C expressions is that the sub-expression on the left hand side of the operator takes the value of the right hand side (this includes the assignment operators as the most common ones). These operators require that their left hand side be an l-value (meaning that it is modifiable). Quite frequently, the l-value is a simple variable occurrence, but these sub-expressions can be arbitrarily complex. In such cases, the D/U representation needs to be constructed carefully to include all the defined and used variables appropriately. One important issue is the handling of pointer dereference expressions of the form `*p`. Strictly speaking, the data pointed to by a pointer is not dependent on the address itself. However, we will apply a conservative approach and include such pointers as well (this approach is also used by some other algorithms). In Figure 1, we list some other cases and the way we treat them in our representation (following the principles for dereference variables introduced above).

```
a[i]       = r; //        d1:{r,i,a}
*(p+x)     = r; //        d2:{r,p,x}
m.a        = r; //        d3:{r}
p->a       = r; //        d4:{r,p}
```

Figure 1
Handling of field accesses

Clearly, this is a conservative approach as, for example, the array name `a` and the index variable `i` both appear in the use-set of the first statement; however from a computational point of view only the data at the address pointed to depends on `r`. Although debatable, here we shall choose this approach to be able to compute a conservative-type of dependence which can be used, for instance, to assist impact analysis.

Pointers, pointer dereferences, address-of and arrays. Our algorithm computes the dependences on memory locations, which makes the handling of pointers and related structures straightforward, but there are several special features worth mentioning. As we said previously, in the case of pointer dereferences both the pointer and the dereferenced object will be included. Using the address-of operator does not induce a new dependence because the address itself can be viewed as a constant value. All the other operations with pointers are treated in the same way as in the case of regular variables. Arrays can be handled in a similar way as pointers since they can be interpreted as pointers with appropriate offsets corresponding to the index. The only extension is that the variable(s) used in the index operators are also treated as used variables. Multiple pointers and indirections can be handled in the same way as well.

The handling of function pointers does not require major modifications to the presented algorithms, but we omitted these details from the formal algorithm (in the next section) to aid readability, and we describe them more fully here. Statically, we cannot determine the called function, so in the D/U representation we cannot use *arg(f,n)* and *ret(f)* variables either. Instead, we use their special form in which the actual names are not given, just some symbolic names of the form *arg(?,n)* and *ret(?)*. These temporary variables will be resolved upon the execution of the slicing algorithms as soon as the called functions become known.

Structs and unions. C language *unions* and *bitfields* can be handled in a conservative way. Namely, we will treat unions and bitfields as scalar variables because when we define a field we virtually define all the others as well. Bitfields can introduce multiple dependences due to overlaps in memory regions, which will be handled in a similar way to that with type-casts, described later on. In the case of *structs*, however, we want to preserve the individual tracking of the dependences of the fields as in the case with arrays. For this, we follow a similar approach to the handling of arrays because the structs can also be interpreted as memory regions with a fixed base address and offsets corresponding to the fields. That is, for each field access we create a distinct dereference variable, which we can use separately in the D/U sets.

The handling of the individual struct variables as parts of expressions is more complicated because the expression operations in this case will correspond to all the fields together (struct copying). In this case, we model the dependences for each field access combination (which may be recursive). The struct variable itself will not be part of the D/U sets, but all the references to it will be transformed to the actual field accesses for all the fields. Figure 2 provides examples for handling structs, members and dereferences (in the commented lines below line 2, we can see how the fields are modelled).

```
   struct S s,t,*p,*q;
   t.a = ...
   t.b = ...
1. q      = &t;   //     q : {}
2. s      = *q;   //     : {}
// s.a    = q->a; // d2:{q,d1}
// s.b    = q->b; // d4:{q,d3}
// ...
3. x      = s.a;  //     x : {d5}
4. p      = &s;   //     p : {}
5. y      = p->b; //     y : {p,d6}
```

Figure 2
Handling of struct variables

For instance, during a slice computation, the runtime addresses of `d5` and `d2` will be the same, which will result in correct dependences between `s.a` and `t.a`.

Type casts. Type casts can cause a problem for our slicing algorithms in cases where the sizes of the original type and the new type are different. The basic methods discussed above will lose any dependences among overlapping memory regions of the objects (e.g. structure members). The problems related to type casts can be handled only by maintaining the length of the referenced memory addresses as well as their starting address. We did not include this in the formal description of the algorithms for the sake of clarity, but we will overview the basic method here. The D/U representation does not include any specific extensions, but in the execution trace we will output the dereference addresses and the sizes of the variables in question, which will form regions instead of single addresses (`sizeof` can be used in the instrumented code for this purpose). The slicing algorithm will then take into account each byte of the referenced memory region, which will result in not losing any dependences; and this will be suitable for all kinds of type casts, including casts between scalars and pointers.

3.3 Instrumentation and the Trace File

The purpose of code instrumentation is to produce a semantically equivalent code that, upon execution, produces a *trace* of the execution. The trace records the executed i^j actions and other information required by the slicing algorithm. It is a linear sequence of elements with various meanings, which is, upon execution, stored in a file for later processing. The sequence can be described with a context free grammar shown in Figure 3.

$$
\begin{array}{rcl}
\langle trace\text{-}file \rangle & ::= & \{ \langle global\text{-}var \rangle \}\ \langle main\text{-}function \rangle \\
\langle global\text{-}var \rangle & ::= & G\ (\ \mathbf{id}\ ,\ \mathbf{addr}\ ) \\
\langle local\text{-}var \rangle & ::= & D\ (\ \mathbf{id}\ ,\ \mathbf{addr}\ ) \\
\langle function \rangle & ::= & FB\ (\ \mathbf{id}\ )\ \{ \langle function\text{-}body \rangle \}\ FE \\
\langle main\text{-}function \rangle & ::= & FB\ (\ \mathtt{main}\ )\ \{ \langle function\text{-}body \rangle \}\ FE \\
\langle function\text{-}body \rangle & ::= & \langle local\text{-}var \rangle \\
 & | & \langle action \rangle \\
 & | & \langle block\text{-}scope \rangle \\
\langle block\text{-}scope \rangle & ::= & BB\ (\ \mathbf{bnum}\ )\ \{ \langle function\text{-}body \rangle \}\ BE\ (\ \mathbf{onum}\ ) \\
\langle action \rangle & ::= & E\ (\ \mathbf{inum}\ ,\ \mathbf{jnum}\ )\ \{ \langle action\text{-}suffix \rangle \} \\
\langle action\text{-}suffix \rangle & ::= & P\ (\ \mathbf{addr}\ ) \\
 & | & \langle function \rangle
\end{array}
$$

Figure 3
Formal description of the trace

The order of elements in the trace is determined by the execution of the instrumented program. First the data for all of the global variables are dumped (mark *G* with the variable name and its actual address). Then the execution is traced starting with the `main` function. On entering a function, a function-begin

mark with the function name (*FB*) is generated, and on exiting it a function-end mark (*FE*) is generated. During the execution of a function body, three kinds of events can occur: the data for a local variable (*D*) is generated in a similar way to that for the globals, or a nested block (corresponding to a syntactic block in a C program) is generated with the delimiting marks (*BB* with a unique block serial number and *BE* with the identifier of an outer block), or an executable instruction (action) is traced. The delimiting marks are not generated only for the blocks according to the syntax with { and }, but for each jump instruction into or out of some blocks and single statement sub-instructions as well. The block identifier that comes with *BE* is the number of the block in which the next executable instruction is located. Usually, it is the block containing the current one, but in the case of unstructured jumps it may be any block in the current function.

An action is generated for each C instruction (expression) and it consists of two parts. The main part (*E*) designates the executed instruction number *i* and the execution step *j*. In addition, an optional list of information (the action suffix) related to the current instruction may be generated. Here, there are two types of action suffixes. If a function call is a part of the expression of the current action, the trace for the whole function will be dumped as a suffix for the current action. This can clearly result in a large amount of recursive data structures being generated, which may be similar at different instances if the invocation is similar. This could be optimized in an implementation by applying some kind of a compression; however, here we do not implement such a feature. The other kind of action suffixes will be generated whenever a pointer dereference is encountered in the expression of the current action. The accessed memory address is dumped into the trace using *P*. For each action, the additional dereferences will correspond to the relevant dereference variables in the D/U. Note that the order in which the *P* marks will be generated is the same as the way they are executed, and this order must also be the same as the corresponding dereference variables are listed in the D/U representation. This property will be exploited by the slicing algorithm.

To get the required contents of the trace file, the source code needs to be *instrumented* at several locations. At each relevant point, a call to an instrumenting function is generated, which will place the necessary marks into the trace file. We chose the instrumented code to be C++ rather than C for practical reasons.[3] For example, some instrumentor functions are easier to implement as template functions, and we can also put the calls to the instrumentor functions before the variable declarations. The instrumentor functions are provided in additional source and header files, which need to be included in the linking phase when the program is built. To implement the instrumentation for each trace element, several practical solutions had to be elaborated, for instance: block and function delimiting marks had to be placed at various places due to possible

[3] Note, that this solution might be problematic when certain language features are used in the original C program that are incompatible with the selected C++ compiler.

jumps; local and global variables are dumped using the address-of operator; action marks are generated for each expression using the comma-operator; dereference marks are generated by a C++ template function that returns the pointer to a type passed in the template parameter, etc. Figure 4 shows an excerpt from the C program bzip, its instrumented version and a part from the generated trace file.

Since the instruction numbers are generated incrementally, we need to maintain a data structure to map the instruction numbers to the physical file line numbers (line numbers will be essential in presenting the actual results of slicing). The method of mapping line numbers to instruction numbers depends on the actual implementation of the static phase. In our toolset, we used the information taken from our static analyzer for this purpose, which takes into account both the fully qualified file names and the absolute line numbers. Here, we can use line information got from both the preprocessed file and the original file locations.

```
Int32 nb, na, mid;
nb = 0;
na = 256;
do {
   mid = (nb + na) >> 1;
   if (indx >= cftab[mid]) nb = mid; else na = mid;
}
```

```
D_VA(&nb,"nb");
D_VA(&na,"na");
D_VA(&mid,"mid");
D_EH( 1259,D__ec++); D_EB() ,(nb = 0);
D_EH( 1260,D__ec++); D_EB() ,(na = 256);
do
{ /*BlockGuard*/
{
D_SB(243);
D_EH( 1262,D__ec++); D_EB() ,(mid = (nb+na)>>1);
D_EH( 1263,D__ec++); if ( D_EB() ,(indx>=(*D_P(&cftab[(mid)]))) )
{ /*BlockGuard*/
D_EH( 1264,D__ec++); D_EB() ,(nb = mid)
;} /*BlockGuard*/
else
{ /*BlockGuard*/
D_EH( 1265,D__ec++); D_EB() ,(na = mid)
;} /*BlockGuard*/
;
D_SC(242);
}
;} /*BlockGuard*/
```

```
D nb 0x0012F018
D na 0x0012F010
D mid 0x0012F014
E 1259 9578
EB
E 1260 9579
EB
SB 243
E 1262 9580
EB
E 1263 9581
EB
P 0x0012F3C0
E 1265 9582
EB
SC 242
```

Figure 4

Instrumentation and trace file example

3.4 Global Algorithm for C

The extended global algorithm for slicing C programs with the solutions to the problems elaborated on earlier can be seen in Figures 5 and 6. Here, the notation *TR >> tr* is used to denote the reading of the next trace element *tr* from the trace *TR*, which is viewed as a stream of elements, as described in Section 3.3. Other formalisms are self-explanatory. Note that for the sake of clarity we omitted such supporting activities as error handling and synchronization support between the trace and the algorithm.

The algorithm begins with the program GlobalAlgorithmForC, which has two input parameters; namely, program *P* that is to be sliced and input `x` for which the dynamic slices will be computed globally. First the trace is produced (in a file), which is read sequentially (lines 1-2). The algorithm is driven by the elements found in the trace, but its structure must be in sync with the static D/U representation (see, for example, function calls and dereference marks). The function calls are captured in the trace recursively, so they are also handled by the algorithm by recursively calling the function ProcessFunction when such a call is found. The main program of the algorithm (after storing the addresses of global variables on the scope stack written in lines 3-6) starts by processing the `main` function in line 7.

During processing, a helping structure is maintained for the local and global scalar variables. This structure (*sc*) is a stack of scopes that are entered dynamically upon execution. The scope stack is maintained in the function ProcessFunction, as dictated by the trace. Namely, a new function scope is created on the top when entering a function (*FB*) in lines 12-13. As we saw earlier, the block beginning (*BB*) and ending marks (*BE*) are found in the trace in the case of structured control flow and for unstructured jumps as well (lines 14-17). Therefore, a new scope for a block is created only if it has not already been created for the current function. Otherwise, the current scope pointer is simply set to this block. Since jumps into blocks are possible, they cannot be deleted upon exiting (only the current scope pointer is set), but the whole function scope is deleted when exiting (*FE*).

The other two activities performed in ProcessFunction are the storing of the addresses of local variables in the stack (*D*, lines 10-11) and the processing of the execution actions (*E*, lines 18-19) by the function ProcessAction. ProcessAction takes an action i^j, computes the corresponding dynamic dependence sets of the defined memory addresses and variables and outputs the corresponding slices. The DU^C items are processed for the statement *i* starting with the first one and the so-called *dynamic D/U item i. d'$_k$: U'$_k$* is computed for each step (*for* loop in lines 24-39). Then, the usual operations for computing the dynamic dependence sets are performed [5, 6, 7] (here, *DynDep* stores the actual dependences, while *LS* and *LD* denote the last defining statement number and execution step, respectively). First the used variables are processed (lines 26-33), then the dependence set for the defined variable is computed and output in lines 34-39.

```
program GlobalAlgorithmForC(P, x)
inputs:   P = program
          x = input of P for which dynamic slices should be computed
outputs: dynamic slices for all (x, i^j, V_i) criteria
          (j = 1 ... J, V_i = union of DU^C(i) sets)
globals:  tr = trace element
          sc = scope stack
begin
 1 Store trace TR
 2 TR >> tr
 3 Create global scope on the top of sc
 4 while tr = G(id, addr)
 5   Store id with addr on the top of sc
 6   TR >> tr
   endwhile
 7 ProcessFunction(main)
end
procedure  ProcessFunction(f)
begin
 8 while tr ≠ FE
 9   case tr of
10     D(id, addr) :
11       Store id with addr on the top of sc
12     FB(f) :
13       New function scope on sc
14     BB(block-num) :
15       If ∄ block scope with block-num for actual function on sc then create it
16     BE(outer-num) :
17       Set current scope pointer in sc to outer-num
18     E(i, j) :
19       ProcessAction(i, j)
     endcase
20   TR >> tr
   endwhile
21 Delete all block scopes and the function scope on the top of sc
22 TR >> tr
end
```

Figure 5
Global algorithm for C

Each static D/U variable is resolved with the help of the Resolve function (lines 30, 34). Resolving means finding the memory addresses which the scalar and dereference variables point to at the *j*-th step. Addresses of scalars are looked up in the scope stack by using the usual lookup rules for the function at the top of the stack (lines 41-42 of the Resolve function). The actual addresses of memory dereferences are taken from the trace (*P*), taking into account the fact that the order in which the addresses are dumped into the trace must be the same as the order the static D/U lists the dereference artificial variables (lines 43-45). All other variables (e.g. predicates) will be the same after resolution (lines 46-47).

```
procedure ProcessAction(i, j)
local:    PR = set of predicate variables
begin
23 S := ∅
24 for all (d_k, U_k) ∈ DU^C(i) items
25   PR = ∅
26   for all u_kl ∈ U_k
27     if u_kl = ret(g) for some function g then
28       TR ≫ tr
29       ProcessFunction(g)
       endif
30     u'_kl = Resolve(u_kl)
31     if u'_kl is a predicate variable then PR = PR ∪ {u'_kl}
32     else U'_k = U'_k ∪ {u'_kl}
     endfor
33   U'_k = U'_k ∪ {p}, for which LD(p) = max{LD(r)|r ∈ PR}
34   d'_k = Resolve(d_k)
35   DynDep(d'_k) = ⋃_{u'∈U'_k} (DynDep(u') ∪ {LS(u')})
36   LS(d'_k) = i
37   if d'_k is a predicate variable then LD(d'_k) = j
38   S = S ∪ DynDep(d'_k)
39   Output S as the dynamic slice for (x, i^j, V_i)
  endfor
end
procedure Resolve(x)
begin
40 case x's type of
41   scalar:
42     return x' = lookup variable x in sc and get its address
43   dereference:
44     TR ≫ tr
45     return x' = addr from tr in the form of P(addr)
46   all other artificial:
47     return x' = x
  endcase
end
```

Figure 6
Global algorithm for C (continued)

The other modification for processing one action is that the control dependences are handled in the way described in Section 3.2. Namely, we determine the active predicate by choosing the one from the set U_k that was the most recently defined in line 33 (*PR* contains all static predicate dependences, from which the one with maximal *LD* is taken. ProcessAction also implements the handling of function calls by invoking ProcessFunction recursively, if a function call return variable (*ret*) is found in the D/U. In this case, the trace is processed until the function returns (lines 27-29).

3.5 Implementation and Measurements

We implemented the presented algorithm in a prototype tool and performed experiments about the feasibility of the approach on real-world programs. We used five small to medium size C programs from the open source domain, whose main parameters can be observed in Table 1.

Program	Lines of Code	Statements	Static variables	Scalar variables	Predicate variables	Dereference variables
bcdd	442	78	179	31%	24%	2%
unzoo	2,900	932	1,896	26%	34%	5%
bzip	4,495	2,270	4,184	25%	30%	5%
bc	11,554	3,441	6,898	19%	34%	6%
less	21,488	5,373	10,605	18%	41%	4%

Table 1
Basic program properties

The number of variables found in the program and their types are relevant to the performance of the algorithm. The last four columns of the table overview the total number of static variables in the programs and how their types are distributed. However, the actual computation complexity of the slicing algorithm is mostly determined by the dynamic properties of program elements, which we present in Table 2. The first two columns show the number of test cases we used in our experiments and the average length of the corresponding execution histories, respectively. The next two columns show the average number of dynamic variables (such as memory locations used) and the sizes of the use sets occuring in each step during execution. These two properties are primarily responsible for the actual dependence set sizes and ultimately the space and time costs of the algorithm. The resulting dependence set sizes (the dynamic slices) are shown in the last column in percentage relative to the program size.

Program	Test cases	Avg. actions	Avg. dynamic variables	Avg. use set size	Avg. dependence set size (wrt. program size)
bcdd	5	623.4	34	5.4	18.27%
unzoo	13	169,557.3	1,173	8.9	5.17%
bzip	18	14,245.7	985	8.1	4.35%
bc	49	5,807.3	634	12.6	3.37%
less	14	101,178.5	2,117	6.9	4.80%

Table 2
Dynamic properties of the programs and the slicing algorithm

The length of the execution naturally influences the expected number of dynamic variables. However, the use set sizes and dependence set sizes typically do not depend on this property, but on the logical structure of the program and its computations. Hence, we may conclude that the performance of the algorithm in each step will not be dependent on the length of the execution, which is one of the primary benefits of the method compared to previous approaches.

Conclusions

The dynamic slicing approach presented above does not require a complete dependence graph to be built as a preprocessing step, but instead our algorithm makes use of customized data structures. This has obvious advantages in practical situations and will presumably make the approach scalable and feasible as well. However, other technical issues remain to be solved (for instance, handling the C "long jump" construct) and an optimized version should be developed before making the algorithm available as a prototype tool to other researchers.

Other possible ways of improving the basic algorithms include the idea of trace block summaries [16]. This could be exploited in the implementation for debugging applications; and this is what we plan to investigate in the near future.

References

[1] Weiser, Mark. Program Slicing. IEEE Transactions on Software Engineering, SE-10(4):352–357, 1984

[2] Xu, Baowen, Qian, Ju, Zhang, Xiaofang, Wu, Zhongqiang, and Chen, Lin. A Brief Survey of Program Slicing. ACM SIGSOFT Softw. Eng. Notes, 30(2):1-36, 2005

[3] Agrawal, Hiralal and Horgan, Joseph R. Dynamic Program Slicing. In Proceedings of the ACM SIGPLAN'90 Conference on Programming Language Design and Implementation, number 6 in SIGPLAN Notices, pp. 246-256, White Plains, New York, June 1990

[4] Zhang, Xiangyu, Gupta, Rajiv, and Zhang, Youtao. Cost and Precision Trade-offs of Dynamic Data Slicing Algorithms. ACM Transactions on Programming Languages and Systems, 27(4):631-661, July 2005

[5] Beszédes, Árpád, Gergely, Tamás, and Gyimóthy, Tibor. Graph-less Dynamic Dependence-based Dynamic Slicing Algorithms. In Proceedings of the Sixth IEEE International Workshop on Source Code Analysis and Manipulation (SCAM'06), pp. 21-30, September 2006

[6] Beszédes, Árpád, Gergely, Tamás, Szabó, Zsolt Mihály, Csirik, János, and Gyimóthy, Tibor. Dynamic Slicing Method for Maintenance of Large C Programs. In Proceedings of the 5th IEEE European Conference on Software Maintenance and Reengineering, pp. 105-113, March 2001

[7] Gyimóthy, Tibor, Beszédes, Árpád, and Forgács, István. An Efficient Relevant Slicing Method for Debugging. In Proceedings of ESEC/FSE'99, number 1687 in Lecture Notes in Computer Science, pp. 303-321, Springer-Verlag, September 1999

[8] Beszédes, Árpád, Gyimóthy, Tibor, Lóki, Gábor, Diós, Gergely, and Kovács, Ferenc. Using Backward Dynamic Program Slicing to Isolate Influencing Statements in GDB. In Proceedings of the 2007 GCC Developers' Summit, pp. 21-30, July 2007

[9] Szegedi, Attila and Gyimóthy, Tibor. Dynamic Slicing of Java Bytecode Programs. In Proceedings of the Fifth IEEE International Workshop on Source Code Analysis and Manipulation (SCAM'05), pp. 35-44, IEEE Computer Society, September 2005

[10] Beszédes, Árpád, Faragó, Csaba, Szabó, Zsolt Mihály, Csirik, János, and Gyimóthy, Tibor. Union Slices for Program Maintenance. In Proceedings of the IEEE International Conference on Software Maintenance (ICSM 2002), pp. 12-21, IEEE Computer Society, October 2002

[11] Tip, Frank. A Survey of Program Slicing Techniques. Journal of Programming Languages, 3(3):121-189, September 1995

[12] Horwitz, Susan, Reps, Thomas, and Binkley, David. Interprocedural Slicing using Dependence Graphs. ACM Transactions on Programming Languages and Systems, 12(1):26-61, 1990

[13] Agrawal, Hiralal. Towards Automatic Debugging of Computer Programs. PhD thesis, Purdue University, 1992

[14] Venkatesh, G. A. Experimental results from dynamic slicing of C programs. ACM Transactions on Programming Languages and Systems, 17(2):197-216, March 1995

[15] Korel, Bogdan and Laski, Janusz W. Dynamic Program Slicing. Information Processing Letters, 29(3):155-163, October 1988

[16] Zhang, Xiangyu, Gupta, Rajiv, and Zhang, Youtao. Precise Dynamic Slicing Algorithms. In Proceedings of the 25th International Conference on Software Engineering, pp. 319-329, May 2003

[17] Muchnick, Steven S. Advanced Compiler Design and Implementation. Morgan Kaufmann, 1997

[18] Galli, Tamás, Chiclana, Francisco, Carter, Jenny, and Janicke, Helge. Towards Introducing Execution Tracing to Software Product Quality Frameworks. Acta Polytechnica Hungarica, 11(3):5-24, 2014

Constrained Data-Driven Model-Free ILC-based Reference Input Tuning Algorithm

Mircea-Bogdan Radac[1], Radu-Emil Precup[1], Emil M. Petriu[2]

[1]Department of Automation and Applied Informatics, Politehnica University of Timisoara, Bd. V. Parvan 2, RO-300223 Timisoara, Romania
E-mail: mircea.radac@upt.ro, radu.precup@upt.ro

[2]School of Electrical Engineering and Computer Science, University of Ottawa, 800 King Edward, Ottawa, Ontario, Canada, K1N 6N5
E-mail: petriu@eecs.uottawa.ca

Abstract: This paper proposes a data-driven Iterative Reference Input Tuning (IRIT) algorithm that solves a reference trajectory tracking problem viewed as an optimization problem subjected to control signal saturation constraints and to control signal rate constraints. The IRIT algorithm incorporates an experiment-based stochastic search algorithm formulated in an Iterative Learning Control (ILC) framework in order to combine the advantages of model-free data-driven control and of ILC. The reference input vector's dimensionality is reduced by a linear parameterization. Two neural networks (NNs) trained in an ILC framework are employed to ensure a small number of experiments in the gradient estimation. The IRIT algorithm is validated by two case studies concerning the position control of a nonlinear aerodynamic system. The results prove that the IRIT algorithm offers the significant control system performance improvement by few iterations and experiments conducted on the real-world process. The paper successfully merges the use of ILC in both model-free reference input tuning and NN training.

Keywords: constraints; Iterative Reference Input Tuning algorithm; linear parameterization; mechatronics; neural networks

1 Introduction

The reference trajectory tracking problem can be considered as a reference input design of an initial Control System (CS) with a priori tuned feedback controllers for stability and disturbance rejection. Therefore, the reference trajectory tracking is defined as an open-loop optimal control problem. The data-driven solving of this Optimization Problem (OP) can be carried out in the Iterative Learning Control (ILC) framework, where the sequence of reference input signal samples is updated at each iteration. In this setting, the reference input tuning is regarded as

the optimization variable and the solution to the OP is based on a gradient search algorithm. The gradient information is obtained experimentally without using any knowledge on the process.

In the context of the above features, the proposed approach ensures the data-driven model-free iterative tuning. Hence, our approach shares some similarities with other related approaches to iterative and adaptive model-free control, with several advantages versus the model-based controller tuning [1, 2].

This paper applies ILC to both reference input tuning and neural network (NN) training. The ILC-based solving of optimal control problems is formulated in [3] and [4], time and frequency domain convergence analyses are conducted in [5], the stochastic approximation is treated in [6], and the output tracking is discussed in [7]. The affine constraints are handled in [8] by the transformation of ILC problems with quadratic objective functions (o.f.s) into convex quadratic programs. The system impulse response is estimated in [9] using input/output measurements from previous iterations and next used in a norm-optimal ILC structure that accounts for actuator limitations by means of linear inequality constraints. A learning approach that gives the parameters of motion primitives for achieving flips for quadrocopters is proposed in [10], but it makes use of approximate simple models of the process. Similar formulations with reinforcement learning for policy search using approximate models and signed derivative are given in [11]. NNs applied to ILC in a model-based approach are also reported in [12].

Our recent results given in [13] and [14] are focused on an experiment-based approach to the reference trajectory tracking that takes into account control signal saturation constraints, it employs an Interior Point Barrier (IPB) algorithm, and both simulated and experimental case studies are included. We have applied ILC in [15] to the reference input tuning subject to control signal saturation constraints and control signal rate constraints using max-type quadratic penalty functions, and the results have been validated on a simulated case study related to a nonlinear aerodynamic system. The ILC-based training of NNs has been proposed in [16] in order to reduce the number of experiments of IFT with operational constraints for nonlinear systems, and tested by means of an experimental case study. The IFT with operational constraints applied to data-driven controllers tuned for a reduced sensitivity has been suggested in [17]; an NN identification mechanism has provided the gradient information used in the search algorithm, a perturbation-based approach has been involved in the estimation of second-order derivatives, and the results have been validated by a simulated case study and compared with SPSA and with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update algorithm.

This paper is built upon these results, and the main contribution with respect to the state-of-the-art is an experiment-based Iterative Reference Input Tuning (IRIT) algorithm that solves the constrained reference trajectory tracking. This is advantageous because: the dimensionality of the reference input vector is reduced

by a linear parameterization that enables cost-effective controller designs and implementations, the NN-based identification mechanism applied to the nonlinear CS leads to a simple, effective and general IRIT algorithm with a reduced number of experiments, the involvement of ILC in IRIT and NN training makes our approach a special case of supervised learning according to the relationships discussed in [14]. This strong involvement determines our IRIT-based CSs to benefit of the advantages of ILC highlighted in a mechatronics application. The proposed solution is model-free as opposed to the model-based solutions for constrained ILC presented in [8], [18], [19].

The paper is organized as follows. The next section presents the formulation of the problem that concerns the reference trajectory tracking problem solved in the data-driven optimal ILC framework. Section 3 deals with the model-free estimation of o.f.'s gradient. Section 4 proposes the model-free constrained optimal control problem and gives the formulation of the IRIT algorithm. Section 5 motivates the use of the NN-based approach in gradient estimation. Section 6 validates the IRIT algorithm by two simulated case studies that deal with the angular position control of a nonlinear aerodynamic system. The results and their discussion convincingly validate the new IRIT algorithm. The conclusions are highlighted in Section 7.

2 Data-driven Approach to Reference Trajectory Tracking

2.1 Problem Formulation

The CS is characterized by the discrete time Linear Time-Invariant (LTI) Single Input-Single Output (SISO) model

$$y(\boldsymbol{\rho}, r, k) = T(\boldsymbol{\rho}, q^{-1})r(k) + S(\boldsymbol{\rho}, q^{-1})v(k), \tag{1}$$

where k is the discrete time argument, $y(k)$ is the process output sequence, $r(k)$ is the reference input sequence, $v(k)$ is the zero-mean stationary and bounded stochastic disturbance input sequence acting on the process output and accounting for various types of load or measurement disturbances, $S(\boldsymbol{\rho}, q^{-1})$ is the sensitivity function, $T(\boldsymbol{\rho}, q^{-1})$ is the complementary sensitivity function

$$S(\boldsymbol{\rho}, q^{-1}) = 1/[1 + P(q^{-1})C(\boldsymbol{\rho}, q^{-1})], T(\boldsymbol{\rho}, q^{-1}) = 1 - S(\boldsymbol{\rho}, q^{-1}), \tag{2}$$

$P(q^{-1})$ is the process transfer function (t.f.), $C(\boldsymbol{\rho}, q^{-1})$ is the controller t.f., which is parameterized by the parameter vector $\boldsymbol{\rho}$ that contains the tuning parameters of the controller, and q^{-1} is the one step delay operator. The parameter vector $\boldsymbol{\rho}$ will be omitted as follows in certain equations for the sake of simplicity.

An ILC framework to describe the reference trajectory tracking problem is introduced using the lifted form (or super vector) representation. For a relative degree n of the closed-loop CS t.f. $T(\boldsymbol{\rho}, q^{-1})$, the lifted form representation for an N samples experiment length and the matrices in the deterministic case are

$$\begin{aligned}&\mathbf{Y} = \mathbf{T}\,\mathbf{R} + \mathbf{Y}_0, \\ &\mathbf{Y} = [y(n) \quad y(n+1) \quad ... \quad y(N-1)]^T, \mathbf{R} = [r(0) \quad r(1) \quad ... \quad r(N-n-1)]^T, \\ &\mathbf{Y}_0 = [y_{10} \quad y_{20} \quad ... \quad y_{(N-n)0}]^T, \mathbf{T} = \begin{bmatrix} t_1 & 0 & ... & 0 \\ t_2 & t_1 & ... & 0 \\ ... & ... & ... & ... \\ t_{N-n} & t_{N-n-1} & ... & t_1 \end{bmatrix},\end{aligned} \tag{3}$$

where $\mathbf{R}$ is the reference input vector that contains the reference input sequence over the time interval $0 \le k \le N-n-1$, $\mathbf{Y}$ is the controlled output vector, t_i is the i^{th} impulse response coefficient of $T(\boldsymbol{\rho}, q^{-1})$, $\mathbf{T}$ is a lower-triangular Toeplitz matrix, $\mathbf{Y}_0$ is the free response of the CS due to nonzero initial conditions and trial-repetitive disturbances, and T indicates matrix transposition. Zero initial conditions are assumed without loss of generality, and the tracking error vector $\mathbf{E}$

$$\mathbf{E} = \mathbf{Y} - \mathbf{Y}^d = \mathbf{T}\,\mathbf{R} - \mathbf{Y}^d, \tag{4}$$

where $\mathbf{Y}^d$ is the desired reference trajectory vector generated from the desired process output $y^d(k)$. Equation (4) shows that knowledge on $\mathbf{T}$ would provide the optimal solution which makes the tracking error zero, i.e., $\mathbf{R} = \mathbf{T}^{-1}\mathbf{Y}^d$. However, $\mathbf{T}$ can be ill-conditioned, and this matrix is always subject to measurement errors; therefore, $\mathbf{T}^{-1}$ cannot be used. A solution to the iterative estimation of $\mathbf{T}$ in an ILC framework is given in [9].

The control objective is expressed as the following OP that involves the expected normalized norm of the tracking error:

$$\begin{aligned}&\mathbf{R}^* = \arg\min_{\mathbf{R}} J(\mathbf{R}) = E\{\frac{1}{N} \| \mathbf{E}(\mathbf{R}) \|_2^2\} \\ &\text{subject to system dynamics (1) and to some operational constraints,} \\ &J(\mathbf{R}) = E\{\frac{1}{N}(\mathbf{T}\,\mathbf{R} - \underbrace{\mathbf{Y}^d}_{\mathbf{M}})^T(\mathbf{T}\,\mathbf{R} - \mathbf{Y}^d)\} = E\{\frac{1}{N}(\mathbf{R}^T\,\mathbf{Q}\,\mathbf{R} + 2\mathbf{q}\,\mathbf{R} + \alpha)\},\end{aligned} \tag{5}$$

the deterministic formulation of the o.f. $J(\mathbf{R})$ is quadratic with respect to $\mathbf{R}$, where $\mathbf{Q} = \mathbf{T}^T\mathbf{T}$ is a positive semi-definite matrix, $\mathbf{q} = \mathbf{M}^T\,\mathbf{T}$, and $\alpha = \mathbf{M}^T\mathbf{M}$. A gradient descent approach to iteratively solve (5) is

$$\mathbf{R}_{j+1} = \mathbf{R}_j - \gamma_j \tilde{\mathbf{H}}_{\mathbf{R}}^{-1} est\{\frac{\partial J}{\partial \mathbf{R}}\Big|_{\mathbf{R}=\mathbf{R}_j}\}, \tag{6}$$

where the subscript j is the iteration or trial index, $est\{\frac{\partial J}{\partial \mathbf{R}}\Big|_{\mathbf{R}=\mathbf{R}_j}\}$ is the estimate of the gradient of the o.f. with respect to the reference input vector samples, $\tilde{\mathbf{H}}_{\mathbf{R}}^{-1}$ is a Gauss-Newton approximation of the Hessian of the o.f., typically given by a BFGS update algorithm, and γ_j is the step size of the update law (6). When no model information is used for the choice of γ_j in order to guarantee the convergence of the search algorithm [3-9], a small enough value of the step size will usually ensure the convergence. This renders our approach a truly model-free one.

The stochastic convergence of ILC algorithms treated in [5, 6] is related to two imposed stochastic convergence conditions: the estimated o.f.'s gradient is unbiased, and the step size sequence $\{\gamma_j\}_{j\geq 0}$ converges to zero but not too fast. Constant values of the step size can be set in practical experiments, where the theoretical convergence is not targeted and few iterations are aimed. The deterministic formulation of the OP (5) will be employed in the next sections.

2.2 Reducing the Dimensionality of the Reference Input Vector

Using the reference input vector tuning as in [13, 14], the dimension of the search space is usually high, of about hundreds of samples of the reference input signal to be optimized. A linear transformation is considered in order to reduce the reference input vector dimension. A common linear parameterization can be a polynomial fit of a certain order, a Fourier fit or a Gaussian fit, all of them linear in the parameters. For an h_r degree polynomial for which

$$r(k) = \sum_{i=0}^{h_r} k^i \theta_i,\ 0 \leq k \leq N-n-1, \tag{7}$$

the reference input vector is expressed according to the linear transformation

$$\begin{aligned} &\mathbf{R} = \boldsymbol{\Gamma}\,\boldsymbol{\theta}, \\ &\boldsymbol{\Gamma} = [\Gamma_{ij}]_{i=1\ldots N-n,\, j=1\ldots h_r+1},\ \Gamma_{i1} = 1, \\ &\boldsymbol{\theta} = [\theta_0 \quad \ldots \quad \theta_{h_r}]^T. \end{aligned} \tag{8}$$

The OP given in (5) is also quadratic in $\boldsymbol{\theta}$ by the virtue of the linear transformation.

This reduction of the OP dimension may be useful for several reasons. The convergence to a local minimum of the o.f. can be accelerated; in addition, the ill-conditioning of the BFGS update algorithm can be avoided. The idea of reducing the dimension of the learning space in an ILC formulation is also treated in several

approaches in [20-23]. These approaches range from decomposing the reference signal using different types of basis functions, to down sampling the reference signals. However, this problem handling is considered in a model-based context and not in a model-free one as in our case.

3 Model-free Estimation of the Gradient

Using (8) in (5), the o.f. will be quadratic with respect to $\boldsymbol{\theta}$. Hence

$$J(\boldsymbol{\theta}) = \frac{1}{N}(\boldsymbol{\theta}^T \boldsymbol{\Gamma}^T \mathbf{Q} \boldsymbol{\Gamma} \boldsymbol{\theta} + 2\mathbf{q} \boldsymbol{\Gamma} \boldsymbol{\theta} + \alpha) \cdot \tag{9}$$

A gradient search is performed to find the minimum of this function. The analytic solution is not desired because it depends on the matrices $\mathbf{Q}$ and $\mathbf{q}$ that depend on the unknown $\mathbf{T}$. The gradient search using a Gauss-Newton approximation of the Hessian of the new o.f. $J(\boldsymbol{\theta})$ is

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - \gamma_j \tilde{\mathbf{H}}_{\boldsymbol{\theta}}^{-1} \left.\frac{\partial J}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j} \cdot \tag{10}$$

A model-free approach to the gradient estimation is given in [15] and reformulated here. From (9), using the matrix derivation rules and the fact that $\boldsymbol{\Gamma}^T \mathbf{Q} \boldsymbol{\Gamma}$ is symmetric by the virtue of $\mathbf{Q} = \mathbf{T}^T \mathbf{T}$ being symmetric, the gradient of $J(\boldsymbol{\theta})$ with respect to the parameter vector $\boldsymbol{\theta}$ will be

$$\left.\frac{\partial J}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j} = \frac{2}{N}(\boldsymbol{\Gamma}^T \mathbf{T}^T \mathbf{T} \boldsymbol{\Gamma} \boldsymbol{\theta} + \boldsymbol{\Gamma}^T \mathbf{T}^T \mathbf{M}) = \frac{2}{N} \boldsymbol{\Gamma}^T \mathbf{T}^T (\mathbf{T} \boldsymbol{\Gamma} \boldsymbol{\theta} + \mathbf{M}) \cdot \tag{11}$$

But $\mathbf{T} \boldsymbol{\Gamma} \boldsymbol{\theta} + \mathbf{M} = \mathbf{E}$, and the gradient of $J(\boldsymbol{\theta})$ in the deterministic case at each iteration *j* is finally expressed as

$$\left.\frac{\partial J}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j} = \frac{2}{N} \boldsymbol{\Gamma}^T \mathbf{T}^T \mathbf{E}_j \cdot \tag{12}$$

In (12), $\frac{2}{N} \mathbf{T}^T \mathbf{E}_j$ is actually the gradient of $J(\mathbf{R})$ from (5) with respect to $\mathbf{R}$. Therefore, using (8)

$$\frac{\partial J(\mathbf{R}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{2}{N} \frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \frac{\partial J(\mathbf{R})}{\partial \mathbf{R}} = \frac{2}{N} \boldsymbol{\Gamma}^T \frac{\partial J(\mathbf{R})}{\partial \mathbf{R}} \cdot \tag{13}$$

Equation (13) can be interpreted as the chain derivation rule of the function $J(\mathbf{R}(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$, and it will be used later in the paper. Equation (12) suggests that the gradient information can be obtained either by an experimentally measured $\mathbf{T}$ or by using a special gradient experiment at each iteration.

The second approach is preferred in our case and it is next presented. Different solutions to the feed-forward optimal control design problem using finite-differences approximations of the gradient by experiments with perturbed parameters are presented in [24, 25].

The successive updates (10) for the parameterized reference input trajectory are performed in the vicinity of the current iteration reference input trajectory. The linearity assumption and operation can therefore be justified in this case. As we see from (12), the gradient vector is obtained experimentally driving the closed-loop CS in non-nominal operating regimes because the current iteration error $\mathbf{E}_j$ is used as a reference input in the gradient estimation scheme according to [15]. For a linear system this does not affect the quality of the gradient information although it may affect the nominal operation of the CS. In order to allow for near-nominal experimenting regimes to be used with linear systems and to further extend the applicability of the IRIT algorithm to nonlinear systems a perturbation-based approach is proposed to obtain the gradient information near the nominal trajectory. This idea stems from [26], and it is a modified version of the algorithm used in [15]. The model-free gradient estimation algorithm consists of the following steps:

Step A. Record the tracking error at the current iteration in the vector $\mathbf{E}_j$.

Step B. Define the reversed vector $rev(\mathbf{E}_j)$

$$\begin{aligned} &rev(\mathbf{E}_j) = rev([e_j(0) \quad ... \quad e_j(N-n-1)]^T) = [e_j(N-n-1) \quad ... \quad e_j(0)]^T, \\ &e_j(k) = y(n+k) - y^d(n+k), 0 \le k \le N-n-1. \end{aligned} \tag{14}$$

Step C. Apply $\mathbf{\Gamma\theta}_j + \mu \times rev(\mathbf{E}_j)$ as a reference input to the CS and obtain the output vector $\mathbf{Y}_G = \mathbf{T}(\mathbf{\Gamma\theta}_j + \mu \times rev(\mathbf{E}_j))$ where the subscript G stands for "gradient". The scalar coefficient μ is chosen such that the perturbed term $\mu \times rev(\mathbf{E}_j)$ represents only a small deviation around the nominal reference input trajectory $\mathbf{R}_j = \mathbf{\Gamma\theta}_j$.

Step D. Since $\mathbf{Y}_j = \mathbf{T\Gamma\theta}_j$ is known from the nominal experiment, obtain $\mathbf{\Gamma}^T\mathbf{T}^T\mathbf{E}_j$ as

$$\mathbf{\Gamma}^T\mathbf{T}^T\mathbf{E}_j = \frac{1}{\mu}\mathbf{\Gamma}^T rev(\mathbf{Y}_G - \mathbf{Y}_j), \tag{15}$$

and use (15) in (12) to get the gradient $\left.\frac{\partial J}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j}$.

The choice of the parameter μ can be done automatically such that the nominal reference input is not perturbed too much in amplitude.

4 Dealing with Control Signal Saturation and Control Signal Rate Constraints

The operational constraints regarding the saturation of actuators, the saturation of the control signal rate or the bounds on the state variables of the process are very important in many real-world CS applications. Different numerical algorithms can be employed in model-based approaches to solve the OP (5) for such systems. However, a model-free approach is presented as follows.

The lifted form representations allow the expression of a particular form of the OP that can be of interest. Assuming the deterministic case, let $\mathbf{S}_{ur} \in \Re^{(N-m)\times(N-m)}$ be the lifted map that corresponds to the t.f. $S_{ur}(q^{-1}) = C(q^{-1})S(q^{-1})$, where $\Re$ is the set of real numbers. Using the notation m for the relative degree of $S_{ur}(q^{-1})$, $m \le n$, the lifted form representations are [15]

$$\begin{aligned}
&\mathbf{U} = [u(m) \quad u(m+1) \quad ... \quad u(N-1)]^T, \mathbf{R} = [r(0) \quad r(1) \quad ... \quad r(N-m-1)]^T, \\
&\mathbf{U} = \begin{bmatrix} s_1 & 0 & ... & 0 \\ s_2 & s_1 & ... & 0 \\ ... & ... & ... & ... \\ s_{N-m} & s_{N-m-1} & ... & s_1 \end{bmatrix} \cdot \mathbf{R} = \mathbf{S}_{ur} \cdot \mathbf{R}, \Delta\mathbf{U} = [\Delta u(1) \quad \Delta u(2) \quad ... \quad \Delta u(N-n)]^T \\
&= [u(m) - 0 \;\; u(m+1) - u(m) \;\; ... \;\; u(m+N-n-1) - u(m+N-n-2)]^T \\
&= [s_1 r(0) s_2 r(0) + s_1 r(1) - s_1 r(0) \;\; ... \;\; s_{N-n} r(0) + ... + s_1 r(N-n-1) - s_{N-n-1} r(0) \\
&- ... - s_1 r(N-n-2)]^T \\
&= \begin{bmatrix} s_1 & s_2 & s_3 & ... & s_{N-n} \\ 0 & s_1 & s_2 & ... & s_{N-n-1} \\ 0 & 0 & s_1 & ... & ... \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & s_1 \end{bmatrix}^T \cdot \mathbf{R} - \begin{bmatrix} 0 & s_1 & s_2 & ... & s_{N-n-1} \\ 0 & 0 & s_1 & ... & s_{N-n-2} \\ 0 & 0 & 0 & ... & ... \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T \cdot \mathbf{R} = \mathbf{S}_{\Delta ur} \cdot \mathbf{R},
\end{aligned} \tag{16}$$

where $\mathbf{R} \in \Re^{(N-m)\times 1}$ is a vector of greater length than in (3), for which $\mathbf{R} \in \Re^{(N-n)\times 1}$. Therefore, a truncation of $\mathbf{S}_{ur}$ corresponding to the leading principal minor of size $N-n$ is considered such that $\mathbf{S}_{ur} \in \Re^{(N-n)\times(N-n)}$ because we need the same $\mathbf{R}$ of size $N-n$ to be tuned, and this in turn will allow only $N-n$ (out of $N-m$) constraints imposed to $\mathbf{U}$, and the affine constraints $\mathbf{U}_{\min} \le \mathbf{U}(\mathbf{R}) \le \mathbf{U}_{\max}$ and $\Delta\mathbf{U}_{\min} \le \Delta\mathbf{U}(\mathbf{R}) \le \Delta\mathbf{U}_{\max}$ are imposed to $\mathbf{R}$.

The OP, which ensures the reference trajectory tracking with control signal constraints and with control signal rate constraints is expressed as

$$\begin{aligned}
&\boldsymbol{\theta}^* = \arg\min_{\mathbf{R}} \frac{1}{N}(\boldsymbol{\theta}^T \boldsymbol{\Gamma}^T \mathbf{Q}\boldsymbol{\Gamma}\boldsymbol{\theta} + 2\mathbf{q}\boldsymbol{\Gamma}\boldsymbol{\theta} + \alpha),\\
&\text{subject to dynamics (1), and to } \tilde{\mathbf{S}}\boldsymbol{\Gamma}\boldsymbol{\theta} \le \tilde{\mathbf{U}} \text{ and } \tilde{\mathbf{S}}_\Delta \boldsymbol{\Gamma}\boldsymbol{\theta} \le \Delta\tilde{\mathbf{U}}, \text{ where}\\
&\tilde{\mathbf{S}} = [\mathbf{S}_{ur}^T \quad -\mathbf{S}_{ur}^T]^T \in \Re^{2(N-n)\times(N-n)}, \tilde{\mathbf{S}}_\Delta = [\mathbf{S}_{\Delta ur}^T \quad -\mathbf{S}_{\Delta ur}^T]^T \in \Re^{2(N-n)\times(N-n)}\\
&\tilde{\mathbf{U}} = [\mathbf{U}_{\max}^T \quad -\mathbf{U}_{\min}^T]^T \in \Re^{2(N-n)\times 1}, \Delta\tilde{\mathbf{U}} = [\Delta\mathbf{U}_{\max}^T \quad -\Delta\mathbf{U}_{\min}^T]^T \in \Re^{2(N-n)\times 1},\\
&\mathbf{U}_{\min} = [u_{\min}^1 \quad u_{\min}^2 \quad \ldots \quad u_{\min}^{N-n}]^T, \mathbf{U}_{\max} = [u_{\max}^1 \quad u_{\max}^2 \quad \ldots \quad u_{\max}^{N-n}]^T.
\end{aligned} \tag{17}$$

A solver for this type of problems in the deterministic case is the IPB algorithm [8, 14]. As we have shown in [14] for inequality constraints concerning only the control signal saturation, the constrained OP is transformed into an unconstrained OP by the use of the penalty functions. The logarithmic barrier penalty function grows unbounded as the constraints are close to being violated and in the stochastic framework this is always the case. A solution to overcome this problem is given in [27, 28], but with quadratic penalty functions. We propose the following augmented o.f. that accounts for inequality constraints concerning the control signal saturation and the control signal rate:

$$\tilde{J}_{p_j}(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + p_j[\underbrace{\frac{1}{2}\sum_{h=1}^{c}\{[\max\{0,-(\underbrace{\tilde{u}_h - \tilde{\mathbf{s}}_h^T\boldsymbol{\Gamma}\boldsymbol{\theta}}_{q_h(\boldsymbol{\theta})})\}]^2\}}_{\phi(\boldsymbol{\theta})} + \underbrace{\frac{1}{2}\sum_{h=1}^{c}\{[\max\{0,-q_h(\boldsymbol{\Gamma}\boldsymbol{\theta})\}]^2\}}_{\Delta\phi(\boldsymbol{\theta})}], \tag{18}$$

where the positive and strictly increasing sequence of penalty parameters $\{p_j\}_{j\ge 0}$, $p_j \to \infty$, guarantees that the minimum of the sequence of augmented o.f.s $\{\tilde{J}_{p_j}(\boldsymbol{\theta})\}_{j\ge 0}$ will converge to the solution to the constrained OP (17), h, $h = 1...c$, is the constraint index, $q_h(\boldsymbol{\theta}) > 0$ is h^{th} constraint, $\tilde{u}_h$ is h^{th} element of $\tilde{\mathbf{U}}$, and $\tilde{\mathbf{s}}_h^T$ is h^{th} row of $\tilde{\mathbf{S}}$. The OP (17) is solved using a stochastic approximation algorithm that makes use of the experimentally obtained gradient of $\tilde{J}_{p_j}(\boldsymbol{\theta})$. For practical applications, where stochastic convergence is not targeted and a few number of iterations is desired, the penalty parameters can be chosen as $p_j = p = \text{const}$.

The quadratic penalty functions $\phi(\boldsymbol{\theta})$ and $\Delta\phi(\boldsymbol{\theta})$ in (18) corresponding to the control saturation and control rate constraints use the *max* function, which in this case is non-differentiable only at zero. Given that $\phi(\boldsymbol{\theta})$ and $\Delta\phi(\boldsymbol{\theta})$ are Lipschitz and non-differentiable at a set of points of zero Lebesgue measure, the algorithm visits the zero-measure set with probability zero when a normal distribution for the noise is assumed [27]. Therefore

$$\frac{\partial[\max\{0,-q_h(\boldsymbol{\theta})\}]^2}{\partial r(i)} = -2\max\{0,-q_h(\boldsymbol{\theta})\}\cdot\frac{\partial q_h(\boldsymbol{\theta})}{\partial r(i)}. \tag{19}$$

Using the gradient of $\phi(\mathbf{R})$ with respect to $\mathbf{R}$ given in [15], the linear transformation (8) and the chain derivation rule with respect to vectors lead to the expression of the gradient of $\phi(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:

$$\frac{\partial \phi(\mathbf{R}=\boldsymbol{\Gamma}\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\Gamma}^T \mathbf{S}_{ur}^T \boldsymbol{\zeta}(\boldsymbol{\theta}). \tag{20}$$

The gradient of $\Delta\phi(\mathbf{R}=\boldsymbol{\Gamma}\boldsymbol{\theta})$ in (18) with respect to the $N-n$ elements of $\mathbf{R}$ (from 0 to $N-n-1$) is given in [15], and the linear transformation (8), next the chain derivation rule with respect to vectors lead to the expression of the gradient of $\Delta\phi(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:

$$\frac{\partial \Delta\phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\Gamma}^T (\mathbf{M}_1 - \mathbf{M}_2) \Delta\boldsymbol{\zeta}(\boldsymbol{\theta}) \cdot \tag{21}$$

Using (19), (20) and (21), the expression of the gradient of the o.f. (18) at the current iteration *j* is

$$\left.\frac{\partial \tilde{J}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j} = \frac{2}{N} \boldsymbol{\Gamma}^{\mathrm{T}} \mathbf{T}^T \mathbf{E}_j + p_j \{\boldsymbol{\Gamma}^{\mathrm{T}} \mathbf{S}_{ur}^T [\underbrace{\boldsymbol{\zeta}(\boldsymbol{\theta}_j) + \Delta\boldsymbol{\zeta}(\boldsymbol{\theta}_j) - \Delta\bar{\boldsymbol{\zeta}}(\boldsymbol{\theta}_j)}_{\boldsymbol{\psi}(\boldsymbol{\theta}_j)}]\}, \tag{22}$$

where $\boldsymbol{\theta} \in \Re^{(h+1)\times 1}$, $\boldsymbol{\zeta}(\boldsymbol{\theta}_j)$ is considered in [15] as $\boldsymbol{\zeta}(\boldsymbol{\theta})$ is seen as a function of $\boldsymbol{\theta}$ via the transformation (8), $\Delta\boldsymbol{\zeta}(\boldsymbol{\theta}_j)$ is considered in [16] as $\Delta\boldsymbol{\zeta}(\boldsymbol{\theta})$ is seen as a function of $\boldsymbol{\theta}$ via the transformation (8), and $\Delta\bar{\boldsymbol{\zeta}}(\boldsymbol{\theta}_j)$ is a one step ahead vector of dimension $N-n$:

$$\Delta\bar{\boldsymbol{\zeta}}(\boldsymbol{\theta}_j) = [\Delta\boldsymbol{\zeta}(\boldsymbol{\theta},2) \quad ... \quad \Delta\boldsymbol{\zeta}(\boldsymbol{\theta},N-n) \quad 0]^T. \tag{23}$$

As shown in [16], the matrix term in the expression of $\frac{\partial \Delta\phi(\mathbf{R})}{\partial \mathbf{R}}$ and the structure of the matrices justifies the use of a single gradient experiment, with $rev(\boldsymbol{\zeta}+\Delta\boldsymbol{\zeta}-\Delta\bar{\boldsymbol{\zeta}}) = rev(\boldsymbol{\psi}(\boldsymbol{\theta}_j))$ injected as the reference input to the CS, taking advantage of the dimensionality of the map $\mathbf{S}_{ur}^T$. The same approach will be used as in the gradient estimation algorithm given in Section 3 in order to constrain the evolution of the dynamic system in the vicinity of the nominal trajectory.

The feasibility is not preserved during the tuning since the constraints are weighted in the o.f. only when they are violated. The feasibility is not a problem because our approach allows the initialization of a solution that is not initially feasible. But this causes the nonlinear behaviour when the constraints are active and therefore it is not recommended. However, in the long-term run, as the sequence $\{p_j\}_{j\geq 0}$ increases, the gradient due to the constraints that are violated is decisive, and the reference trajectory tracking objective is neglected with the expense of fulfilling OP's constraints. The constraints are active and they vary only subjected to the random effects of the noise affecting the closed-loop system.

For non-minimum-phase systems, the iterative reference input update (6) or (10) may lead to unbounded growth of the reference input's amplitude because this update will try to compensate for the non-minimum-phase character of the system response. In terms of the analytical solution to the reference trajectory tracking problem $\mathbf{R} = \mathbf{T}^{-1}\mathbf{Y}^{d}$, this corresponds to filtering $\mathbf{Y}^{d}$ through the inverse of an unstable map $\mathbf{T}$. We propose three solutions to this problem, briefly outlined as follows. The first solution requires that the desired trajectory should have a non-minimum-phase character. The second solution uses a regularization factor in the definition of the original o.f. given in (5), for example as the weighted norm of the reference input $\lambda \| \mathbf{R} \|_2^2$, where $\lambda > 0$ is a scalar weight. This will balance the o.f. and the growth of the amplitude of the reference input will be limited. The third solution is based on the fact that the introduction of constraints on the control signal and on the control signal rate will indirectly limit the amplitude of the reference input as an unbounded reference input will generate an unbounded control input signal. Therefore, out approach indirectly solves the reference trajectory tracking problem for non-minimum-phase systems by taking into account the control signal constraints.

Our IRIT algorithm consists of the following steps:

Step S1. Start with the initial guess of $\mathbf{R}$. Calculate the regressor $\mathbf{\Gamma}$, perform data normalization on the regressor, and fit the initial $\boldsymbol{\theta}$ using a least squares algorithm. Choose the upper and lower bounds for the control signal, the upper and lower bounds for the control signal rate and generate the desired reference trajectory vector $\mathbf{Y}^{d}$. Choose the tolerances tol_N for stopping the stochastic search algorithm. Choose the sequence $\{p_j\}_{j\geq 0}$ and γ_0. Set the iteration index for $\boldsymbol{\theta}$ and $\{p_j\}_{j\geq 0}$ to $j = 0$.

Step S2. Conduct the normal experiment with the current $\boldsymbol{\theta}_j$. Evaluate the o.f. $\tilde{J}(\boldsymbol{\theta}_j)$ with $\boldsymbol{\theta}_j = \boldsymbol{\theta}$ in (17), record the current tracking error $\mathbf{E}_j$, and compute the vector variables $\zeta, \Delta\zeta, \Delta\bar{\zeta}$ as shown in [15].

Step S3. Conduct the first gradient experiment according to the approach given in Section 3 to find the first term in the gradient of the augmented o.f. given in (22), namely $\frac{2}{N}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{T}^{T}\mathbf{E}_j$.

Step S4. Conduct the second gradient experiment in the same way as descried in Section 3. The reversed vector $\boldsymbol{\psi}(\boldsymbol{\theta}_j)$ is used as the reference input applied to the real-world CS to find the second term in the gradient of the augmented o.f. given in (22), namely $\mathbf{S}_{ur}^{T}\boldsymbol{\psi}(\boldsymbol{\theta}_j)$, after which the expression $p_j\{\mathbf{\Gamma}^{T}\mathbf{S}_{ur}^{T}\boldsymbol{\psi}(\boldsymbol{\theta}_j)\}$ is obtained in a straightforward manner because $\mathbf{\Gamma}$ is known.

Step S5. Estimate the gradient using (22), and calculate the next reference input sequence using

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - \gamma_j est\{\frac{\partial \tilde{J}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j}\}. \tag{24}$$

Step S6. If the gradient search has converged in terms of the gradient of the augmented o.f. less then a constant, $est\{\frac{\partial \tilde{J}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j}\} < tol_N$, stop the algorithm. Otherwise, calculate $\mathbf{R}_{j+1} = \boldsymbol{\Gamma}\,\boldsymbol{\theta}_{j+1}$, set $j = j+1$, and next jump to step *S2*.

5 Neural Network-based Gradient Estimation Mechanism

Each iteration in the algorithm given in the previous section requires a normal experiment with the current parameterized reference input. After the normal experiment, the gradient experiments require running perturbed trajectories in the vicinity of the nominal trajectories. These perturbed trajectories are obtained for perturbed reference inputs with small amplitude signals according to the previous sections and to [13-17]. In order to avoid conducting gradient experiments on the real-world CS, a simulation-based mechanism can be used, with identified models instead of the real-world CS. These models are only valid in the vicinity of the current iteration nominal trajectories. No additional experiments are required to collect data in a wide operating range for identification purposes so these models have scope only within the current iteration.

In order to extend the applicability of this approach to smooth nonlinear systems that can be well approximated by linear systems near some operating points, NN-based models can be used for the identification purposes. Our approach has two advantages. First, the closed-loop CS behaviour is usually of low-pass type; therefore, the models usually have simple dynamics. Second, the numerical differentiation issues which occur in noisy environments will be mitigated by our approach. Linear models could have been used as well for gradient estimation since they can be considered as particular cases of nonlinear ones.

Let the nonlinear dynamic maps from the reference input to the controlled output M_{ry} and from the reference input to the control input M_{ru} are supposed to be characterized by the following nonlinear autoregressive exogenous (NARX) models [16, 17]:

$$y(k) = M_{ry}(y(k-1),...,y(k-n_y),r(k-1),...,r(k-n_{ry})),$$
$$u(k) = M_{ru}(u(k-1),...,u(k-n_u),r(k-1),...,r(k-n_{ru})). \tag{25}$$

A more compact representation that takes advantage of the supervector notation is $\overline{\mathbf{Y}} = M_{ry}(\mathbf{R})$ and $\overline{\mathbf{U}} = M_{ru}(\mathbf{R})$. The current iteration trajectories $\{\mathbf{R}_j, \mathbf{U}_j, \mathbf{Y}_j\}$ from the normal experiment are used to identify M_{ry} and M_{ru}, respectively. Using (15) from the model-free gradient estimation algorithm (given in Section 3) in (22), an estimate of the gradient of the augmented o.f. is expressed as

$$est\{\frac{\partial \tilde{J}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_j}\} = \frac{2}{\mu_{\mathbf{Y}}}\boldsymbol{\Gamma}^{\mathrm{T}} rev(\overline{\mathbf{Y}}_{G_j} - \overline{\mathbf{Y}}_j) + p_j\{\frac{1}{\mu_{\mathbf{U}}}\boldsymbol{\Gamma}^{\mathrm{T}} rev(\overline{\mathbf{U}}_{G_j} - \overline{\mathbf{U}}_j)\},$$
$$\overline{\mathbf{Y}}_{G_j} = M_{ry}(\mathbf{R}_j + \mu_{\mathbf{Y}} rev(\mathbf{E}_j)), \overline{\mathbf{Y}}_j = M_{ry}(\mathbf{R}_j), \tag{26}$$
$$\overline{\mathbf{U}}_{G_j} = M_{ru}(\mathbf{R}_j + \mu_{\mathbf{U}} rev(\boldsymbol{\Psi}_j)), \overline{\mathbf{U}}_j = M_{ru}(\mathbf{R}_j),$$

where $\boldsymbol{\psi}(\boldsymbol{\theta}_j)$ is defined in (22), and $\mu_{\mathbf{Y}}, \mu_{\mathbf{U}}$ are scaling factors chosen such that the perturbations are of small amplitude with respect to the current iteration reference input. The superposition principle invoked here is expected to work for small amplitude perturbations of the nominal trajectories at the current iteration.

We are using a feed-forward NN architecture that consists of one hidden layer with a hyperbolic tangent activation function and a single linear neuron. The input-output map is [16, 17]

$$\hat{y}(k+1) = \mathbf{W}^T(k)\boldsymbol{\sigma}(\mathbf{V}(k),\mathbf{x}(k)), \tag{27}$$

where $\mathbf{W}^T = [w_0 \quad w_1 \quad ... \quad w_H] \in \mathbf{R}^{H+1}$ is the vector of output layer weights, $\boldsymbol{\sigma}^T = [1 \quad \sigma_1(\mathbf{V}_1^T\mathbf{x}) \quad ... \quad \sigma_H(\mathbf{V}_H^T\mathbf{x})]$ is the vector of hidden layer neurons outputs, with the hyperbolic activation tangent activation functions $\sigma_m(x) = \tanh(x), m = 1...H$, the first term in $\boldsymbol{\sigma}$ corresponds to the bias of the output neuron, and each hidden layer neuron is parameterized by its vector of weights $(\mathbf{V}^m)^T = [v_m^0 \quad v_m^1 \quad ... \quad v_m^{nu}] \in \mathbf{R}^{nu+1}$, $m = 1...H$, which multiplies the input vector $\mathbf{x}^T = [x_0 \quad x_1 \quad ... \quad x_{nu}]$.

Treating the NN as a nonlinear multi input-multi output dynamical system considered in the iteration domain [16, 17]:

$$\mathbf{W}_{j+1} = \mathbf{W}_j + \mathbf{u}_j^w,$$
$$\mathbf{V}_{j+1}^i = \mathbf{V}_j^i + \mathbf{u}_j^{v^i}, i = 1...H, \tag{28}$$
$$\mathbf{Y}_j(k+1) = \mathbf{W}_j^T\boldsymbol{\sigma}(\mathbf{V}_j^i, \mathbf{x}(k)), k = 0...N,$$

where j is the iteration index, the dynamical system (28) is transformed into a static map from the inputs to the outputs, and the batch training of the NN can be

regarded as a supervised learning approach, that aims the minimization of the tracking error $\mathbf{E}_j = \mathbf{Y}_j - \mathbf{Y}^d$ referred to also as training error.

As shown in [16, 17], the input at each iteration is derived in the framework of norm-optimal ILC as the solution to an OP, is next transformed into another OP by a Taylor series expansion. The optimal vector solution to this OP consists of the increments of the NN weights, expressed as update laws, which actually represent our ILC-based training scheme for NNs. The norm-optimal ILC formulation is more general since the o.f. also includes the regularization term on the weights update, and it offers a degree of freedom in learning.

6 Case Studies and Discussion of the Results

The case studies apply our IRIT algorithm to the controller tuning for a representative mechatronics application, namely the angular positioning of the vertical motion of a twin-rotor aero-dynamical system experimental setup [16]. A rigid beam supports at one end a horizontal rotor which produces vertical motion and at the other end a vertical rotor causing horizontal motion. The horizontal position is considered fixed in this case study. The nonlinear equations that describe the vertical motion are [16]

$$\begin{aligned} &J_v\dot{\Omega}_v = l_m F_v(\omega_v) - \Omega_v k_v + g[(A-B)\cos\alpha_v - C\sin\alpha_v], \\ &\dot{\alpha}_v = \Omega_v, \\ &I_v\dot{\omega}_v = M(U_v) - M_r(\omega_v), \end{aligned} \tag{29}$$

where $U_v(\%) = u$ is the control signal represented by the PWM duty-cycle corresponding to the input voltage range of the DC motor, $-24\,\mathrm{V} \le u \le 24\,\mathrm{V}$, $\omega_v\,(rad/s)$ is the angular speed of the rotor, $\alpha_v\,(rad) = y$ is the process output corresponding to the pitch angle of the beam which supports the main and the tail rotor, $\Omega_v\,(rad/s)$ is the angular velocity of the beam. The expressions of the other parameters and variables related to (29) are given in [16], and the parameter values are also given in [16] as

$$\begin{aligned} &J_v = 0.02421\,kg\,m^2,\ I_v = 4.5\cdot 10^{-5}\ kg\,m^2, k_v = 0.0127\ kg\,m^2/s, \\ &B - A = 0.05\ rad\,kg\,m, l_m = 0.2\,m,\ C = 0.0936\ rad\,kg\,m. \end{aligned} \tag{30}$$

The nonlinear model (50) is not used in the reference input tuning process except for obtaining an initial feedback controller, which can also be obtained by model-free approaches. A discrete-time linear PID controller with the following t.f. is considered:

$$H(q^{-1}) = (0.012 + 0.001q^{-1})/(1-q^{-1}). \tag{31}$$

The reference trajectory is prescribed in terms of the unit step response of a second-order normalized reference model with the t.f. $\omega_n^2/(s^2+2\zeta\omega_n s+\omega_n^2)$ and the parameters $\omega_n = 0.5$ rad/s and $\zeta = 0.7$. The sampling period is $T_s = 0.1$ s and the length of experiments is of $N = 400$ samples. The relative degree of $T(q^{-1})$ is $n = 1$ and the relative degree of $S_{ur}(q^{-1})$ is $m = 0$.

The initial reference input is chosen such that to obtain a CS response that is very different from the targeted reference trajectory. Therefore, the initial reference input is set as a squared signal of amplitude 0.1, and this motion corresponds to a "take-off" manoeuvre followed by a "landing" manoeuvre. The coefficients of the initial polynomial fit obtained via a least squares algorithm of dimension $h_r + 1 = 8$ are grouped in the parameter vector

$$\boldsymbol{\theta}_0 = [0.05\ \ 0.93\ \ -15.62\ \ 101.36\ \ -307.68\ \ 465.59\ \ -342.52\ \ 97.86]^T. \tag{32}$$

The NN architecture used in the identification and subsequently in the gradient estimation consists of one hidden layer with six neurons and one output layer with one neuron. As shown in Section 5, hyperbolic tangent activation functions are employed in the hidden layer, and a linear function is employed as the output neuron activation function. This NN architecture uses the last two outputs and the last two inputs in order to obtain the output prediction. The same simple architecture is used for both M_{ry} and M_{ru}. The inputs of the two NNs are selected as

$$\begin{aligned} \mathbf{x}_{ry}^T(k) &= [1\ \ y(k)\ \ y(k-1)\ \ r(k)\ \ r(k-1)] \text{ for } M_{ry}, \\ \mathbf{x}_{ru}^T(k) &= [1\ \ u(k)\ \ u(k-1)\ \ r(k)\ \ r(k-1)] \text{ for } M_{ru}. \end{aligned} \tag{33}$$

The outputs of the NNs are the closed-loop output and the control signal, respectively.

The training of the two NN architectures is carried out in our ILC framework. Each neuron in the hidden layer has five parameters, i.e., four weights and one bias. The output layer has seven weights including the bias. We trained the weight vectors $\mathbf{W} \in \mathbf{R}^{7\times 1}$ and $\mathbf{V}_i \in \mathbf{R}^{5\times 1}, i = 1...6$. The initial values of the hidden neurons parameters are chosen from a normal distribution centred at zero with variance 1.

The NN-based identification is carried out on the nominal trajectories of the closed-loop for the initial controller parameters presented in the sequel. Only the results concerning the identified map M_{ry} are presented here. For the norm-optimal ILC problem, the weighting matrices were chosen as $\mathbf{R} = \mathbf{I}_{400}$ and $\mathbf{Q} = 0.0001 \cdot \mathbf{I}_{37}$, where $\mathbf{I}_\zeta$ is the general notation for ζ^{th} order identity matrix. The evolutions of the training error throughout the iterations and of the simulated trajectory before and after training are shown in Figure 1.

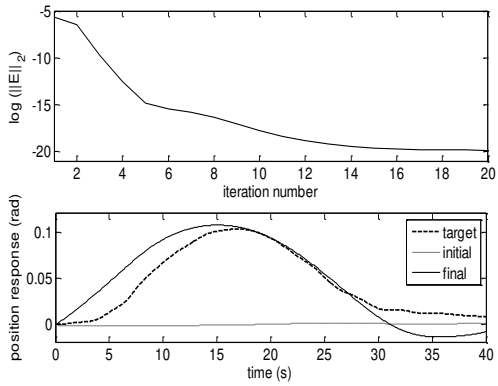


Figure 1

NN training error versus iteration number and controlled output response before and after training

Two simulated case studies are next considered. The first simulated case study deals with the unconstrained optimization, where only the reference input signal is tuned according using our approach in order to ensure the output tracking improvement. A BFGS update was used for the Hessian estimate and the step size was chosen constant equal to 0.12. The final parameter vector that describes the reference input is

$$\boldsymbol{\theta}_{22} = [0.04 \;\; 0.88 \;\; -15.66 \;\; 101.35 \;\; -307.69 \;\; 465.59 \;\; -342.52 \;\; 97.86]^T. \tag{34}$$

Figure 2 gives the initial and final reference input after optimization and the o.f. decrease over the iterations. Only the first five parameters of the parameter vector are changed significantly. The control signal and the final controlled output before and after the optimization are shown in Figure 3 and in Figure 4, respectively. The rise of the reference input on the first 15 s with respect to the initial value and the decrease of the final reference input after 20 s compared to the initial reference input have to be correlated with the output response. This indicates that the reference input is tuned such as to anticipate the low bandwidth of the CS. As an effect, the final controlled output is rising faster under the take-off manoeuvre, and also tracks the reference input more accurately for the landing manoeuvre.

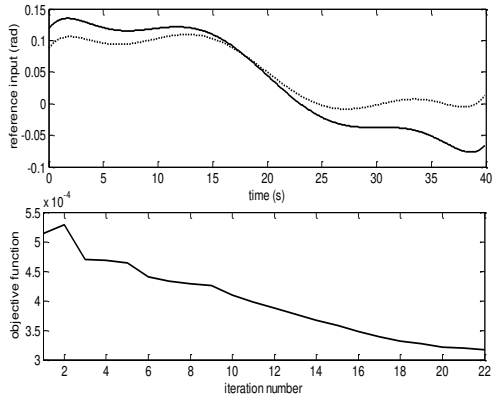


Figure 2

Simulations results expressed as reference input versus time and as o.f. in (5) versus iteration number. The initial reference is dotted and the final reference input is black solid

The second simulated case study corresponding to the optimization with control signal saturation and control signal rate constraints is presented as follows. The two inequality constraints are $-0.05 \le u(k) \le 0.12$ and $-0.01 \le \Delta u(k) \le 0.015$.

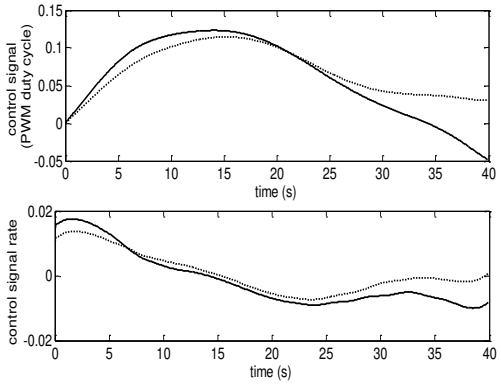


Figure 3

Simulation results for the unconstrained case: control signal responses: initial (dotted) and final (solid)

The algorithm is applied as in the deterministic case as follows. The sequence of penalty parameters in (18) was set to a constant value $p_j = 9$. Two constant values of the step-scaling parameter were used for the gradient descent. When no constraints are violated the step size was set to $\gamma = 0.1$; otherwise, it was set to $\gamma = 0.12$. 400 samples of the reference input are subject to optimization and a total of 796 constraints were used: 798 for control signal saturation and 798 for control signal rate saturation. The final parameter vector that describes the reference input is

$$\boldsymbol{\theta}_{32} = [0.04 \;\; 0.91 \;\; -15.64 \;\; 101.35 \;\; -307.69 \;\; 465.59 \;\; -342.52 \;\; 97.86]^T. \tag{35}$$

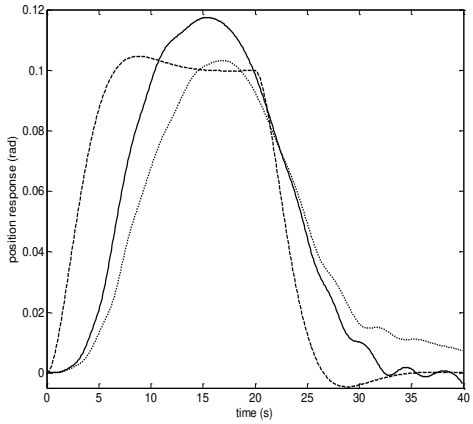


Figure 4

Simulation results for the unconstrained case expressed as position response: initial (black dotted), final response after optimization (black solid) and reference trajectory (black dashed)

The evolution of the reference input during the learning process is presented in Figure 5. The final reference input parameterized by $\boldsymbol{\theta}_{32}$ has higher amplitude for the first 20 sec and forces the take-off motion to respond faster but with higher overshoot. On the other hand, after 20 s, the final reference input drops below the initial reference input trajectory in order to compensate for the slow response in the landing manoeuvre and also to correct the steady-state error.

Figure 5 also shows the sum of penalty functions $\phi(\boldsymbol{\theta}) + \Delta\phi(\boldsymbol{\theta})$ that contributes to the augmented o.f. $\tilde{J}_{p_j}(\boldsymbol{\theta})$ in (18) to be optimized. Since the constraints are violated more, they weight more in the o.f., and they eventually provide a more significant contribution to the gradient of the o.f., thus driving the optimization in the direction of bringing the trajectories within the feasible boundaries. This has a negative impact on the reference tracking criterion. Even with the double approximation involved in the linearity assumption and in the NN-based gradient estimation, the o.f. decreases as shown in Figure 5 and the performance improvements are evident.

The results given in Figure 5 have to be correlated with the control signal responses illustrated in Figure 6. There are several control signal trajectories during the learning process that violate the constraints more heavily. But, our IRIT algorithm brings the trajectories as much as possible close to the boundaries of the feasible region in such situations.

The output trajectory evolution during the learning is presented in Figure 7, where the final output trajectory is closer to the reference trajectory when compared with the initial response. The corresponding final reference input overcomes the difficulty of both the take-off manoeuvre and the landing manoeuvre by anticipating the slow responses.

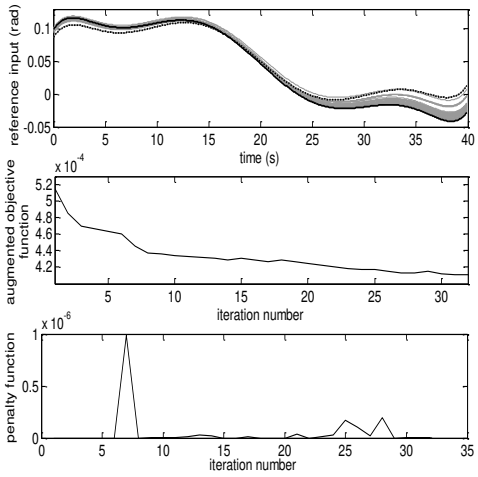


Figure 5

Simulation results expressed as reference input versus time as the learning converges, augmented o.f. (18) versus iteration number and sum of penalty functions. The initial reference is dotted and the final reference input is black solid

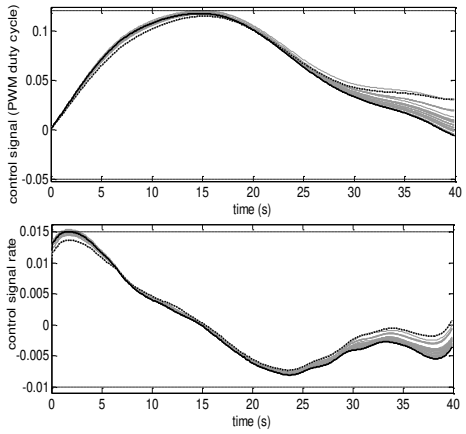


Figure 6

Simulation results for the constrained case expressed as control signal responses: initial (dotted) and final (solid). The constraints are dashed

The advantage of the proposed approach is even more obvious because it can force the controlled output in the vicinity of the reference trajectory; the parameter tuning of the controller can next be carried out using similar model-free iterative techniques in linear or nonlinear formulations such as IFT [17]. Our approach prevents the optimization to get stuck in local minima that are far from the global minimum. Proceeding this way the windsurfing approach is actually solved using our algorithm in a two-degrees-of-freedom tuning setting. Therefore, the CS can be optimized in order to exhibit highly complex manoeuvres.

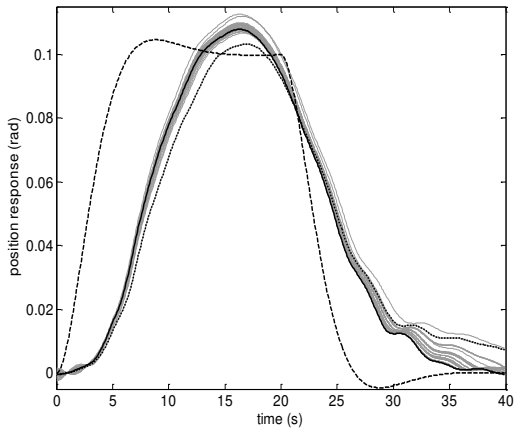


Figure 7

Simulation results for the constrained case expressed as position response: initial (black dotted), final response after optimization (black solid) and reference trajectory (black dashed) and intermediate trajectories (grey)

The optimization of the reference input sequence in the parameterized form leads to less spectacular results then in the case when the non-parameterized reference input is subjected to optimization [13-17]. However, the parameterization effect is that it reduces the many local minima specific to the case when no parameterization is used. It seems that the polynomial approximation employed for the reduction of the reference input dimension has limited potential and other basis functions may be exploited for this purpose such as radial basis functions.

The convergence speed to the solution within few experiments on the real-world process depends on how many constraints are violated at each iteration. This depends on the interplay between the penalty parameter p_j and the step size of the search algorithm.

The discussion presented in this section can be extended because different results will be obtained for other o.f.s and other constraints in the OPs. This depends on the performance specifications and objectives [29-34] for various CS applications [35-40].

The training approach using the ILC framework described in the paper can be extended to other NN architectures. Moreover, it can be employed for the same architecture that is used in this paper, with more than one hidden layer.

Conclusions

This paper has proposed a data-driven algorithm, which solves an optimal control problem in order to ensure the constrained reference trajectory tracking by few experiments conducted on the real-world CS. The new IRIT algorithm has three advantages. First, the closed-loop CS stability is not affected while solving the

trajectory tracking problem. Starting with a given closed-loop controller, the stability is maintained along the iterations of the IRIT algorithm and no additional tests are needed. The change in the controller parameters specific to other tuning techniques including IFT, which usually require attention in order to achieve the bumpless transfer between controllers, is mitigated by our approach. Second, cost-effective controller designs and implementations are achieved because of the linear parameterization that ensures the reduced dimensionality of the reference input vector. Third, our IRIT algorithm is advantageous as it works for smooth nonlinear systems around some operating points

The proposed algorithm can be generalized by considering other data-driven optimization approaches to controller tuning combined with ILC to optimize the reference input sequence. These techniques can yield automated tools for controller design and tuning, with benefits for the CS designers.

Acknowledgement

This work was supported in part by a grant from the Romanian National Authority for Scientific Research, CNCS – UEFISCDI, project number PN-II-ID-PCE-2011-3-0109, by the strategic grant POSDRU/159/1.5/S/137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Program Human Resources Development 2007-2013, and from the NSERC of Canada.

References

[1] A. S. Bazanella, L. Campestrini, D. Eckhard: Data-Driven Controller Design: The H_2 Approach, Springer-Verlag, Berlin, Heidelberg, 2012

[2] Z.-S. Hou, Z. Wang: From Model-Based Control to Data-Driven Control: Survey, Classification and Perspective, Information Sciences, Vol. 235, 2013, pp. 3-35

[3] S. Gunnarsson, M. Norrlöf: On The Design of ILC Algorithms Using Optimization, Automatica, Vol. 37, No. 12, 2001, pp. 2011-2016

[4] D. H. Owens, J. Hätönen: Iterative Learning Control – An Optimization Paradigm, Annual Reviews in Control, Vol. 29, No. 1, 2005, pp. 57-70

[5] M. Norrlöf, S. Gunnarsson: Time and Frequency Domain Convergence Properties in Iterative Learning Control, International Journal of Control, Vol. 75, No. 14, 2002, pp. 1114-1126

[6] M. Butcher, A. Karimi, R. Longchamp: Iterative Learning Control Based on Stochastic Approximation, Proceedings of 17th IFAC World Congress, Seoul, Korea, 2008, pp. 1478-1483

[7] H.-F. Chen, H.-T. Fang: Output Tracking for Nonlinear Stochastic Systems by Iterative Learning Control, IEEE Transactions on Automatic Control, Vol. 49, No. 4, 2004, pp. 583-588

[8] S. Mishra, U. Topcu, M. Tomizuka: Optimization-based Constrained Iterative Learning Control, IEEE Transactions on Control Systems Technology, Vol. 19, No. 6, 2011, pp. 1613-1621

[9] P. Janseens, G. Pipeleers, J. Swevers: A Data-Driven Constrained Norm-Optimal Iterative Learning Control Framework for LTI Systems, IEEE Transactions on Control Systems Technology, Vol. 21, No. 2, 2013, pp. 546-551

[10] S. Lupashin, A. Schöllig, M. Sherback, R. D'Andrea: A Simple Learning Strategy for High-Speed Quadrocopter Multi-Flips, Proceedings of 2010 IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 2010, pp. 1642-1648

[11] J. Z. Kolter, A. Y. Ng: Policy Search via the Signed Derivative, in Robotics: Science and Systems V, J. Trinkle, Y. Matsuoka, J. A. Castellanos, Eds., The MIT Press, Cambridge, MA, USA, 2009, pp. 1-8

[12] Z. Xiong, J. Zhang: A Batch-to-Batch Iterative Optimal Control Strategy Based on Recurrent Neural Network Models, Journal of Process Control, Vol. 15, No. 1, 2005, pp. 11-21

[13] M.-B. Radac, R.-E. Precup, E. M. Petriu, S. Preitl, C.-A. Dragos: Experiment-Based Approach to Reference Trajectory Tracking, Proceedings of 2012 IEEE Multi-Conference on Systems and Control, Dubrovnik, Croatia, 2012, pp. 470-475

[14] M.-B. Radac, R.-E. Precup, E. M. Petriu, S. Preitl, C.-A. Dragos: Data-Driven Reference Trajectory Tracking Algorithm and Experimental Validation, IEEE Transactions on Industrial Informatics, Vol. 9, No. 4, 2013, pp. 2327-2336

[15] M.-B. Radac, R.-E. Precup, E. M. Petriu: Design and Testing of a Constrained Data-Driven Iterative Reference Input Tuning Algorithm, Proceedings of 13th European Control Conference, Strasbourg, France, 2014, pp. 2034-2039

[16] M.-B. Radac, R.-E. Precup, E. M. Petriu, S. Preitl: Iterative Data-Driven Tuning of Controllers for Nonlinear Systems with Constraints, IEEE Transactions on Industrial Electronics, Vol. 61, No. 11, 2014, pp. 6360-6368

[17] M.-B. Radac, R.-E. Precup, E. M. Petriu, S. Preitl: Iterative Data-Driven Controller Tuning with Actuator Constraints and Reduced Sensitivity, Journal of Aerospace Information Systems, Vol. 11, No. 9, 2014, pp. 551-564

[18] C. T. Freeman, Y. Tan: Iterative Learning Control with Mixed Constraints for Point-to-Point Tracking, IEEE Transactions on Control Systems Technology, Vol. 21, No. 3, 2012, pp. 604-616

[19] M. Volckaert, M. Diehl, J. Swevers: Generalization of Norm Optimal ILC for Nonlinear Systems with Constraints, Mechanical Systems and Signal Processing, Vol. 39, No. 1-2, 2013, pp. 280-296

[20] J. X. Xu, Y. Chen, T. H. Lee, S. Yamamoto: Terminal Iterative Learning Control with an Application to RTPCVD Thickness Control, Automatica, Vol. 35, No. 9, 1999, pp. 1535-1542

[21] J. van de Wijdeven, O. Bosgra: Using Basis Functions in Iterative Learning Control: Analysis and Design Theory, International Journal of Control, Vol. 83, No. 4, 2010, pp. 661-675

[22] G. Pipeleers, K. L. Moore: Reduced-Order Iterative Learning Control and a Design Strategy for Optimal Performance Tradeoffs, IEEE Transactions on Automatic Control, Vol. 57, No. 9, 2012, pp. 2390-2395

[23] J. Boolder, B. Lemmen, S. Koekebakker, T. Oomen, O. Bosgra, M. Steinbuch: Iterative Learning Control with Basis Functions for Media Positioning in Scanning Inkjet Printers, Proceedings of 2012 IEEE Multi-Conference on Systems and Control, Dubrovnik, Croatia, 2012, pp. 1255-1260

[24] M. Heertjes, D. Hennekens, M. Steinbuch: MIMO Feed-Forward Design in Wafer Scanners Using a Gradient Approximation-Based Algorithm, Control Engineering Practice, Vol. 18, No. 5, 2010, pp. 495-506

[25] M. J. C. Ronde, G. A. L. Leenknegt, M. J. G. van de Molengraft, M. Steinbuch: Data-Based Spatial Feedforward for Over-Actuated Motion Systems, Mechatronics, Vol. 24, No. 6, 2014, pp. 679-690

[26] J. Sjöberg, F. De Bruyne, M. Agarwal, B. D. O. Anderson, M. Gevers, F. J. Kraus, N. Linard: Iterative Controller Optimization for Nonlinear Systems, Control Engineering Practice, Vol. 11, No. 9, 2003, pp. 1079-1086

[27] I.-J. Wang, J. C. Spall: Stochastic Optimization with Inequality Constraints Using Simultaneous Perturbations and Penalty Functions, International Journal of Control, Vol. 81, No. 8, 2008, pp. 1232-1238

[28] Y. He, M. C. Fu, S. I. Marcus: Convergence of Simultaneous Perturbation Stochastic Approximation for Nondifferentiable Optimization, IEEE Transactions on Automatic Control, Vol. 48, No. 8, 2003, pp. 1459-1463

[29] S. Preitl, R.-E. Precup, J. Fodor, B. Bede: Iterative Feedback Tuning in Fuzzy Control Systems. Theory and Applications, Acta Polytechnica Hungarica, Vol. 3, No. 3, 2006, pp. 81-96

[30] F.-G. Filip, K. Leiviskä: Large-Scale Complex Systems, in Springer Handbook of Automation, S. Y. Nof, Ed., Springer-Verlag, Berlin, Heidelberg, 2009, pp. 619-638

[31] J. Vaščák, K. Hirota: Integrated Decision-Making System for Robot Soccer, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 15, No. 2, 2011, pp. 156-163

[32] R.-E. Precup, S. Preitl, M.-B. Radac, E. M. Petriu, C.-A. Dragos, J. K. Tar: Experiment-Based Teaching in Advanced Control Engineering, IEEE Transactions on Education, Vol. 54, No. 3, 2011, pp. 345-355

[33] M. Bošnak, D. Matko, S. Blažič: Quadrocopter Hovering Using Position-Estimation Information From Inertial Sensors and a High-Delay Video System, Journal of Intelligent & Robotic Systems, Vol. 67, No. 1, 2012, pp. 43-60

[34] D. Rojas, G. Millan, F. Passold, R. Osorio, C. Cubillos, G. Lefranc: Algorithms for Maps Construction and Localization in a Mobile Robot, Studies in Informatics and Control, Vol. 23, No. 2, 2014, pp. 189-196

[35] R.-E. Precup, S. Preitl: Stability and Sensitivity Analysis of Fuzzy Control Systems. Mechatronics Applications, Acta Polytechnica Hungarica, Vol. 3, No. 1, 2006, pp. 61-76

[36] Z. C. Johanyák: Fuzzy Modeling of Thermoplastic Composites' Melt Volume Rate, Computing and Informatics, Vol. 32, No. 4, 2013, pp. 845-857

[37] K. Lamár, J. Neszveda: Average Probability of Failure of Aperiodically Operated Devices, Acta Polytechnica Hungarica, Vol. 10, No. 8, 2013, pp. 153-167

[38] H.-N. Teodorescu: On the Characteristic Functions of Fuzzy Systems, International Journal of Computers Communications & Control, Vol. 8, No. 3, 2013, pp. 469-476

[39] D. Yazdani, B. Nasiri, R. Azizi, A. Sepas-Moghaddam, M. R. Meybodi: Optimization in Dynamic Environments Utilizing a Novel Method Based on Particle Swarm Optimization, International Journal of Artificial Intelligence, Vol. 11, No. A13, 2013, pp. 170-192

[40] E. Osaba, F. Diaz, E. Onieva, R. Carballedo, A. Perallos: AMCPA: A Population Metaheuristic With Adaptive Crossover Probability and Multi-Crossover Mechanism for Solving Combinatorial Optimization Problems, International Journal of Artificial Intelligence, Vol. 12, No. 2, 2014, pp. 1-23

Colourfastness of Multilayer Printed Textile Materials to Artificial Light Exposure

Nemanja Kašiković, Dragoljub Novaković, Igor Karlović, Gojko Vladić, Neda Milić

University of Novi Sad, Faculty of Technical Sciences, Department of Graphic engineering and design, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
knemanja@uns.ac.rs, novakd@uns.ac.rs, karlovic@uns.ac.rs, vladicg@uns.ac.rs, milicn@uns.ac.rs

Abstract: Accuracy of colour reproduction and colourfastness are the most significant parameters of the quality of printed fabrics in the textile industry. The aim of this paper is to gain deeper insights into the colour difference of textiles printed by ink jet method before and after light exposure. In order to accomplish this, 60 samples were printed, using four process colours (CMYK) on three textile materials with different characteristics (fabric weight and thread count). Additionally, number of ink layers varied (from 1 to 5 layers). Afterwards, printed samples were exposed to the light. The colour of the samples was measured using spectrophotometer before and after exposition to the light, colour difference was calculated and afterwards analysed. Colour difference analysis has shown the significant difference of colour fastness within samples that were printed with different number of ink layers and also between different textile materials.

Keywords: ink jet printing; light exposure; colour difference

1 Introduction

Textile printing can be defined as the process of transferring ink to the textile substrate by using specific printing technique. The fact is that printing processes are changing rapidly and intensively [1]. That affects the textile printing industry, causing decrease of textile screen printing market share from 90% in 2000 [2] by introduction of digital ink jet textile printing that offers higher printing speed of short runs, flexibility, creativity and environment safety [3, 4]. It is important to note that using digital printing technique enables better visual effects, as well as no limitation of print formats. Besides that, it is easier to get unified print quality during the production runs [5, 6]. Another advantage of digital ink jet is ability of printing on the great number of different substrates. One of the most often used fabric for digital printing is polyester based, especially for the manufacturing of the flags. Thermal stability, excellent behaviour during exploitation and uniform quality are characteristics that make polyester so applicable [7].

2 Colour Dyeing and Printing of Textiles

The quality of the produced image should be comparable to the quality of the printed images that are produced by the means of rotary screen printing or other printing technologies used for textile printing. The final print quality in digital textile printing depends on:

-Ink and ink system (used dyes, binders, monomers, or oligomers).
-Print head (resolution, jet straightness and nozzle reliability).
-Substrate: Depending on the type of fabric, different approaches are needed.
-Surface properties affect penetration of the inks, surface pre- and post-treatment could improve image quality, and for example dirt could cause printing artefacts.
- Fabric fixing mechanism: Motion of the fabric affects registration between dots and cause artefacts, like banding or stitching lines.
- Image processing. [8]

The choice of textile substrate influences image quality (i.e., inter-color bleed, dot quality, color, visual perception, etc.), ink drying time, and fastness (light, wet, gas, etc.). Fabric substrates are three-dimensional structures, and low viscosity inks can wick into macro-capillaries between yarns and fibers. Inks can also diffuse into the micro-capillaries in fibers.The wicking and diffusion rates are controlled by the surface tension of ink, ink viscosity, yarn and fabric structures, and the polymer morphology of the fiber. Ultimately, dye molecules in the ink droplets must be fixed on or near the surface of the textile fiber substrate for sharp and brilliant color images. The fixing mechanism depends on the dye/fiber combination [9]. Different dyes are used for the coloration of different fibers. Colour durability, such as wet-fastness and light-fastness, differs for each combination of dye and substrate. An overview of these combination is given in [10]. Several thesis investigated the structual and optical characteristics of printed textiles on the the colour reproduction [11] as well the technological improvement of digital ink jet printing regarding the aformentioned influencing factors [12] and the possibility of using novelty materials in post treatment of digital prints of textile materials for better print quality [13]. In another paper by [14] the influences of exposure of a four-layer stack of woven polyester fabrics to atmospheric pressure air plasma treatment was investigated regarding colourfastness. The research of anti bleeding property and colourstrength revealed that the atmospheric pressure plasma could substantially penetrate two layers of polyester fabrics to observably improve their ink jet printing effect.This effect also contributed to the deeper and more vivid colour and better anti-bleeding performance.

2.1 Light Exposure of Printed Textile Samples

Textile materials used for flag printing are exposed to different environment influences such as heat, UV radiation, moisture, washing, etc. during exploitation.

One of the factors that has an impact on the product quality, is ability to retain colour while exposed to light over long time. This represents a great issue because final appearance of the product cannot be predicted [15]. Light exposure of printed samples causes colour differences, although structural modification of material is also possible [16]. Researchers trying to improve the lightfastness of the printed colors on the textile material deploy two possible solutions:either they can choose dyes with better lightfastness, and/or to use a UV absorber. In a paper [17] it was stated that by using UV absorbers, both water insoluble and soluble the lightfastness of inkjet printed cotton fabrics with reactive inks could improve. The water soluble UV absorber had better lightfastness improvement than the water insoluble UV absorber based on the conditions examined. Exposure of printed samples to light can be done according to standards: ISO105 – B02, ISO105-B01-1999, ISO105-B03 -1997, ISO105-B04-1997, ISO105-B05-1996, ISO105-B06-1999, ISO105-B07:2009, ISO105-B08:1999. Majority of researchers use ISO105 – B02 standard [18, 19, 20, 21, 22]. The difference between samples, exposed and unexposed to light, is usually determined visually, using the blue wool reference. However, visual judgment can not provide information accurate enough in all circumstances. The more objective way to measure colour difference changes for different colourfastness test (light,wash, rub) is a spectrophotometric method which was used by several researchers [23, 24, 25].

3 Materials and Methods

This paper researches the influence of light on colour prints made on three textile substrates. The experiment demanded step-by-step procedure and measurement and gathering of relevant data for detailed analysis of how light influences on printed samples. Samples were printed using digital ink jet printing technique (Mimaki JV22-160 printing system with J-Eco Subly Nano inks). The used ink are dispersed dye digital ink for direct ink jet material printing and they have good fastness properties, which makes them suitable for outdoor products printing. The interaction of the polyester and the dye is hydrophobic with solid state mechanism. The fixation is by heat when the dyes go to gaseous state and are absorbed by the printed polyester fabric and when the dyes are cooled down it enables good bonding to the polyester fibre as they can get inbetween the tightly paded polymer chains.

The colour values of samples before and after light exposure were specified by spectrophotometric measurement according to which the colour differences were calculated. As flags three types of polyester fabrics were used, mostly because of their specific characteristics including superior strength and resistance [26]. For all samples material characteristics were determined according to standards: fabric weight using standard ISO 3801, thread count (ISO 7211-2) and material composition (ISO 1833). Characteristics of used materials are shown in Table 1.

Table 1
Characteristics of textile materials used in the experiment

Tested samples	Material composition (%)	Fabric weight (g/m2)	Thread count (p/10 cm)	
			Warp	Weft
Material 1	Polyester 100 %	110,6	170	120
Material 2	Polyester 100 %	101,5	160	100
Material 3	Polyester 100 %	141,3	260	120
Examination	ISO 1833	ISO 3801	ISO 7211-2	

In order to analyse influence of light and heat a test card was prepared. Dimensions of test card were150 x 10 cm and consisted of 4 patches of colour, size 35 x 10 cm. Colour values of patches were: patch 1 – 100% cyan, patch 2 – 100% magenta, patch 3 – 100% yellow, patch 4 – 100% black. Since printing system Mimaki JV 22 – 160 is able to make variations in number of printed ink layers, print test card was used in five variations, from one to five ink layers. It was assumed that increase of ink layers will improve colour fastness during exploitation and longer exposure to various influences. To measure colorimetric properties of samples the spherical spectrophotometer Datacolor Spectraflash SF 600® PLUS - CT was used. This spectrophotometer features D_{65} standard light and standard observer 10°, with measuring aperture of 16 mm. During spectrophotometric measurement, all textile samples were folded three times and set on mat white background with $L^* > 92$ and $C^* < 3$ (standard ISO13655). Every sample was measured 10 times and average value was calculated accordingly. Before starting measurement procedure spectrophotometer was tested in terms of preciseness and accuracy according to standard ASTM E2214-08 (2008) and obtained results were in tolerable extent specified by device specification. Light exposure of samples was conducted according to ISO 105-B02 standard (method 2), Xenon test chamber Alpha by manufacturer Atlas was used to simulate influence of light. Samples were exposed according to predefined conditions (temperature, exposure time, relative moisture). After exposing the samples, spectrophotometric measurements were made in order to determine colour fastness of textile samples. In order to determine statistically how material characteristic and number of ink layer influence the colour differences (CIE LCH colour space) caused by light exposure ANOVA and T tests were used. All these analyses gave guidance how to create model for simulation of light influences on specific colour that is printed on specific textile sample.

4 Results and Discussion

After printing process the colour coordinates CIE L (lightness), a, b, C (chroma) and H (hue) colour coordinates (CIE Lab and LCH colour spaces) were determined for printed sample with variations of numerous ink layers. Results in

Table 2 represent comparison of samples printed with five ink layers and other samples to using colour difference ΔE76 formula.

Table 2
Spectrophotometer measurement results

Sample	L	a	b	C	h	ΔE	Pearson correlation (L and ink layer	Pearson correlation (C and ink layer number)
1-1C	48.29	-7.29	-31.13	31.97	256.82	16.7		
1-2C	43.78	-6.09	-27.7	28.36	257.6	11.1		
1-3C	39.94	-5.82	-25.95	26.6	257.36	6.9		
1-4C	36.15	-5.29	-24.75	25.31	257.94	2.89		
1-5C	33.5	-4.72	-23.74	24.2	258.77	/	-0.99	-0.97
1-1M	46.74	48.71	0.88	48.72	1.03	16.3		
1-2M	41.34	44.94	2.95	45.03	3.76	9.56		
1-3M	39.24	41.96	2.59	42.04	3.53	6.2		
1-4M	36.54	40.27	3.45	40.42	4.9	2.91		
1-5M	34.37	38.54	4.33	38.78	6.41	/	-0.98	-0.98
1-1Y	78.28	-2.73	65.72	65.77	92.38	16.5		
1-2Y	74.64	1.78	68.71	68.73	88.52	10.1		
1-3Y	72.02	4.39	70.92	71.05	86.46	6.03		
1-4Y	69.81	6.78	71.71	72.03	84.6	2.73		
1-5Y	68.03	8.84	71.67	72.22	82.97	/	-0.99	0.94
1-1K	36.04	-0.33	-1.18	1.23	254.56	13.8		
1-2K	30.01	-0.65	-0.91	1.12	234.55	7.85		
1-3K	25.68	-0.4	-1.02	1.1	248.63	3.52		
1-4K	23.35	-0.58	-0.95	1.11	238.68	1.19		
1-5K	22.16	-0.51	-0.88	1.02	239.54	/	-0.96	-0.91
2-1C	44.21	-4.32	-31.74	32.03	262.25	18.9		
2-2C	37.11	-2.44	-27.27	27.38	264.88	10.4		
2-3C	34.42	-0.02	-24.02	24.02	269.95	6.3		
2-4C	30.73	-1.23	-23.17	23.2	266.96	2.74		
2-5C	28.62	-0.46	-21.6	21.6	268.79	/	-0.97	-0.95
2-1M	43.07	49.59	1.55	49.62	1.79	18.7		
2-2M	36.49	42.8	3.12	42.92	4.17	9.29		
2-3M	32.81	38.87	5.51	39.26	8.06	4.55		
2-4M	32.79	38.66	5.12	39	7.55	4.21		
2-5M	29	32.36	5.28	32.79	9.26	/	-0.95	-0.96
2-1Y	77.17	4.82	75.02	75.18	86.32	20.8		
2-2Y	71.77	4.82	75.02	75.18	86.32	15.0		
2-3Y	67.91	6.93	72.02	72.36	84.5	10.0		
2-4Y	64.85	8.49	68.91	69.43	82.98	5.77	-0.99	
2-5Y	62.31	8.95	65.05	65.66	82.17	/		-0.96
2-1K	27.28	-0.3	-0.14	0.33	205.02	6.4		
2-2K	23.81	-0.43	-0.14	0.45	198.29	2.96	-0.93	-0.77

2-3K	22.23	-0.18	-0.19	0.26	227	1.36		
2-4K	21.52	-0.1	-0.09	0.14	222.64	0.66		
2-5K	20.97	0.07	-0.18	0.19	293.04	/		
3-1C	52.3	-0.53	-32.08	32.09	269.06	17.9		
3-2C	42.67	-0.31	-31.27	31.27	269.43	8.67		
3-3C	38.89	0.73	-30.16	30.17	271.38	4.65		
3-4C	36.69	2.45	-28.95	29.05	274.84	2.14		
3-5C	34.95	1.75	-27.91	27.97	273.59	/	-0.93	-0.99
3-1M	51.34	38.48	-8.58	39.43	352.19	21.1		
3-2M	39.15	41.25	-3.06	41.36	355.75	11.6		
3-3M	36.62	33.33	-4.57	33.64	352.19	5.09		
3-4M	34.49	31.06	-4.06	31.33	352.56	2.91		
3-5M	32.28	31.87	-2.35	31.96	355.78	/	-0.90	-0.87
3-1Y	79.74	-3.25	68.35	68.42	92.72	16.5		
3-2Y	75.14	2.06	68.87	68.9	88.29	10.0		
3-3Y	72.56	3.81	68.68	68.78	86.83	7.13		
3-4Y	69.69	6.53	66.88	67.2	84.42	2.88		
3-5Y	67.68	7.59	65.12	65.56	83.35	/	-0.99	-0.83
3-1K	35.17	3.42	0.85	3.52	13.89	12.8		
3-2K	26.84	1.64	-0.3	1.67	349.55	4.33		
3-3K	26.03	1.58	-0.27	1.6	350.32	3.53		
3-4K	24.15	1.53	-0.17	1.54	353.8	1.67		
3-5K	22.52	1.55	-0.51	1.63	341.99	/	-0.90	-0.72

Note: Labels in the Sample column are structured so that the first number represents textile material; the second number represents number of ink layers while the letter represents colour that the sample was printed in.

Results presented in table 2 show that increasing number of ink layers during print leads to decreasing of lightness and saturation at all four process colours. Changes of lightness and saturation have linear tendency with a high value of correlation coefficient. Therefore, colour difference value will rise with change of ink layer number.The lowest values of ΔE were calculated for black colour for all materials, while samples printed by cyan, magenta and yellow showed higher values of ΔE, similar across all materials. Colour difference values where mostly greater than 5, ΔE>5, except in comparison with samples printed with four and five ink layers. Based on linear tendency of lightness and saturation changes it can be assumed that further increase in number of ink layer number will increase value ΔE also.

4.1 Spectrophotometric Measurement after Light Exposure

After printed samples had been exposed to light according to ISO105-B02 standard (method 2) (temperature, lightness, moisture) the spectrophotometric measurements were conducted in order to determine how the exposure to light influenced changes in sample colour. The earlier researches regarding similar subject [22] proved that increased number of ink layers is increasing light fastness, although limited to visual assessment of colour changes. It was expected that spectrophotometric measurements confirm results gained by visual analysis. This

practically means that samples assessed with higher grade of colour fastness would have lower colour difference values after light exposure. The table 3 represents resulting differences of colour coordinates ΔL, Δa, Δb, ΔC and Δh of samples.

Table 3
Differences of colour coordinates before and after light exposure

Sample	ΔL	Δa	Δb	ΔC	Δh
1-1C	0.87	2.58	2.78	3.24	4.76
1-2C	1.53	2.28	2.73	3.15	4.48
1-3C	0.79	2.15	2.84	3.23	4.37
1-4C	1.2	2.38	2.23	2.65	5.71
1-5C	0.69	1.83	2.47	2.78	4.06
1-1M	1.59	1.07	0.57	1.11	0.6
1-2M	1.2	1.46	0.45	1.49	0.44
1-3M	0.85	1.28	0.52	1.31	0.51
1-4M	1.42	0.53	-0.11	0.52	-0.16
1-5M	0.88	0.66	0.38	0.68	0.36
1-1Y	1	-0.43	-2.25	2.27	0.3
1-2Y	1.39	-0.74	-1.76	1.78	0.82
1-3Y	2.02	0.15	1.19	1.2	0.01
1-4Y	1.39	-0.29	-0.52	0.55	0.24
1-5Y	1.37	-0.21	0.06	0.04	0.21
1-1K	4.02	1.8	-0.57	1.84	-
1-2K	3.33	1.63	-0.78	1.8	-
1-3K	3	1.78	-0.18	1.61	-
1-4K	2.17	2.07	0.05	1.74	-
1-5K	1.74	2.35	0.43	1.77	-
2-1C	2.14	1.36	-0.32	0.11	3.91
2-2C	1.53	1.92	0.03	0.31	4.82
2-3C	0.39	1.99	0.27	0.62	4.17
2-4C	0.52	1.23	-0.53	0.37	3.84
2-5C	1.12	0.76	-0.32	0.22	2.5
2-1M	0.73	4.13	0.28	4.14	-0.16
2-2M	2.48	2.83	0.08	2.83	-0.23
2-3M	2.76	0.79	0	0.78	-0.07
2-4M	1.58	2.18	0.23	2.19	0
2-5M	0.81	2.15	-0.22	2.13	-0.35
2-1Y	4.4	-1.22	6.42	6.26	1.74
2-2Y	3.92	-1.55	4.73	4.6	1.78
2-3Y	3.84	-0.66	4.56	4.43	1.22
2-4Y	3.09	-1.33	2.76	2.7	1.39
2-5Y	1.81	-0.92	-0.21	0.21	0.92
2-1K	2.64	0.71	0.01	0.69	-
2-2K	2.19	1.07	0.21	1.02	-
2-3K	1.9	1.35	-0.04	1.34	-
2-4K	1.92	1.26	0.14	1.18	-
2-5K	1.82	1.12	0.43	1.01	-

3-1C	2.72	0.64	0.4	0.46	1.69
3-2C	2.01	0.89	1.19	1.32	1.8
3-3C	2.17	0.41	0.92	0.96	0.81
3-4C	1.7	0.84	1.04	1.13	1.85
3-5C	2.16	-0.07	-0.09	0.1	-0.17
3-1M	2.74	4.77	0.97	4.81	0.82
3-2M	1.74	4.76	1.07	4.81	0.98
3-3M	2.24	4.03	0.45	4.05	0.37
3-4M	1.55	3.76	0.67	3.8	0.55
3-5M	2.06	2.43	0.07	2.43	0.15
3-1Y	4.24	-0.86	4.07	3.96	1.2
3-2Y	3.71	-0.73	1.75	1.7	0.83
3-3Y	3.15	-0.18	2.26	2.23	0.38
3-4Y	2.62	-1.03	1.17	1.13	1.09
3-5Y	1.78	-0.35	-1.9	1.9	0.41
3-1K	5.49	1.99	-0.22	1.89	-
3-2K	5.07	2.58	0.13	2.56	-
3-3K	4.63	2.2	0.96	2.35	-
3-4K	4.76	2.09	0.24	2.09	-
3-5K	1.69	2.27	0.54	2.26	-

Note: Labels in the Sample column are structured so the first number represents textile material; the second number represents number of ink layers while the letter represents colour that the sample was printed in. The value of hue change Δh was not calculated for black colour because of its achromacy.

Values of ΔE are grouped according to colour and presented on Figure 2.

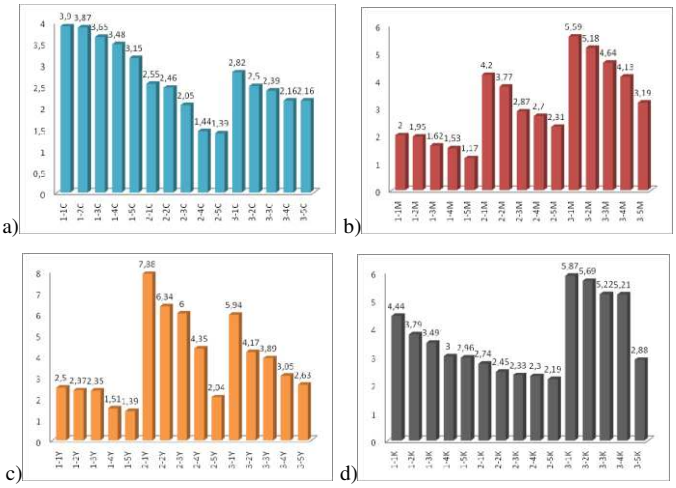


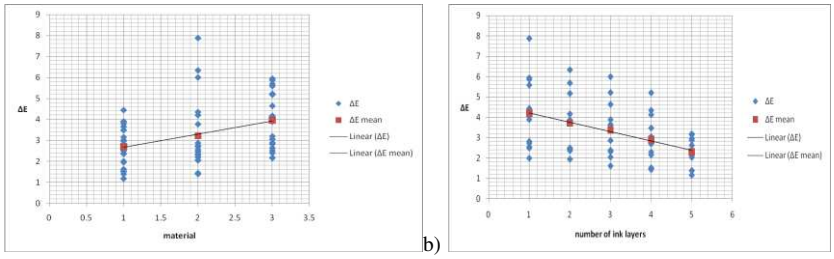
Figure 2

Value ΔE between unexposed and light exposed samples – all textiles printed by: a) cyan, b) magenta, c) yellow d) black colour

Figure 2 indicates that increased number of ink layers during printing cause better light fastness compared to samples printed with lower number of ink layers, as it was concluded by preliminary researches [22]. This can be concluded since the highest value of ΔE was recorded after light exposure of the sample that was printed in one ink layer, while the lowest value of ΔE was marked after light exposure of the sample printed in five ink layers.

In order to determine how ink layers, material characteristics (fabric weight and thread count) and type of ink influenced colour difference (ΔE), ANOVA analysis was conducted followed by Tukey tests. Although ANOVA test, as all parameter tests, has assumption about normal distribution of sample group, it is very robust and in case of bigger samples it does not cause any significant problems [27]. The two-way ANOVA with material characteristics and ink layer number as between subject factors showed no interactive effect of the two factors, Using one-way ANOVA statistically significant colour difference was determined depending on type of textile material, $F(2,57) = 4.130$, $p=0.021$. Post-hoc tests showed that statistical difference existed between material 3 ($M= 3.97$, $SD= 1.36$) and material 1 ($M= 2.71$, $SD= 0.99$), while material 2 ($M= 3.22$, $SD= 1.73$) did not differ significantly from other samples. Furthermore, statistically significant difference in change of lightness was determined depending on number of ink layers, $F(4,55) = 3.552$, $p=0.012$. Post-hoc tests showed that there was significant difference between samples with one ($M= 4.20$, $SD= 1.81$) and with five ink layers ($M= 2.29$, $SD= 0.7$), while the group of samples with two ($M= 3.71$, $SD= 1.44$), three ($M= 3.37$, $SD= 1.36$) and four layers ($M= 2.90$, $SD= 1.21$) did not differ from one to another and did not differ from samples with one and with five ink layer.

Figure 3 represents graphs of colour difference that are grouped by influence factors: a) textile and b) number of ink layers.



a) b)

Figure 3

Colour difference graphs grouped by factors: a) material, b) number of ink layers

The Pearson correlation coefficients between values of lightness and chroma before and after light exposure including all 60 samples were both high -0.998, with corresponding coefficient of determination of 0,995 pointing to possibility of creating simulation model for colour appearance after light exposure. This is an important result since it means that according colour fastness behaviour of just one sample, the lightness and chroma changes of any sample can be predicted.

The graphics of these linear correlations with equations are show on Figure 4. From obtained results can be deducted that significant difference in colour change after light exposure, ΔE, is caused by different colour values of samples before light exposure on samples with different thread count and printed with different ink layer number.

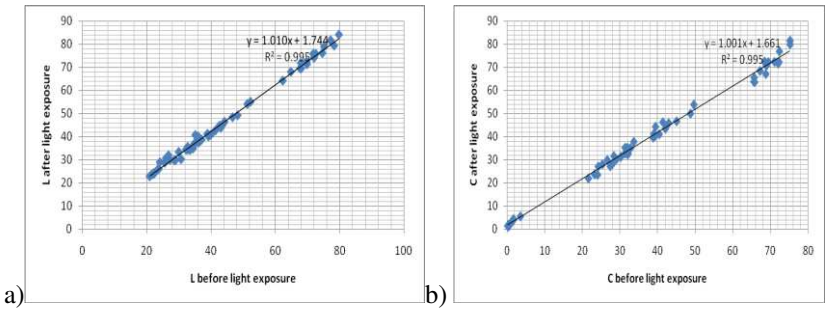


Figure 4

Correlation of values before and after light exposure for: a) lightness, b) chroma

Conclusion

Researching the light influence on printed materials has been a field of concern to many researches in textile dying and printing. The aim of these researcher were to prolong the use of the printed or dyed product with intact colour deviations. The use of digital ink jet printers enable multi passs printing with different number of layers. This reasearch focused on spectrophotometric measurement assessment of the influence of multiple printed layers. Exposure to light of printed samples showed that increased number of ink layers increases colour fastness, spectrophotometric measurement showed lower colour difference. The highest values of ΔE were calculated for samples made by only one ink layer, while the lowest values were marked on light exposed samples with five ink layers. This can be explained by a greater number of ink particles on the surface. Moreover, colour differences between samples printed with five ink layers and all other samples were calculated. The greatest colour difference was noted between samples with one and five ink layers ($\Delta E > 6$), and the smallest one was between samples made of four and five ink layers ($\Delta E < 6$).

Also, characteristics of substrate have been proven to be important influential factor for light fastness and therefore cannot be ignored.

According to results of the research, it is recommended to take into consideration number of ink layers when choosing digital printing techniques in order to achieve high quality. If extended colourfastness is needed additional layers of the ink can be applied to the substrate. The other point is that spectrophotometric measurement can be better solution then visual assessments, since it provides more accurate results. Numerical values gathered from instrumental measurements are more suitable for quantification of values and colour management and are

more easily controlabble by digital ink jet printer than in classic textil dying. Regardless of the results, the importance of visual assessment cannot be disputed since the end user is a human.

Acknowledgement

This work was supported by the Serbian Ministry of Science and Technological Development, Grant No.:35027 "The development of software model for improvement of knowledge and production in graphic arts industry"

References

[1] Neral B., Šostar Turk S., Schneider R. Efikasnost mikrovalnog fiksiranja reaktivnog bojila C.I. Reactive Red 24 u digitalnom tisku pamučnih tkanina, In: Tekstil, 2007, Vol. 56, Issue 6, 2007, p. 358

[2] Horrocks, A. R. (ed.). Handbook of Technical Textiles, The Textile Institute, Woodhead Pub. Ltd., Cambridge, UK, 2000

[3] Choi P. S. R., Yuen C. W. M., Ku S. K. A., Kan C. W. Digital Ink-jet Printing for Chitosan-treated Cotton Fabric, In: Fibers and Polymers, 2005, Vol. 6, Issue 3, p. 229

[4] Masaru O., Kazuhide Y., Yukio A. Textile Printing by the Ink-Jet Printer, In: Nihon Gazo Gakkaishi/Journal of the Imaging Society of Japan, 2010, Vol. 49, Issue 5, p. 417

[5] Owen P. Digital Printing: A World of Opportunity from Design to Production, In: AATCC Review, 2003, Vol. 3, Issue 9, p. 10

[6] Xue C. H, Shi M. M, Chen H. Z. Preparation and Application of Nanoscale Micro Emulsion as Binder for Fabric Inkjet Printing, In: Colloids and Surfaces A: Physicochemical. Eng. Aspects, 2006, Vol. 287, Issue 1-3, p. 147

[7] Ristić N., Ristić I., Jovaničić P., Jocić D. One-Bath Dyeing of Polyester/Cotton Blend with Reactive Dye after Alkali and Chitosan Treatment, In: Industria Textila, 2012, Vol. 63, Issue 4, p. 190

[8] Phillips T. Revolutionizing Textile Decoration and Finishing with Ink Jet Technology [online]. Talk given at Textile Coating & Laminating Conference. Xennia Technology Ltd: Cannes; November 2010. [Online] URL: http://www.xennia.com/uploads/pppRevolutionisingTextileDecoration-Nov2010-sml.pdf. [Accessed 6 April 2013]

[9] Kim Y. K. Effect of Pretreatment on Print Quality and Its Measurement, Digital Printing of Textiles, H. Ujiie Ed., Woodhead Publishing, Cambridge, England, p. 256

[10] Provost J, Freche M., Hees U., Kluge M., Weiser J.: Ink-Textile Interactions in Ink Jet Printing-The Role of Pretreatments, 2003 [Online] URL: http://provost-inkjet.com/resources/SDC%2B%2BInk%2BJetPretreatment%2B4th%2BDec%2B03.pdf [Accessed 13 April 2013]

[11] Jihyun B.: Color in Ink-Jet Printing: Influence of Structural and Optical Characteristicsof Textiles., PhD thesis, North Carolina State University, 2007 [Online] URL: http://repository.lib.ncsu.edu/ir/bitstream/1840.16/4425/1/etd.pdf [Accessed 13 April 2013]

[12] Choi P.S. Development of Optimum Printing System for Digital Ink Jet Printing, Hong Kong Polytechnic University , Master thesis, 2006 [Online] URL: http://hdl.handle.net/10397/3157 [Accessed 11 May 2013]

[13] Momin H. N. Chitosan and Improved Pigment Ink Jet Printing on Textiles, PhD Thesis, School of Fashion and Textiles, RMIT University, 2008, [Online] URL: http://researchbank.rmit.edu.au/eserv/rmit:6752/Momin.pdf [Accessed 13 May 2013]

[14] Zhang C. M., Fang K. J.: Influence of Penetration Depth of Atmospheric Pressure Plasma Processing into Multiple Layers of Polyester Fabrics on Ink Jet Printing, Surface Engineering, Volume 27, Number 2, March 2011 , pp. 139-144(6)

[15] Herascu N., Simileanu M., Radvan R. Colour Changes in the Artwork Materials Aged by UV radiation, In: Romanian Reports in Physics, 2008, Vol. 60, Issue 1, p. 95

[16] Morent R., De Geytern, Verschuren J., De Clerck K., Kiekens P., Leys C. Non-Thermal Plasma Treatment of Textiles, In: Surface and Coatings Technology, 2008, Vol. 202, Issue 14, p. 3427

[17] Yiqi Y., Vamshi N.: Improvement of the Lightfastness of Reactive Inkjet Printed Cotton, Dyes and Pigments, 74,1, 2007, pp. 154-160, http://dx.doi.org/10.1016/j.dyepig.2006.01.030

[18] Vizárová K., Reháková M., Kirschnerová S., Peller A., Simonb P., Mikulásik R., Stability Studies of Materials Applied in the Restoration of a Baroque Oil Painting, In: Journal of Cultural Heritage, 2011, Vol. 12, Issue 2, p. 190

[19] Varesano A., Tonin C. Improving Electrical Performances of Wool Textiles: Synthesis of Conducting Polypyrrole on the Fiber Surface, In: Textile Research Journal, 2008, Vol. 78, Issue 12, p. 1110

[20] Zarkogianni M., Mikropoulou E., Varellab E., Tsatsaroni E. Colour and Fastness of Natural Dyes: Revival of Traditional Dyeing Techniques, In: Coloration Technology, 2010, Vol. 127, Issue 1, p. 18

[21] Gun A. D., Tiber B. Colour, Colour Fastness and Abrasion Properties of 50/50 Bamboo/Cotton-blended Plain-knitted Fabrics in Three Different Stitch Lengths, In: Textile research journal, 2011, Vol. 81, Issue 18, p.1903

[22] Kašiković N., Novaković D., Karlović I, Vladić G. Influence of Ink Layers on the Quality of Ink Jet-printed Textile Materials, In: Tekstil ve konfeksiyon, 2012, Vol. 22, Issue 2, p. 115

[23] Kan C. W., Yuen C. W. M.,Tsoi W. Y.,Chan C. K.: Ink-Jet Printing for Plasma-treated Cotton Fabric with Biomaterial, ASEAN Journal of Chemical Engineering, Vol. 11, No. 1 (2011) pp 1-7

[24] Rat B., Majnarić I., Možina K.: Visibility of Care Labelling Code Symbols, Tekstil: Journal of Textile & Clothing Technology; Jun 2011, Vol. 60, Issue 6, p. 251

[25] Mikuž M., Šostar-Turk S., Pogačar V., Transfer of Ink-jet Printed Textiles for Home Furnishing into Production with Rotary Screen Printing Method, FIBRES & TEXTILES in Eastern Europe January / December 2005, Vol. 13, No. 6 (54)

[26] Zhang C, Fang K. Surface Modification of Polyester Fabrics for Inkjet Printing with Atmospheric-Pressure Air/Ar Plasma, In: Surface and Coatings Technology, 2009, Vol. 203, Issue 14, p. 2058

[27] Pallant J. SPSS Survival Manual, Allen & Unwin, Crows Nest NSW 2065, Australia, 2011

Experiments with a Newly Developed Biogas Reactor Block

Mónika Bakos-Diószegi[1], Miklós Horváth[1]

[1]Óbuda University, Donát Bánki Faculty of Mechanical and Safety Engineering, Népszínház utca 8, H-1081 Budapest, Hungary, E-mail: dioszegi.monika@bgk.uni-obuda.hu; horvath.miklos@bgk.uni-obuda.hu

Abstract: Increasing energy security can be found in the principal objectives of all countries. There is a competition in the research of long-term available renewable energy sources and their high efficiency utilization, and also in time. From the renewable energy sources the biomass, as an inexhaustible commodity was used in the experiments. Digestion was chosen from the several methods to explore green energy from waste and then to produce valuable biogas. However, the experiments related to anaerobic fermentation of organic materials take a long time, up to several months, so the results gained from these carefully prepared and accurately conducted experiments are definitely valuable. By optimizing the fermentation equipment it was possible to increase the time-efficiency of the measurements. The results are given in this paper.

Keywords: renewable energy; biogas; biogas reactor block

1 Introduction

Due to its natural features and agricultural and livestock past for centuries, Hungary is rich in biomass. This organic material is worth taking into consideration when selecting a method to produce biogas, which has not been used significantly in the country, so far. Experiments related to anaerobic fermentation of organic materials take a long time, so the results gained from these carefully prepared and accurately conducted experiments are definitely valuable. The fermentation block, the technological equipment used to conduct experiments, was optimized in order to increase the time-efficiency of the measurements. Some prototypes of the construction under development were tested several times, and finally a proprietary biogas unit was developed and installed, which is introduced below.

2 Pursuit of Energy Security

The word “security” comes from the Latin phrase "sercus", which also has worry-free, careless meanings. According to Barry Buzan, the famous English political scientist, security is the possibility and ability to survive and subsist against the threats of existence. Buzan divided the phrase security into five sectors: military, political, economic, social and environmental. The Military Science Encyclopedia adds the humanitarian, environmental and disaster recovery sectors to them. [1]

One of the basic pillars of economic security is energy, without which the Member States of the European Union, among them Hungary, would not exist. Everyday life of individuals and states would be impossible without energy. Consequently, the pursuit aiming ensure of energy supply, is also involved in energy security. The excessive energy demand of the world influences everyday life of developed societies. The largest consumers are affected the most by the lurch of energy security. More than half of energy consumption of the European Union is imported. 53.8% of energy consumption of the 27 member states has come from external sources. Net exporter during this period has only been Denmark. [2]

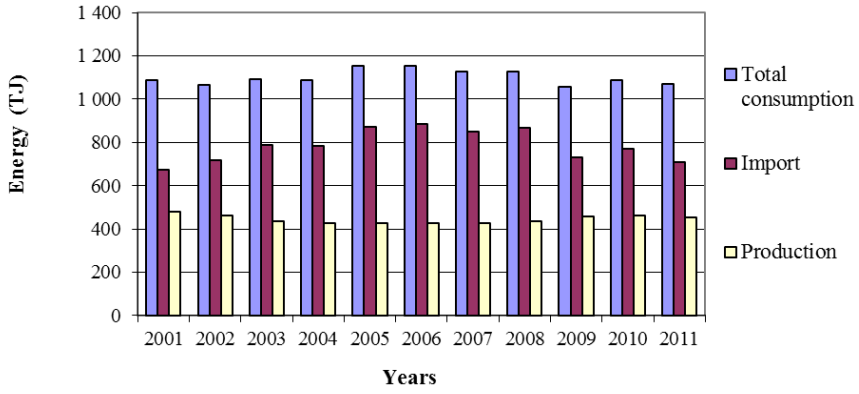


Figure 1
Energy balance of Hungary

The lack of coverage of domestic energy consumption derives especially from the increased consumption of oil and natural gas. Due to the gas program, that started at the beginning of the early 1990s, both the institutions and the population switched to gas consumption, significantly. As a result the gas dependence has increased by an additional 20% from 1995.

The increased demand for raw materials has brought a large amount of gas imports, which made our country even more dependent to unpredictable foreign relations in the region. On the basis of the energy balance in 2011, it can be concluded that 66% of the country’s energy consumption is imported. [3]

The growth of renewable energies in electricity production has been observed since 2003. Although it increased from 1% to 8.1% by 2010, Hungary is still among the last ones from the EU member states.

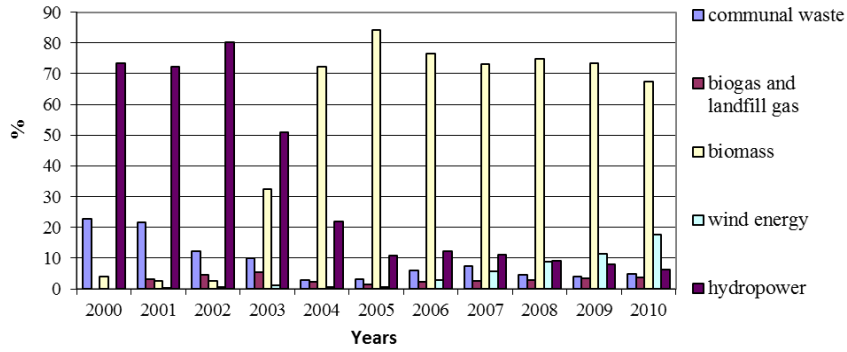


Figure 2
The proportion of renewable energy sources in electricity production

From the renewable energy sources the biogas has been used only from 2001 with a small amount of 6 TJ. This number was multiplied by 2010 to 1516 TJ, but still not close to the value inherent in the domestic biogas potential.

For comparison, in Sweden about 50% of the heat energy is produced from biogas today. [4]

Environmental considerations also encourage constructions of biogas power plants all over the world.

Methane, which is 21 times more damaging greenhouse gas than carbon dioxide, contributes with a ca. 20% to the anthropogenic greenhouse effect. More than 50% of methane originates from animal husbandry. Its average degradation time in the atmosphere (8 years) is much shorter than that of carbon dioxide (50-200 years). [5] A solution to the efforts to reduce or prevent its free access to the atmosphere can be the controlled fermentation of crop and livestock by-products. The methane, generated in power plants, will not enter the atmosphere, and what is more, it is the perfect raw-material for energy production.

3 Conditions of Organic Material Fermentation

Despite the fact that the application and utilization of anaerobic degradation of organic materials is not new, the research of biogas is its infancy both in Hungary and in the world. The chemical-biological process is simple, since the fermentation of organic material can take place anywhere, where the conditions of anaerobic rotting are given.

The degradation of complex organic materials takes place in several steps. First, the compound - sugar-chain polymer is broken up using exoenzymes of bacteria involved in the hydrolysis. The volatile organic acids obtained as the product of degradation are transformed further to acetate and hydrogen by acetogenic microorganisms. Approximately 70% of the final obtained methane is converted from acetate. The final step takes place strictly in anaerobic conditions, during which the methanogen microorganisms use up the ready-made volatile acids and produce methane and carbon-dioxide. The acetogens and methanogens exist in close symbiosis. The acetogenic bacteria prepare the nutrients for methanogens, which consume also the hydrogen. However, the presence of hydrogen affects unfavourably the operation of acetogens, so during their degradation methanogens stimulate the functioning of tribes located in front of them in the food chain. Consequently, the back and forth dependence between the different groups of microorganisms can be considered as syntrophic relationship, i.e. they not only help each other but they are interdependent in their diet.

In the process of biogas formation it is very important to provide parameters that stimulate degradation.

Biological conditions:

1. organic material
2. presence of acidogenic and methanogenic bacteria, including appropriate living conditions for them. [6]

The necessary living conditions:

1. anaerobic environment,
2. appropriate solid content of raw material (2-9% in the case of wet process),
3. constant, balanced fermentation temperature, (anaerobic biodegradation takes place either on mesophilic (30-35 ° C) or thermophilic temperatures (50-55 ° C) in a biogas unit that includes the fermenters),
4. continuous agitation to avoid a gas barrier film, local acidification and temperature difference, as well as to prevent foaming,
5. sufficiently chopped and mixed raw material(for better gas yield),
6. slightly alkaline pH, sufficient C / N ratio of the base material.

The aim of the research was to produce biogas in laboratory conditions.

During the experiments the effect of another factor, the stirring was also investigated. Altough it is not a condition of anaerobic degradation, digesters using organic materials should be stirred. The main effects of stirring in the case of biogas fermenters are that it prevents solids from settling in the fluid, homogenize the medium and makes biological processes more intensive. Thus, the surface available for bacteria responsible for gas production will increase and the volume of biogas produced will also increase. However, the energy demand of stirring is high, so its application is uneconomic. Laboratory experiments carried out in Germany proved that there was not significant difference in biogas yields

during anaerobic batch fermentation of corn silage when the stirring was intermittent and continuous. However, with intermittent stirring - 10 min stirring 230 minutes break - 29% energy saving was achieved in relation to the continuously stirred technology. [7]

In Denmark, a similar study was conducted with cow manure degradation. In this case, the volume of the generated methane was 12.5% higher when intermittent stirring was applied, compared with continuous stirring. [8]

On the basis of the above listed results and our own experiments the intermittent stirring was chosen, which is optimal in terms of biogas yield and energy use, too.

Conditions, discussed above had to be ensured for proper degradation. The system has operated more and more optimal, little by little, and was modified according to the results of the previous test case.

4 Implementation of the Experiment

Conditions needed to the optimal degradation process, mentioned above, are provided by a fermenter. In biogas plants this task is carried out by a huge concrete tank, in which the biogas is generated in temperature controlled, hermetically sealed and stirred conditions, usually with continuous input/output of the base material.

Model 1

In our laboratory it was an anaerobically sealed, heat-resistant glass vessel of 1 litre that served as a fermenter (1), in which we could measure the temperature and the pH continuously (3). There was a gas collecting bag (8), silicone tubes (4), a 3-way valve (6) and also a sampling cap (7) necessary to sampling gas. At the beginning we could not solve stirring (2), and we could not afford buying a gas meter (5).

The gas sampling component functioned properly, which was confirmed by the gas chromatographic analysis. It was a good choice to use the commercially available bags for collecting the gas, because it can be recycled, avoiding environment pollution. The hermetically sealed vessel, the pH-meter, the silicone and connected fittings also operated properly.

The primary issue raised by putting the basic concept into practice was the measuring the quantity of the gas generated. We could not afford buying a gas mater suitable for detecting low flow, laboratory quantity gas, because of their high prices. It was also a problem that the fermentation liquid has not been heated yet, and so it could not provide the optimal mesophilic (37 ^{0}C) degradation

temperature. Furthermore, it appeared urgent to solve the stirring, to provide a homogeneous degradation process.

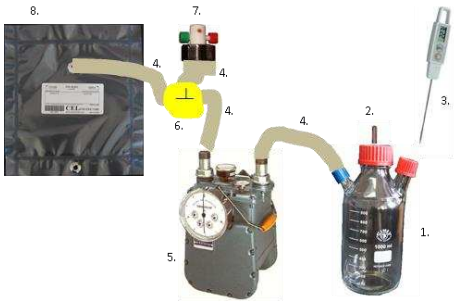


Figure 3
Model 1 [9] [10] [11]

Model 2

Model 2 was designed to eliminate the shortcomings and problems arose during operating Model 1. To provide insulation and heating combined with stirring the construct shown below was developed (Figure 4).

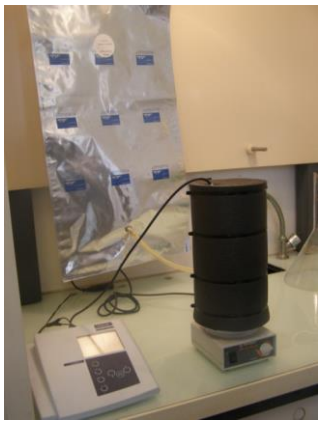


Figure 4
Model 2

The installed heating magnetic stirrer appeared to be a good choice for heating; however it was not suitable to provide the sufficient homogeneity of the substrate. The generally resulting 7-8% dry matter content was too high, so it was necessary to find a new solution for stirring the raw material. The polyfoam insulation was acceptable for experiment, but we have not been able to find a good solution to measure the quantity of the generated gas in that phase of the research.

Model 3

The main difference compared to the first two variants is a construction suitable for measuring the volume of the generated gas and for keeping the constant temperature of 37°C. The homogeneous and constant temperature was achieved by placing the fermenter into a dryer which was used as an incubator cabinet (Figure 5). The dryer ensured the constant 37 °C +/-1°C, required to the measurements, perfectly.

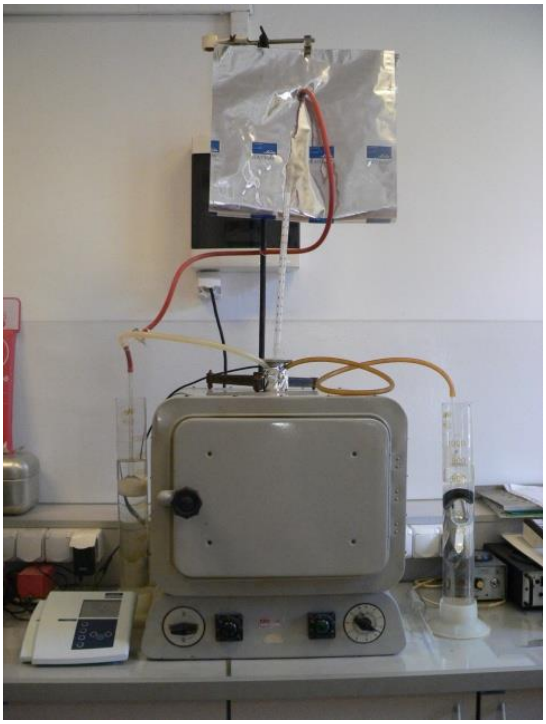


Figure 5
Model 3

To determine the volume of gas produced, the volume displacement method was used, which is very easy to implement and its accuracy is in accordance with standard recommendations.

The volume of the generated gas equals to the volume of liquid displaced, and its pressure without compression is 101325 Pa. The apparatus consisted of a graduated cylinder of 1000 ml, and placed inside the „ squealer", an open ended glass cylinder with a smaller cross-sectioned, tight nozzle. The gas flowed from the fermenter into the „squealer" from the silicone tube, which was led out of the dryer. As a result, the volume of the supersaturated NaCl solution, which was

originally at the same level in the graduated cylinder, has changed. The biogas flow generated by fermentation pumped a volume equal to its own volume from the "squealer" to the graduated cylinder. The volume of the gas produced was measured with an accuracy of 10 ml.

Figure 6
The dryer with the fermenter inside

Model 3 could perfectly provide the constant experimental temperature and also the measurement of the volume of gas generated.

However, during testing Model 3 it was found that to ensure homogeneity of the sewage sludge and preventing foaming, stirring must be provided. Earlier it was replaced by shaking the vessels in the same intervals.

5 Implementation of a New Construction

Significant changes were made in the construction of the new model. Our main objective was to make a device that complies with measuring procedures in the fermentation standard (VDI 4630). According to the standard the degradation process has to take place in several glass vessels simultaneously. We planned to build 8 fermenter blocks, but the dryer appeared to be too small for them, and so a plastic crate filled with water was used for controlling the temperature of the fermenter. The number of the openings of fermenter glasses was also changed. 4 openings with threaded silicone caps were needed for the stirring, input, output and gas outlet.

Figure 7
The complete construction

It was also a significant change compared to the previous constructions that the stirring was solved in the 8 fermentation glasses. The 8 stirrers were built into the largest, center pipe, worked simultaneously, driven by a wiper motor. The required number of rotation was provided by a drive unit consisting of 8 gears.

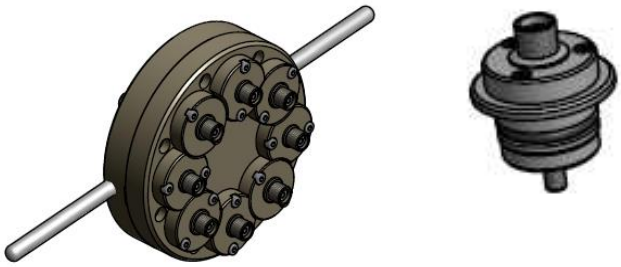


Figure 8
The gear box and the gas-tight connector plug

The shaft coming out of the reduction gearbox was connected to the stirrers by Bowden cables with form-locking connection and with gas-tight connector plugs, built in the caps. The other connectors of the glass vessels were fitted with quick coupling, which ensured adequate gas barrier.

Figure 9
The glass fermenter

Although we kept the principle of positive displacement flow, the „squealer" construction was no longer used for the gas flow measurements. From each fermenter (1) placed in a water bath a silicone tube (2) led the generated gas (3) out. This tube was led to the top of another glass bottle which contained supersaturated NaCl solution (4). There was another tube (5) from the bottom of that bottle, through which the gas pumped the solution to the graduated cylinder (6). This way, it was possible to measure the volume of the generated gas accurately, on a daily basis, according to the standard specifications.

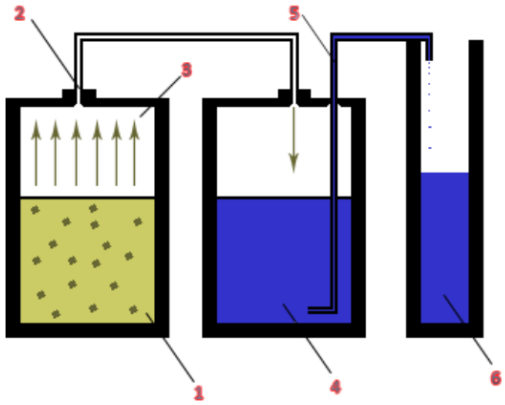


Figure 10
Gas volume meter

To keep the temperature of the water bath constant in the plastic crate a heating element, a thermostat and circulating vanes were used. When the temperature of the wate fell below the settled value of 37 °C, the sensor switched on the heating and circulating at the same time.

Figure 11
Insulation and water circulating

To increase the inertia of the temperature and to prevent the fermenters floating up in the water, a polyfoam board (with holes cut for the bottles) was fitted on the bottom of the crate.

6 Test Experiment with the Newly Developed Biogas Fermenter Block

The experiment was performed by using four fermenters of net one liter each, which were placed into water bath. Sludge taken from waste water channels in Budapest was used as seeding sludge, and chopped then wheat straw, fractionated by sieves, was used as substrate.

The inoculum of 2,55% dry matter and 64.63% organic dry matter content ensured the microorganism medium which was to help the degradation.

The wheat straw had 95.78% dry matter and 93.41% organic dry matter content. Each fermenter was examined at net volume of 700 ml.

Fermenters 1 and 2: with wheat straw of 0,125-2 mm particle size and sludge, without mixing.

Fermenters 3 and 4: with wheat straw of 0,125-2 mm particle size and sludge, with 2 min stirring 60 minutes break.

Figure 12
The four fermenters filled with sludge

The middle threaded orifices of bottles 1 and 2 were closed by silicone sealing sheets. In bottles 3 and 4 there were stirrers that were hanged in as far as 1/3 of the bottle height from the bottom and they were to make the digested substrate homogeneous.

The fermenters with wheat straw-seeding sludge substrates were placed into a temperature controlled bath of 37^0C.

The produced gas from the four fermenters was led into four additional collector bottles where because of the increased pressure of the gas the liquid was displaced. The displaced liquid was collected in measuring cups, so the volume of the produced gas could be read with an accuracy of 1 ml.

During the experiment the continouosly generated gas was collected in a special gas storage bag by using a three-way pipe. In the course of the repeated 24-day experiment samples were taken from the daily „fresh" gas.

It can be concluded that due to the stirring, that made the substrate more homogenous, the efficiency of digestion increased.

The gas yield increased from 180 ml/goTS to 223 ml/goTS. This value corresponds to the value recorded in the literature for average biogas yield of wheat straw. [12]

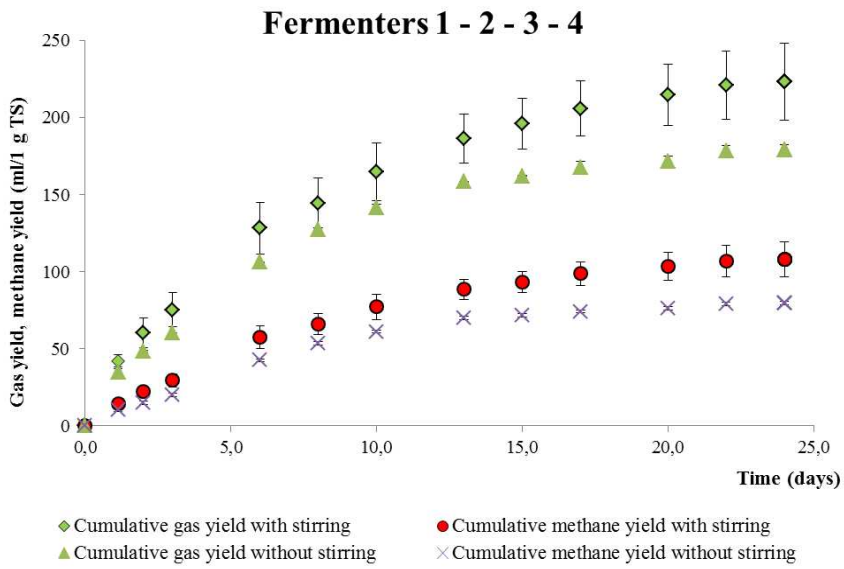


Figure 13
Gas yield and methane yield with (1 and 2) and without (3 and 4) stirring

Fermenters 1 and 2 showed a minimum variance in respect of both the biogas and methane yields. At the end of the 24-day parallel measurements a 3.2 ml/goTS biogas and a 2.5 ml/goTS methane yield deviation was experienced.

A more significant deviation was recorded in respect of the biogas and methane yields in the case of fermenters 3 and 4. The total variance of the cumulative gas yield value reached 22.1 ml/goTS, which was 9.9% of the generated total value of 223 ml/goTS biogas. In the case of methane content the cumulative variance was 11.4 ml/goTS, which was 10.6% of the generated total value of 107.9 ml/goTS.

Summary

The results of the experiment proved that the operation of fermenters 1 and 2 without stirring was stable and gas-tight. Fermenters 3 and 4, which were equipped with stirring vanes, and which produced approximately 10% variation in both, biogas and methane yields, require further constructional modification.

At the end of the developing process the final, automatic and discontinuous biogas reactor block was achieved. The developing process was performed according to Guideline VDI 4630 (Fermentation of organic materials, Characterisation of the substrate, sampling, collection of material data, fermentation tests) [13].The realization was performed in several stages, which were considered as milestones. In order to make our research work even more effective and representative our further objective is to make the operation of the biogas fermenter block more reliable, regarding gas-tightness stability.

Acknowledgements

Thank you for helping certain subtasks of the developments: Klaudia Kormos, Ferenc Haraszti, Zoltan Laky and Otto Etler.

References

[1] Magyar Hadtudományi Társaság: Hadtudományi Lexikon I. kötet. Szabó József (főszerk.). – Budapest, 1995. – ISBN 963 04 5227 8

[2] Statisztikai tükör IV. évfolyam 58. szám http://www.ksh.hu/docs/hun/xftp/stattukor/energiaarak0409.pdf

[3] Magyarország megújuló energia hasznosítási cselekvési terve 2010-2020 A 2020-ig terjedő megújuló energiahordozó felhasználás alakulásáról, Nemzeti Fejlesztés, 2010 december Budapest http://geotermia.lapunk.hu/tarhely/geotermia/dokumentumok/national_renewable_energy_action_plan_hungary_hu.pdf

[4] Biogas Road Map for Europe, http://www.aebiom.org/IMG/pdf/Brochure_BiogasRoadmap_WEB.pdf

[5] Biogas from waste renewable resources, Dieter Deublein, Angelika Steinhauser; http://zorgbiogas.com/upload/pdf/Biogas_from_Waste_and_Renewable_Resources.pdf

[6] Bai A.: A biogáz, 2007 Száz magyar falu könyvesháza, ISBN 978-963-7024-30-6

[7] Alexandra Kowalczyk, Eva Harnisch, Sebastian Schwede, Mandy Gerber, Roland Span - Different mixing modes for biogas plants using energy crops Applied Energy (2013) Volume 112, December 2013, pp. 465-472

[8] Prasad Kaparaju, Inmaculada Buendia, Lars Ellegaard, Irini Angelidakia - Effects of mixing on methane production during thermophilic anaerobic digestion of manure: Lab-scale and pilot-scale studies, Bioresource Technology - Volume 99, Issue 11, July 2008, pp. 4919-4928

[9] http://www.conrad.hu/pic.php?pid=123320&image=1

[10] http://www.celscientific.com/files/Tedlar_Bag.jpg

[11] Laboratóriumi gázáramlás mérő berendezés http://www.elster-americanmeter.com/en/707.html

[12] Kaltwasser, B. J. - Biogáz előállítás és hasznosítás. Budapest, Műszaki Könyvkiadó, 1983

[13] VDI 4630 (2006): Fermentation of organic materials – Characterisation of the substrate, sampling, collection of material data, fermentation test

Application of AdaBoost Algorithm in Basketball Player Detection

Branko Markoski*, Zdravko Ivanković*, Ladislav Ratgeber, Predrag Pecev*, Dragana Glušac***

*University of Novi Sad, Technical faculty "Mihajlo Pupin", Djure Djakovica bb, 23000 Zrenjanin, Republic of Serbia, markoni@uns.ac.rs, zdravko@tfzr.uns.ac.rs, pecev@tfzr.uns.ac.rs, gdragana@tfzr.uns.ac.rs

**PTE-ETK University of Health Sciences Pécs - Doctor School, Havihegyi út 4, 7627 Pécs, Hungary, ratgeber@ratgeber.hu

Abstract: Video materials contain huge amount of information. Their storage in databases and analysis by various algorithms is a constantly developing area. This paper presents the process of basketball game analysis by AdaBoost algorithm. This algorithm is mainly used for face and body parts recognition, and was not tested on player detection in basketball. It consists of a linear combination of weak classifiers. In this paper, we used stumps, i.e. decision trees with only one level as such classifiers. The aim of this research is to assess the accuracy of this algorithm when applied in player detection during basketball games. We examined the capabilities of AdaBoost algorithm on a video footage obtained from the single moving camera, without any previous processing. First training was performed using images of a basketball player's entire body (head, legs, arms and torso), while the second training was performed using images of a head and torso. By applying the algorithm to the given set of images that include head and torso, the algorithm obtained an accuracy of 70.5%. Training on the set of entire body images was not successful due to the large amount of background that goes into the training, and which represents noise in training process. This research concluded that AdaBoost could not be applied to object detection in sports events. We also concluded that this algorithm gives much better results when applied on simpler objects (like face recognition) and that its application could be in detection of players' body parts or as a first step in object detection in order to eliminate as much area as possible. Its application in detecting players' upper body or entire players gives large number of false positive, which makes algorithm inapplicable in real situations.

Keywords: AdaBoost; Object detection; Basketball

1 Introduction

With the advance of information technology, the amount of created, transmitted and stored multimedia content constantly increases. As a result, the multimedia content is widely used in many applications. Therefore, there arises the need for

its organization and analysis, both from commercial and academic aspects. Computer vision represents a technology that can be applied in order to achieve effective search and analysis of video content.

Computer vision represents a process that consists of several phases [1]. First phase is initialization, which is a process of removing background and extracting objects of interest by creating their models using markers, images or predefined shapes. Next phase is tracking, which is a process of object recognition in successive frames. This phase lasts until the object leaves observed area, or until the tracking is terminated. Third phase is pose estimation, which in process of human recognition and represents analysis of the arms, legs, torso and head, according to which the object is classified into one of the previously defined poses. The final phase is recognition. Recognition can be achieved by recognizing person's face or some other characteristic feature. Computer vision, therefore, consists of four phases (initialization, tracking, pose estimation and recognition), but the subject of this study is the first phase, e.g. initialization, which represents the model creation.

The model is created by players' images, which are obtained using specially developed software. This software stores one frame from video material in every 0.5 seconds on provided location in computer. From those frames, we cut rectangles that contain basketball players. These rectangles will be used in training process. Training is done with AdaBoost algorithm, which represents an often-used algorithm in the shape recognition. It is primarily used for face and body parts recognition. The aim of this study is to assess its capabilities in order to detect players in basketball games. Images of basketball players are objects with high degree of diversity, depending on whether the player has the ball, plays defense, shoots on the basket, jumps for the ball etc. Therefore, the training set has many variations, and the goal of this research is to assess whether AdaBoost can be successfully applied in such training process. Its main feature is execution speed, which is especially important when analyzing basketball games. Coaches often want to have complete analysis of the game as soon as possible, in order to make some changes in their team play. For that reason, the use of this algorithm in the process of basketball game analysis would be useful. AdaBoost requires a large training set. Therefore, six thousand positive examples (images of basketball players) and six thousand negative examples (images that do not contain any basketball player) were used. We have combined positive and negative images in testing process (images of basketball players are "glued" over negative examples). Images combined in this way are used in order to assess the performance of AdaBoost algorithm.

The second chapter in this paper provides an overview of application of data mining in sports, as well as research related to computer vision. Basketball player recognition belongs to the broader group of analysis in computer vision that is called human motion capture. The third chapter explains AdaBoost algorithm and its variations in order to achieve better performance. The fourth chapter contains

the procedure for training AdaBoost algorithm, which consists of creating training set, marking positive and negative examples, training and testing. The fifth section contains conclusion remarks and further stages of computer vision that can be applied in order to create solution that can be applicable in practice.

2 Data Mining in Sport

Data mining in sport is experiencing rapid growth in recent years and is gradually attracting the attention of largest sports associations. Baseball team Boston Red Sox and football club AC Milan were among the first organizations that started to apply the benefits of data mining in the sports. Special merits for the introduction of data mining in the sport belong to Dean Oliver, who introduced this methodology in basketball [2], and Bill James, who did the same in baseball [3].

Before the implementation of data mining, sports organizations have relied almost exclusively on human factor. They believed that experts in a given field (coaches, managers, scouts) could successfully convert collected data into practical knowledge. As the amount of collected information increased, these organizations started looking for methods that are more practical. Appropriate usage of large amounts of data available to sports organizations led from engaging additional statisticians to adopting techniques of data mining. Application of data mining can lead to better overall team performance, by analyzing the behavior of players in certain situations, determining their individual impact, revealing the opponent's tactics and pointing possible weaknesses in play.

According to Schumaker et al. [4] in the next few years, the application of data mining in sports will face several challenges and obstacles. The biggest obstacle will be to overcome opposition to new technologies that is present in some sports organizations.

While the use of statistics in decision-making is certainly an improvement over the use of instinct of coaches, managers and scouts, statistics alone can go in the wrong direction without knowledge of the problem domain. The first part of the problem is to determine the performance metrics. A large number of existing sport metrics can easily be used inappropriately. Ballard [5] has presented a typical example of inaccuracy in data collection in basketball. He gives an example of a jump in defense, which represents the number of times a player catches the ball in defense after opponent’s unsuccessful shot. In order to record the jump in defense, teammates have to block opponent players and keep them away from the basket, but only the player who catches the ball is awarded with the rebound. The second part of the problem is to find interesting patterns in data. These patterns may display movement and intentions of opponent players, reveal the beginning of injury during training or predict outcome of observed game. A practical method in finding those patterns could be application of neural networks [6] [7].

With the development of technology, sports events have become available in digital form as part of multimedia databases. Search of video and multimedia content is becoming more common in sports due to large number of available tools. Automated methods of detection are used for parsing video content, and translating it into a form that can be searched [4] [8].

Traditional sports statistics has quickly become insufficient in comparison to the advantages of multimedia technology [9]. In recent years the usage of videos for recording certain events for later analysis, has become a common place. For example, baseball players in American professional league visit a team multimedia room and study different ways in which the pitcher sends the ball, in order to prepare for new game or to play correctly during the match [10]. Another technology that allowed faster analysis and transfer of video materials and knowledge obtained by their analysis is the internet [11] [12]. Due to these technologies, videos are almost immediately available to players, coaches and scouts.

Multimedia analysis of sports events is presented in menu scientific papers. Among them, there are almost no papers about application of AdaBoost in basketball games. Lehuger et al. have applied AdaBoost in soccer [13], while Zahid et al have applied AdaBoost in baseball [14].

Our goal is to develop system that would automatically gather knowledge from footages of basketball games. In order to achieve this, the first step is to recognize players on the court. Next steps would involve court detection, ball and basket detection and analysis of ball trajectory.

3 Adaboost Algorithm

Boosting takes its origin from the theoretical framework for studying machine learning called "PAC" (Probably Approximately Correct) developed by Kearns and Valiant [15]. They were the first who questioned whether "weak" learning algorithm, which behaves slightly better than random guessing, could be a building block for general accurate "strong" learning algorithm.

AdaBoost, short for Adaptive Boosting, is a machine-learning algorithm first formulated by Freund and Schapire [16]. It is adaptive in the sense that classifiers that come in next for execution are adjusted according to those instances that were wrongly classified width the previous classifiers. It is sensitive to noisy data and information that does not belong to the required set. However, in some situations, this algorithm may be less susceptible to memory input set in comparison to many other algorithms. AdaBoost calls the weak classifiers repeatedly, performing a series of $t = 1,...,T$ classifiers. In each execution, "weight" calculated by incorrectly classified examples increases (or, alternatively, weights of each

correctly classified examples decreases). New classifiers are constrained to focus on those examples that were incorrectly classified by previous classifiers. This is a meta-algorithm that can be used together width a number of other algorithms in order to improve performance. Pseudo-code for AdaBoost is given in Listing 1.

Listing 1. Pseudo code for AdaBoost algorithm.

Given: $(x_1, y_1),...,(x_m, y_m)$ where $x_i \in X, y_i \in Y = \{-1,+1\}$

Initialize : $D_1(t) = 1/m$

For $t = 1,...,T$:

- Train weak learner using distribution D_t
- Get weak hypothesis $h_t : X \to \{-1,+1\}$ width error

  $\varepsilon_t = \Pr_{i \sim D_t}[h(x_i) \neq y_i]$

- Choose:

  $$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

- Update:

  $$D_{t+1}(t) = \frac{D_t(i)}{Z_t} \times \begin{cases} \varepsilon^{-\alpha_t}, if\ h_t(x_i) = y_i \\ \varepsilon^{\alpha_t}, if\ h_t(x_i) \neq y_i \end{cases} = \frac{D_t(i)\exp(-\alpha_t\, y_i\, h_t(x_i))}{Z_t}$$

  Where Z_t is a normalization factor (t is chosen so that D_{t+1} will be a distribution). The output is the final hypothesis:

$$H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t\, h_t(x)\right)$$

The algorithm receives as input some training set $(x_1, y_1),...,(x_m, y_m)$ where each x_i belongs to a particular domain or instance space X, and each label y_i is in label space Y. In most cases, it is assumed that $Y = \{-1,+1\}$, except when looking at extending of AdaBoost with more classes. AdaBoost calls "weak" learning algorithm repeatedly in $t = 1,...,T$ execution steps. One of the basic ideas of the algorithm is to maintain a distribution or set of weights over the training set. Weighting distribution in training example *i* in step t is denoted by $D_t(i)$. At the beginning, all the weights are placed on the same value, but at each step, the weights of incorrectly classified examples are increased and the weak learning algorithm is forced to focus on more difficult examples in training set.

The task of the weak learning algorithm is to find a weak hypothesis $h_t : X \to \{-1,+1\}$, which corresponds to the distribution D_t. The accuracy of the weak hypothesis is measured by its error:

$$\varepsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i] = \sum_{i:h_t(x_i) \neq y_i} D_t(i) \quad (1)$$

The previous expression shows that the error is measured in accordance to the distribution D_t over which the weak learning algorithm was trained. In practice, the weak learning algorithm can be any algorithm that uses weights D_t from the training set.

Looking at the example of recognizing players in basketball games, x_i represents a player (standing motionless on the court, shooting to the basket, jumping for the ball, playing defense, attempting dribbling penetration, etc.), while labels y_i show whether a given detection represents a basketball player or something else in the frame. Weak hypothesis is presumption that certain objects are players, and sub collections examined by that hypothesis are selected according to the distribution D_t.

When it comes to the hypotheses h_t, AdaBoost determines parameter α_t. Intuitively, α_t measures importance assigned to the hypothesis h_t. Listing 1 shows that $\alpha_t \geq 0$ if $\varepsilon_t \leq 1/2$, and that α_t increases its value for error ε_t becomes smaller.

Next step is to update distributions D_t by using rules shown in Listing 1. The effect of this rule is to increase the weight of examples that are misclassified by the hypothesis h_t, and to reduce weights of well-classified examples. Thus, the weights are trying to concentrate on "harder" examples.

The final hypothesis H is weighted majority of votes of T weak hypotheses, where α_t represents weight assigned to hypothesis h_t.

In practice, AdaBoost has many advantages. It is fast, simple and easy to program. AdaBoost does not have any parameters that have to be adjusted separately (except for the number of steps T). On the other hand, actual performances of boosting in a particular problem are largely dependent on data and weak learning algorithm. In accordance with theory, boosting can give wrong results when there is not enough data for training, when weak hypothesis are very complex, or when weak hypothesis are too weak.

In 2001, Viola and Jones [17] presented their work that was a milestone in the implementation of AdaBoost algorithm. Their work had three important contributions in fast and accurate image analysis:

- Appliance of integral images
- Learning classification functions
- Cascade creation

3.1 Stump

In this paper, we used stump as a weak learning algorithm. Stump is a machine-learning model that consists of a decision tree width one level. In other words, it is a decision tree width one internal node (root node) that is directly related to end-nodes. Stump gives a prediction based on the value of a single input. It is sometimes called 1-rule. An example of stump is shown in Fig. 1.

Depending on the type of input characteristic, there are several variations. For nominal characteristics, a stump can be created that contains a leaf for each possible characteristic value, or stump width only two leafs where first leaf corresponds to one category of results and second leaf corresponds to all other categories. For binary characteristic, these two approaches are the same.

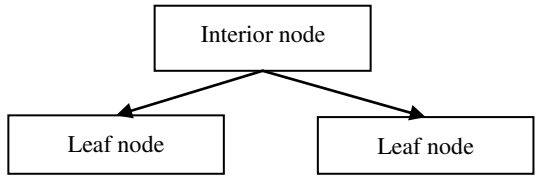


Figure 1
Functional structure of the system that analyzes the movements of the human body

4 Creating Model of Basketball Players

In order to create a model of basketball players, AdaBoost algorithm is applied over the training set, without any previous processing. During the process of training a new classifier, it is necessary to go through several stages:

1. Image acquisition
2. Example creation
3. Training
4. Testing

4.1 Image Acquisition

Different sources can be used in image acquisition process. Since the aim of this paper is player detection in basketball games, the training process uses videos broadcasted by television stations. These videos are stored in multimedia database. In order to apply AdaBoost algorithm we use additional software. The purpose of this software is to capture one frame from the video material in every 0.5 seconds and store it in predefined location on user's hard disk. We used

SampleCreator software, which is implemented in C# programming language and .NET 4.0 framework. Frames were captured and stored using DirectShow technology. A game in NBA league lasts 48 minutes (four quarters of 12 minutes). In addition to this playtime, there are number of interruptions (fouls, time outs, breaks between periods). Including those, we can assess empirical value of the average game duration around 100 minutes. By applying this method of storing frames from single game, we get 100 * 60 * 2 = 12.000 frames. This provides a sufficient amount of data to train AdaBoost algorithm. This algorithm works width black and white pictures so color of shirt itself is not a major problem and an already trained algorithm can be applied on large number of games.

By using DirectShow technology, we obtain images that will mainly serve as positive examples (images that contain objects of interest). However, during the game we have frames that do not contain any player. These frames can be used as negative examples. In addition, negative examples could be found in other sources. It can be any picture without basketball players. Fig. 2 shows positive and negative examples obtained from a basketball game.

Figure 2
Positive and negative example obtained from the basketball game

Besides positive and negative examples, some frames may not fall into any of the above categories. In most cases, those frames contain basketball players with the poor display quality. Their use in the training process could direct AdaBoost algorithm in the wrong direction, because the algorithm is sensitive to the noise data. Those frames should be excluded from the training process.

4.2 Example Creation

Examples are objects of interest, which are used in the training process. Depending on the look of required objects, different authors have used different sizes of training sets: 5000 examples in face detection [18], 6000 examples in pedestrian detection [19]. Training set creation is performed by cutting objects from the frames of basketball game. SampleCreator software is used for this purpose. It allows marking objects, while maintaining the ratio between width and height. Fig. 3 shows the process of marking the whole basketball players that will be used in training, as well as the process of marking player's upper body.

Different ratios of height and width were used in marking process. In marking the whole players, ratio of 2.2 was empirically determined (e.g. width 100 px, height 220 px), while in marking the upper body, that ratio was 1.8 (e.g. width 100 px, height 180 px). The ratios were determined in order to optimally cover player on the observed frame. The ratio of player in one training set must be fixed, because all images in training process are scaled to same size. If ratio is not the same, some images are stretched while others are elongated.

Figures show that not all players are marked as positive examples. The reason is that some objects may adversely affect the training if they are not shown clearly, if they are entering or leaving the frame, or if other objects obscure them. If we compare images on Fig. 3, it is evident that we have not marked the same players. Some players are not marked because of the background that would enter in the training process, which would adversely affect the AdaBoost algorithm. Unmarked players will not interfere in the training process because we crop positive examples from the picture and "glue" them on negative examples at pre-defined locations.

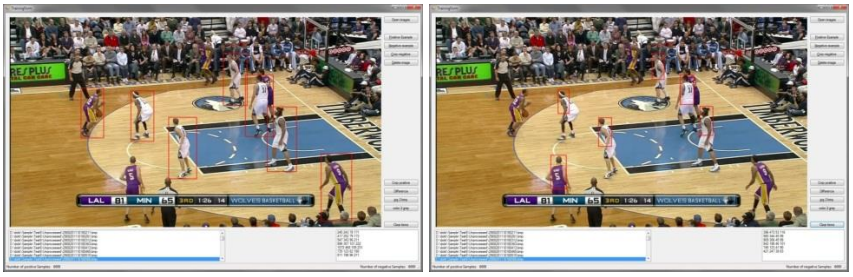


Figure 3
Marking of whole body and upper body of basketball players by SampleCreator software

4.3 Training Example Creation

In order to put labeled examples into a training set, they should be “cut” from the observed images. In addition, example size can be changed, or we can apply function that will distort the image in certain limits. The program is not able to create more training examples based on one picture. It is therefore necessary to have a sufficient number of examples, which is in various applications ranging from 5000 to 7000. The marked areas, which will be used in the training process, are shown in Figs. 4~5.

Figure 4

Images of whole basketball players' body that will be used in the training process

From the previous image, it can be seen that background takes a large area in the training set. Size of this area is different depending to whether a player is close to the paint (the area on the ground painted a darker color), close to the audience (the audience is much darker than the floor), or whether there are other players in its vicinity. In addition, during the game, basketball player can be in different positions depending on whether he is running, having the ball, playing defense or taking a shoot. Depending on that, their limbs can be in different position, which makes a training process much harder. Because of this, training was performed width another set of examples that included basketball player's upper body. In this training, limbs were excluded, which reduced dimensionality of the search problem. It also reduces the amount of background that is essentially a noise. Those examples are shown in Fig. 5.

In our research we have used 6000 positive examples that contain players whole body and 6000 positive examples of players upper body in order to train algorithm.

Figure 5
Images of basketball player upper body that will be used in the training process

4.4 Testing Example Creation

Testing set can be created by putting one positive example over a negative example. In testing process, we used a new set of 1000 positive examples. During this process, a positive image can be resized or distorted. In this way, the application knows the exact position of the positive examples, based on which it can provide assessment of whether the algorithm recognized the positive example on the observed image. Examples of test images are given in Fig. 6.

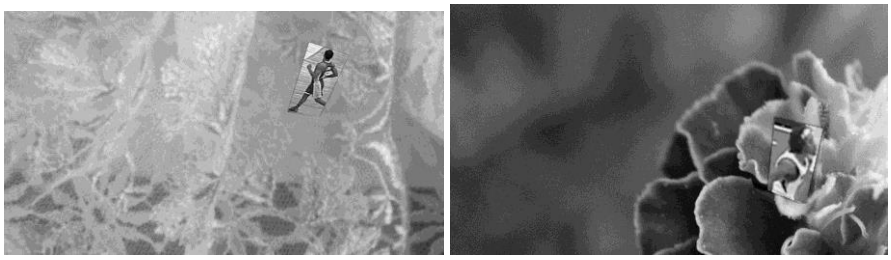


Figure 6
Testing examples width whole body and upper body of basketball player

4.5 Training

In training process, we apply AdaBoost algorithm over the previously marked examples. Regarding the size of examples in training set, Kuranov et al. [18] have shown that the best results are achieved when the dimensions of examples that

contain faces are reduced to 20x20 pixels. In the training process with images that contain whole basketball players, the ratio between the width and height cannot be 1:1. Therefore images were reduced to 20x44 pixels in training with the entire basketball players (574,479 features), and 20x36 pixels for training with player's upper body (392,394 features).

The same authors have also suggested a 20-stages training. If as a training parameters we use degree of false positive of 0.5 and detection rate of 0.999, after the entire training we can expect degree of false positive of $0.5^{20} \approx 9.6e$ - 07, and the detection rate of . $0.999^{20} \approx 0.98$

During AdaBoost training on the examples that have symmetry (human face), we can apply some type of optimization that significantly speeds up the processing. This is due to the fact that in these cases only one-half (left of right) of the Haar feature is used. However, although players are objects that have symmetry, when observed during the game and in all positions in which they could be found, the application of symmetry in the training would not lead to desired results. In order to achieve higher accuracy, we used the extended set of Haar features [20] (vertical features and features rotated by 45 degrees).

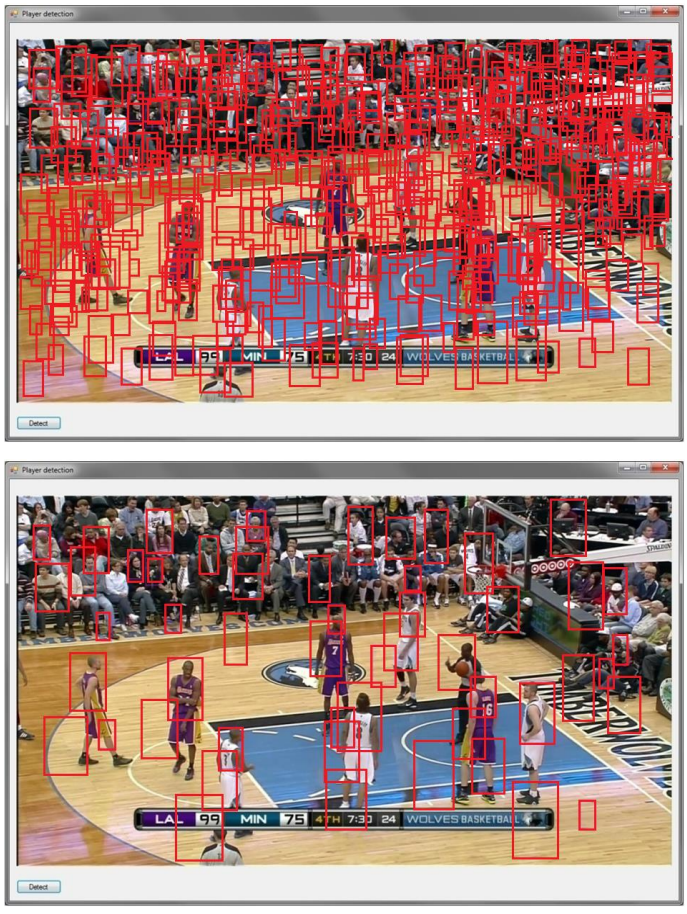
GAB (Gentle AdaBoost) classifier is used in the training that puts less emphasis on outliers [21]. The main reason was the work of Kuranov et al. [18], in which they have proven that GAB algorithm achieves highest results in object detection. This classifier is also the fastest one considering the time required for training.

During the training process, we used six thousand positive examples and six thousand negative examples. Training can be completed in several sub-stages when the minimal desired degree of search is fulfilled, or when the degree of false positive is reached, because the additional phases will certainly reduce this level (0.99 after current phase * 0.99 for next phase = 0.981 after next phase). A valid algorithm is the one that rejects all incorrect examples.

During the training process over the set that included images of whole basketball players, the algorithm failed to reduce the level of false positive below 0.5. The algorithm continues training until the percentage of the original examples used in training falls to 0%, i.e. until all the examples are used. This would lead to interruption of the training without the wanted result. The reason is, primarily, a large degree of diversity that is encountered in the training set. This diversity is reflected in the position of players depending on whether they are standing, walking or running, and whether they have the ball, play an active defense or shooting on the basket. Another reason is a relatively large amount of background, which is located in the training examples, and which represents a noise.

Training on the set that contains players' upper body was successfully finished. This set does not include player arms and legs and is therefore much more balanced. In addition, the amount of background in these images is far lower, as well as the noise.

If the results of training are included in the application and executed on an arbitrary example taken from a basketball game, we get the result as shown in Fig. 7. In figure we can see three images. First image is taken after twelve of twenty total stages in training process. After this step, all players are detected, but also very large number of false positive is detected (algorithm detects almost everything on the image). The second image is taken after eighteen steps in training process. After this stage, algorithm also detects all players, but there are almost twenty times more false positive detections. From the third image it can be seen that the algorithm correctly detects approximately two thirds of the players. In addition, the algorithm still gives a large number of false positive in the analysis of the audience (identified as basketball players). There are seven times more false positive then true positive detections. Among the identified objects are also some unexpected results, such as parquet parts shown in one color. The reason for such poor result is large number of diversity in training set.



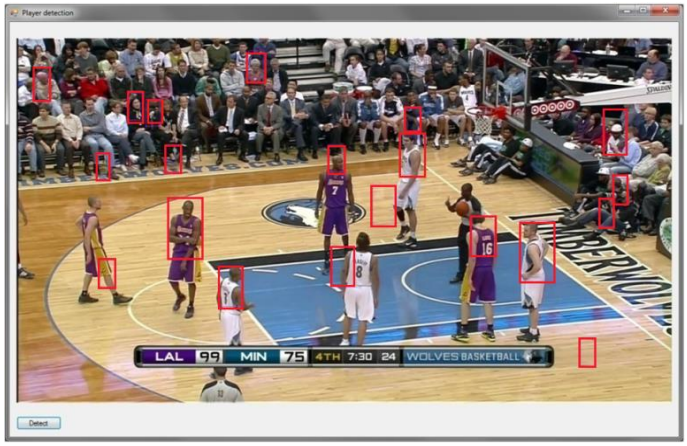


Figure 7
Result of algorithm application

4.6 Testing

In order to assess the performances of a trained classifier, a set of testing images was used. Those images have precise locations for each object of interest. The tool that is used to assess performance as input receives a collection of tagged images over which the classifier is applied. The performance of output is obtained by calculating number of found objects, number of objects that are not found and number of objects that are incorrectly classified as positive.

In this paper, we measured performances of training over the set that contains upper body of basketball players, because training over the set that contains whole players was not possible. With the obtained results, a ROC (Receiver Operating Characteristic) curve was created. It graphically represents sensitivity, i.e. ratio of true positive versus false positive, for a binary classification system.

The output of the testing tool shows the number of objects that are identified (Hits), the number of objects that are not identified and which represents false negative (Missed), and the number of objects that are classified as positive, but actually are not required objects (False). Looking at all images that were used in the training process, we get that the algorithm has successfully classified 705 objects from total number of 1000 used objects (70.5%). Detection is classified as positive if it overlaps at least 60% with real area. Algorithm did not recognize the remaining 295 objects (29.5%). In addition, the algorithm has recognized 7600 objects as requested objects, though they are not. This means that the algorithm, for each successfully recognized object, averagely has to recognize seven to eight points as positive. These results can be characterized as expected when compared with other researches that have used the AdaBoost algorithm. Baluja and Rowley have used the same algorithm, trying to recognize gender based on the images that

contain people faces [22]. The training gave an accuracy of 80%. This is somewhat higher accuracy, but the images used in testing used a face only, which implicates minor amount of noise. When the training sets contain images with variable background, the accuracy of training process is decreasing. Yuan et al. showed differences in accuracy in recognizing faces when the background changes [23]. When the images contained different amounts of light, the resulting accuracy was 68.4%. In recognition of computer-generated characters with a constant background, AdaBoost achieves an accuracy of 90.5% [24].

When comparing our results to other papers that are also interested in player detection, our work is most similar to Lu et al. [25]. They recognize players from basketball games broadcasted via television stations by applying CRF (Conditional Random Fields algorithm) over DPM (Deformable Part Model). In their study they obtained an accuracy of 73% in player detection but much less number of false positive detection. Lehuger obtained an accuracy of 78.03% by applying AdaBoost algorithm, but algorithm was applied on soccer videos where the background is almost constant and player overlapping is much lower. Those results are shown in Table 1.

The complete output obtained after testing can be presented using ROC curves as given in Fig. 8. This curve shows recognition we can expect when we allow a certain degree of false positives. Figure shows that when we allow seven or more false positive, the level of hits is about 70%. By reducing allowed number of false positive, the percent of true positive is reduced as well. This decrease is approximately linear up to the value 2 for false positives, where the percentage of correct hits is just below 55%. By further reducing the level of false positive, the level of true positive decreases exponentially and if we do not allow false positive, i.e. when this value approaches zero, we can expect only about 25% of successful recognition.

Table 1
Detection results of applying AdaBoost algorithm on different training sets

Researcher(s)	Training set	Result
Baluja and Rowley [22]	Face recognition	80%
Yuan et al. [23]	Face recognition with changing background	68.4%
Hoe et al. [24]	Recognition of computer-generated characters	90.5%
Lehuger et al. [13]	Player recognition in soccer	78.03%
Our work	Player recognition in basketball	70.55

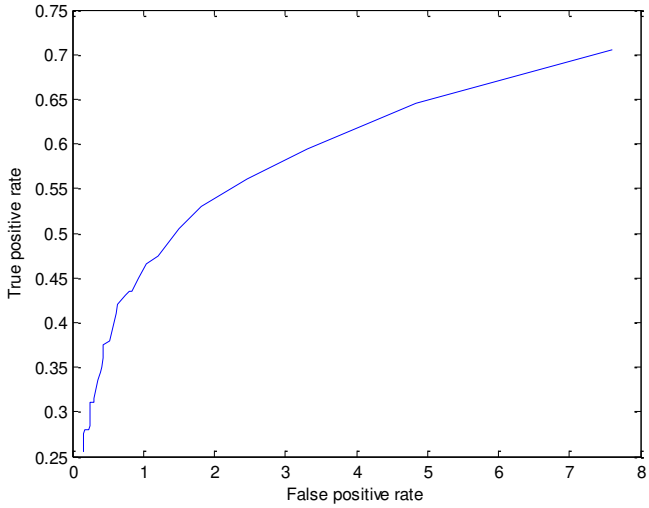


Figure 8
Performance presented by ROC curve

Conclusion

Application of computer vision in analysis of sport events is quite a common practice, especially in recent years. Basketball, as one of the most popular sports, does not deviate from this trend. Unlike some other sports, it is played almost exclusively in halls, which simplifies the process of analysis due to constant lighting. Additional benefits in analysis of basketball games are the facts that all players in any team have jerseys of the same color, and that a large number of players are constantly in the camera view field.

Three approaches can be applied in analyzing basketball games: an analysis using markers that are placed on basketball players, an analysis using multiple synchronized cameras that cover the whole court, and image analysis using single camera. This paper reports the third type of analysis, i.e. analysis of games using video broadcasted by TV stations. This type of analysis brings the greatest amount of assumptions and inaccuracies, because players are often obscured by other players, or are outside the current view field of active camera. Another drawback is the fact that in this type of analysis camera and objects of interest are both moving.

In player detection process, we used Gentle AdaBoost algorithm, which is trained on two sets of examples. First set of images represents entire basketball players' body (head, torso, arms and legs), while another set of images represents basketball players' upper body (head and torso). AdaBoost failed to create a classifier based on images from the first training set. The reason is the large difference in training examples. The appearance of basketball players varies greatly depending on whether they walk, run, play defense, dribble the ball or

shoot on the basket. In addition, all these variations occur when a basketball player is turned in some direction (towards the camera, towards a basket ...). In contrast, another training set, which included only head and torso of basketball players, contained a much smaller degree of variation. This resulted in successfully finished training process by AdaBoost algorithm, which produced a classifier that can be applied in basketball player detection. Nevertheless, the algorithm still has a number of areas in the pictures marked as basketball players (false positive).

This paper presented a degree of applicability of AdaBoost algorithm in the recognition of basketball players, without any prior processing. Obtained results are not applicable in real life situations because of low detection rate and very high rate of false positive detections. AdaBoost algorithm was applied for several reasons. The first is its speed, which is very important in players' recognition in order to get game knowledge as fast as possible. Another reason is very successful application of this algorithm in pedestrian detection on the CCTV footage. Our assumption was that the same algorithm could be successfully applied to players' recognition because that is also movement of people which is captured with the camera. However, the players move in different directions, they can be obscured by other players; a camera that captures them can pan and zoom in order enable viewers better perspective on the current action, and which is most important, their limbs (arms and legs) can be found in almost every position depending on whether they play offense, defense, jump for a ball etc.. When we take a look at pedestrian detection, we have pan, their hands are near the body and leg movements are the same. These are the reasons why AdaBoost did not give applicable results in the recognition of players in basketball games.

In order to improve obtained performances, we could apply background subtraction techniques that would leave only the objects that are likely to represent basketball players, on which the algorithm would then be applied. Further improvement would be achieved by mapping areas of interest, i.e. play field in observed application. This would remove everything that is not on the play field, which would make the search faster and more accurate. Another possible improvement would be training of AdaBoost algorithm for body parts (head, legs, arms, torso), which can then be combined in order to identify players.

AdaBoost primary characteristic is its speed. Some other algorithms achieve better results, if we observe number of identified objects, but their execution is much slower. Identifying players in basketball games is an activity that needs to be done in a short period, because coaches want to have analysis of both teams during the match, in order to make right decisions. Therefore, AdaBoost appears to be a proper solution. One possible solution to achieve better results and fewer false positive requires application of another algorithm that has better accuracy but worse execution time after classification by AdaBoost algorithm. This would not affect the overall performance, because most of the areas in the picture are already rejected by fast AdaBoost algorithm.

Acknowledgment

Research was partially supported by the Ministry of Science and Technological Development of Republic of Serbia, through project no. 171039.

References

[1] Moeslund, T., Hilton, A. and Kruger, V. "A Survey of Advances in Vision-based Human Motion Capture and Analysis", *Computer Vision and Image Understanding*, Vol. 104, No. 2, 2006, pp. 90-126

[2] Dean Oliver, Using Statistics to Make Forecasts, Ph.D., University of North Carolina, 1994

[3] Bill James, "The Bill James Historical Baseball Abstract", Villard 1985

[4] Schumaker, R., Soliman, O. and Chen, H. "Sports Data Mining", *Springer*, 2010

[5] Ballard, C. "Measure of success", *Sports Illustrated*, 2005

[6] Ivankovic, Z., Rackovic, M., Markoski, B., Radosav, D. and Ivkovic, M. "Appliance of Neural Networks in Basketball Scouting", *Acta Polytechnica Hungarica*, Vol. 7, No. 4, 2010, pp. 167-180

[7] Ratgeber, L., Markoski, B., Pecev, P., Lacmanovic, D. and Ivankovic, Z. "Comparative Review of Statistical Parameters for Man's and Women's Basketball Leagues in Serbia", Acta Polytechnica Hungarica, Vol. 10, No. 6, 2013, pp. 151-170

[8] Vasiljevic, P., Markoski, B., Ivankovic, Z., Ivkovic, M., Setrajcic, J. and Milosevic Z. "Basket Supervisor – Collecting Statistical Data in Basketball and Net Casting", *Technics Technologies Education Management*, Vol. 6, No. 1, 2011, pp. 169-178

[9] Lajos Izsó - Péter Tóth: Applying Web-mining Methods for Analysis of Student Behaviour in VLE Courses, Acta Polytechnica Hungarica, 5(4), 2008, pp. 79-92

[10] Lewis, M. "Moneyball: The Art of Winning an Unfair Game", *W. W. Norton & Company*, 2003

[11] Markoski, B., Ivankovic, Z., Pecev, P., Milosevic, Z. and Istrat, V. "Transfer of Basic Statistical Parameters in Basketball by Internet" *Techniques, Informatics and Education TIO 06*, 2011, pp. 792-797

[12] Vasiljevic, P., Ivankovic, Z., Milosevic, Z., Pecev, P. and Markoski, B. "Ajax Web Application for Basketball Statistics", *Information and Communication Technologies for Small and Medium Enterprises*, Arandjelovac, Serbia, 2011

[13] Lehuger, A., S. Duffner, and C. Garcia. "A Robust Method for Automatic Player Detection in Sport Videos." *Compression et Représentation des Signaux Audiovisuels (CORESA'07)*, Montpellier, France, November 2007

[14] Mahmood, Zahid, Tauseef Ali, and Shahid Khattak. "Automatic Player Detection and Recognition in Images using AdaBoost." Applied Sciences and Technology (IBCAST), 2012 9th International Bhurban Conference on IEEE, 2012

[15] Kearns, M. and Valiant, L.G. "Cryptographic Limitations on Learning Boolean Formulae and Finite Automata", *Journal of the Association for Computing Machinery*, Vol. 41, No. 1, 1994, pp. 67-95

[16] Freund, Y. and Schapire, R. E. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, Vol. 55, No. 1, 1997, pp. 119-139

[17] Viola, P. and Jones, M. "Rapid Object Detection using a Boosted Cascade of Simple Features", *Computer Vision and Pattern Recognition*, 2001, pp. 511-518

[18] Kuranov, A., Lienhart, R. and Pisarevsky, V. "An Empirical Analysis of Boosting Algorithms for Rapid Object with Extended Ser of Haar-like Features", *Intel Technical Report MRL-TR-July02-01*, 2002

[19] Viola, P., Jones, M. J. and Snow, D. "Detecting Pedestrians using Patterns of Motion and Appearance", *International Journal of Computer Vision*, Vol. 63, No. 2, 2005, pp. 153-161

[20] Lienhart, R., Kuranov, A. and Pisarevsky, V. "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection", *Pattern Recognition*, 2003, pp. 297-304

[21] Friedman J., Hastie T., Tibshirani R., "Additive Logistic Regression: A Statistical View of Boosting", The Anals of Statistics, Vol. 28, No. 2, 2000, pp. 337-407

[22] Baluja, S. and Rowley, H "Boosting Sex Identification Performance", *International Journal of Computer Vision*, Vol. 71, No. 1, 2007, pp. 111-119

[23] Yuan, Z., Lu, Z. and Pan, H. "LPP-AdaBoost-based Face Detection in Complex Backgrounds", *Fourth International Conference on Computer Sciences and Convergence Information Technology*, 2009, pp. 451-456

[24] Li, L., Hoe, K. E., Yu, X., Dong L. and Chu X. "Human Upper Body Pose Recognition Using Adaboost Template for Natural Human Robot Interaction", *Canadian Conference Computer and Robot Vision*, 2010, pp. 370-377

[25] W. L. Lu, J. A. Ting, K. P. Murphy, J. J. Little "Identifying Players in Broadcast Sports Videos using Conditional Random Fields". *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3249-3256

TP Model-based Robust Stabilization of the 3 Degrees-of-Freedom Aeroelastic Wing Section

Béla Takarics, Péter Baranyi

3D Internet-based Control and Communications Research Laboratory
Institute for Computer Science and Control
Hungarian Academy of Sciences
1111 Kende u. 13-17. Budapest, Hungary
E-mail: takarics.bela@sztaki.mta.hu, baranyi.peter@sztaki.mta.hu

Abstract: Active stabilisation of the 2 and 3 degrees-of-freedom (DoF) aeroelastic wind sections with structural nonlinearities led to various control solutions in the recent years. The paper proposes a control design strategy to stabilise the 3 Dof aeroelastic model. It is assumed that the aeroelastic model has uncertain parameters in the trailing edge dynamics and only one state variable, the pitch angle is measurable, therefore, robust output feedback control solution is derived based on the Tensor Product (TP) type convex representation of the aeroelastic model. The control performance requirements include robust asymptotic stability and constraint on the l_2 norm of the control signal. The control performance requirements are formulated in terms of Linear Matrix Inequalities (LMIs). As the first step of the proposed strategy, the TP type model is obtained by executing TP transformation. As the second step, LMI based control design is performed resulting in controller and observer solution defined with the same polytopic structure as the TP type model. The validation and evaluation of the derived control solutions is based on numerical simulations.

Keywords: aeroelastic wing, robust LMI-based multi-objetive control, TP model transformation, qLPV systems

Nomenclature

The variables used in the paper are defined as below:

- a = non-dimensional distance from the mid-chord to the elastic axis
- b = semi-chord of the wing – m
- c_h = the plunge structural damping coefficients – Nms/rad

- c_{l_α} = aerofoil coefficient of lift about the elastic axis
- c_{l_β} = trailing-edge surface coefficient of lift about the elastic axis
- $c_{m_\alpha,eff.}$ = aerofoil moment coefficient about the elastic axis
- $c_{m_\beta,eff.}$ = trailing-edge moment coefficient about the elastic axis
- c_α = the pitch structural damping coefficient – Nms/rad
- h = plunging displacement – m
- I_α = the mass moment of inertia – kgm^2
- k_h = the plunge structural spring constant
- $k_\alpha(\alpha)$ = non-linear stiffness contribution
- L = aerodynamic force – N
- M = aerodynamic moment – Nm
- m = the mass of the wing – kg
- U = free stream velocity – m/s
- x_α = the non-dimensional distance between elastic axis and the center of mass
- α = pitching displacement – rad
- β = control surface deflection – rad
- ρ = air density – kg/m^3

1 Introduction

Stabilisation of aeroelastic wing section is an actively investigated research area by aerospace and control engineers with a general overview given in [1]. The Nonlinear Aeroelastic Test Apparatus (NATA) model with 3 degrees-of-freedom (DoF) and unsteady aerodynamics was designed in [2, 3] with several control solution approaches found in [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] to name a few. These papers include adaptive control, nonlinear backstepping adaptive control, neural network based approach, optimal infinite-horizon control law, full-state feedforward/feedback control and other control design approaches. A mixed H_∞/H_2 scheduling control system was presented in [16]. An improved 3 DoF NATA model with Linear Quadratic Regulator (LQR) control solution was proposed in [17].

Tensor Product (TP) model transformation based approach was utilised with the application of Linear Matrix Inequalities (LMIs) in several papers. Full state feedback control for the 2 DoF NATA is proposed in [18], which was improved with output

feedback control in [19]. The control performance was further improved by manipulation of the convex hull of the polytopic model in [20]. The 2 DoF NATA was modelled with nonlinear friction in [21], which was utilised for TP model based control design in [22]. A TP model based output feedback control solution is given in [23], which is based on the improved 3 DoF aeroelastic model presented in [17].

The aim of the paper is to propose a control design strategy to robustly stabilise the NATA model given in [17] with uncertain parameters. Besides, the designed control solution has to fulfil criteria of having bounded l_2 norm of the control signal. It is assumed that the only one state, pitch angle and the free-stream velocity are measurable, therefore, output feedback control solution is utilised.

TP type convex polytopic representation of the quasi-Linear Parameter Varying (qLPV) NATA model is obtained by TP model transformation, which is immediately applied for LMI-based control design. TP model transformation is capable of determining various convex representations of the same qLPV model, as well as it can allow the qLPV model to be defined by analytical equations, soft-computing representation or given by numerical data sets. The control design and performance criteria are formulated in terms of LMIs and the control solution results in controller and observer defined by a common polytopic structure of the qLPV model.

The paper shows that defining the uncertainties of qLPV models with various structures has a large influence of the LMI feasibility tests resulting in different control performance solutions.

The paper is structured as follows: the equations of motion and the qLPV representation of the 3 DoF NATA model are given in the following section. The proposed control design methodology is introduced in Section 3 followed by the control design results in Section 4. Section 5 provides numerical simulations with evaluation and the conclusions are provided at the end of the paper.

2 qLPV Model of the 3 DoF NATA Model

The present investigation utilises the NATA model introduced by [16, 17]. The model has three degrees of freedom: plunge h, pitch α and trailing-edge surface deflection β and the equations of motion are the following:

$$\begin{pmatrix} m_h+m_\alpha+m_\beta & m_a x_a b+m_\beta r_\beta+m_\beta x_\beta & m_\beta r_\beta \\ m_a x_a b+m_\beta r_\beta+m_\beta x_\beta & \hat{I}_\alpha+\hat{I}_\beta+m_\beta r_\beta^2+2x_\beta m_\beta r_\beta & \hat{I}_\beta+x_\beta m_\beta r_\beta \\ m_\beta r_\beta & \hat{I}_\beta+x_\beta m_\beta r_\beta \hat{I}_\beta m x_\alpha b & I_\alpha \end{pmatrix} \begin{pmatrix} \ddot{h} \\ \ddot{\alpha} \\ \ddot{\beta} \end{pmatrix} + \quad (1)$$

$$\begin{pmatrix} c_h & 0 & 0 \\ 0 & c_\alpha & 0 \\ 0 & 0 & c_{\beta_{servo}} \end{pmatrix} \begin{pmatrix} \dot{h} \\ \dot{\alpha} \\ \dot{\beta} \end{pmatrix} + \begin{pmatrix} k_h & 0 & 0 \\ 0 & k_\alpha(\alpha) & 0 \\ 0 & 0 & k_{\beta_{servo}} \end{pmatrix} \begin{pmatrix} h \\ \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} -L \\ M \\ k_{\beta_{servo}} \beta_{des} \end{pmatrix}.$$

Based on [17] $k_\alpha(\alpha) = 25.55 - 103.19\alpha + 543.24\alpha^2$. The quasi-steady aerodynamic force and moment is given as:

$$L = \rho U^2 b C_{l_\alpha} \left(\alpha + \frac{\dot{h}}{U} + \left(\frac{1}{2} - a\right) b\frac{\dot{\alpha}}{U}\right) + \rho U^2 b c_{l_\beta}\beta \tag{2}$$
$$M = \rho U^2 b^2 C_{m_{\alpha,eff.}} \left(\alpha + \frac{\dot{h}}{U} + \left(\frac{1}{2} - a\right) b\frac{\dot{\alpha}}{U}\right) + \rho U^2 b C_{m_{\beta,eff.}}\beta.$$

L and M above are valid for the low-velocity regime. The trailing-edge servo-motor dynamics based on [17] can be defined as:

$$\hat{I}_\beta \ddot{\beta} + c_{\beta_{servo}}\dot{\beta} + k_{\beta_{servo}}\beta = k_{\beta_{servo}}\mathbf{u}_\beta. \tag{3}$$

With the combination of equations (1), (3) and (2) one results in:

$$\underbrace{\begin{pmatrix} m_h + m_\alpha + m_\beta & m_a x_a b + m_\beta r_\beta + m_\beta x_\beta & m_\beta r_\beta \\ m_a x_a b + m_\beta r_\beta + m_\beta x_\beta & \hat{I}_\alpha + \hat{I}_\beta + m_\beta r_\beta^2 + 2x_\beta m_\beta r_\beta & \hat{I}_\beta + x_\beta m_\beta r_\beta \hat{I}_\beta m x_\alpha b \\ m_\beta r_\beta & \hat{I}_\beta + x_\beta m_\beta r_\beta \hat{I}_\beta m x_\alpha b & I_\alpha \end{pmatrix}}_{\mathbf{M}_{eom}} \begin{pmatrix} \ddot{h} \\ \ddot{\alpha} \\ \ddot{\beta} \end{pmatrix} + \tag{4}$$
$$+ \underbrace{\begin{pmatrix} c_h + \rho b S C_{l_\alpha} U & \left(\frac{1}{2} - a\right) b\rho b S C_{l_\alpha} U & 0 \\ -\rho b^2 S C_{m_{\alpha,eff}} U & c_\alpha - \left(\frac{1}{2} - a\right) b\rho b^2 S C_{m_{\alpha,eff}} U & 0 \\ 0 & 0 & c_{\beta_{servo}} \end{pmatrix}}_{\mathbf{C}_{eom}} \begin{pmatrix} \dot{h} \\ \dot{\alpha} \\ \dot{\beta} \end{pmatrix} +$$
$$+ \underbrace{\begin{pmatrix} k_h & \rho b S C_{l_\alpha} U^2 & \rho b S C_{l_\beta} U^2 \\ 0 & k_\alpha(\alpha) - \rho b^2 S C_{m_{\alpha,eff}} U^2 & -\rho b^2 S C_{m_{\beta,eff}} U^2 \\ 0 & 0 & k_{\beta_{servo}} \end{pmatrix}}_{\mathbf{K}_{eom}} \begin{pmatrix} h \\ \alpha \\ \beta \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ k_{\beta_{servo}} \end{pmatrix}}_{\mathbf{F}_{eom}} \mathbf{u}.$$

where: $\mathbf{M}_{eom}$ is the mass matrix of the equation of motion, $\mathbf{C}_{eom}$ is the damping matrix of the equation of motion, $\mathbf{K}_{eom}$ is the stiffness matrix of the equation of motion, $\mathbf{F}_{eom}$ is the forcing matrix of the equation of motion.

The equation above was converted into qLPV state space formulation as:

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \\ x_6(t) \end{pmatrix} = \begin{pmatrix} \dot{h} \\ \dot{\alpha} \\ \dot{\beta} \\ h \\ \alpha \\ \beta \end{pmatrix} \quad \text{and} \quad \mathbf{u}(t) = u_\beta,$$

with the state and input matrices given as:

$$\mathbf{A}(\mathbf{p}(t)) = \begin{pmatrix} -\mathbf{M}_{eom}^{-1}\mathbf{C}_{eom}(\mathbf{p}(t)) & -\mathbf{M}_{eom}^{-1}\mathbf{K}_{eom}(\mathbf{p}(t)) \\ -\mathbf{I} & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{M}_{eom}^{-1}\mathbf{F}_{eom} \\ 0 \end{pmatrix}. \quad (5)$$

In case $x_5(t) = \alpha$ is the only measurable state the output and feed-through matrices are the following:

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{D} = 0. \quad (6)$$

The system matrix can be constructed in the following way:

$$\mathbf{S}(\mathbf{p}(t)) = \begin{pmatrix} \mathbf{A}(\mathbf{p}(t)) & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \quad (7)$$

The system parameters are taken from [17], and they are the following:

$m_h = 6.516\ kg$; $m_\alpha = 6.7\ kg$; $m_\beta = 0.537\ kg$; $x_\alpha = 0.21$; $x_\beta = 0.233$; $r_\beta = 0\ m$; $a = -0.673\ m$; $b = 0.1905\ m$; $\hat{I}_\alpha = 0.126\ kgm^2$; $\hat{I}_\beta = 10^{-5}$; $c_h = 27.43\ Nms/rad$; $c_\alpha = 0.215\ Nms/rad$; $c_{\beta_{servo}} = 4.182 \times 10^{-4}\ Nms/rad$; $k_h = 2844$; $k_{\beta_{servo}} = 7.6608 \times 10^{-3}$; $\rho = 1.225\ kg/m^3$; $C_{l_\alpha} = 6.757$; $C_{m_{\alpha,eff}} = -1.17$; $C_{l_\beta} = 3.774$; $C_{m_{\beta,eff}} = -2.1$; $S = 0.5945\ m$.

3 The Proposed Control Design Methodology

3.1 Reconstruction of the TP type polytopic model

TP model transformation with its mathematical background and application in LMI based control design was introduced and elaborated in [24, 25, 26, 27, 23]. The most important definitions corresponding to TP model transformation and TP type polytopic representation are the following:

Definition 1 *(Finite element TP type convex polytopic model - TP model):* $\mathbf{S}(\mathbf{p}(t))$ *in (7) for any parameter is given as the parameter-varying convex combination of LTI system matrices* $\mathbf{S} \in \mathbb{R}^{O \times I}$.

$$\mathbf{S}(\mathbf{p}(t)) = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} .. \sum_{i_N=1}^{I_N} w_{n,i_n}(p_n(t)) \mathbf{S}_{i_1,i_2,..,i_N} = \mathscr{S} \overset{N}{\underset{n=1}{\boxtimes}} \mathbf{w}_n(p_n(t)), \quad (8)$$

where $\mathbf{p}(t) \in \Omega$. *The coefficient tensor* $\mathscr{S} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n \times O \times I}$ *has* $N+2$ *dimensions, it is constructed from the LTI vertex systems* $\mathbf{S}_{i_1,i_2,\ldots,i_N}$ *(8) and the row vector* $\mathbf{w}_n(p_n(t))$

contains one variable and continuous weighting functions $w_{n,i_n}(p_n(t))$, $i_n = 1 \dots I_N$. In order to get convex representation the weighting functions satisfy the following criteria:

$$\forall n, i, p_n(t) : w_{n,i}(p_n(t)) \in [0,1]; \tag{9}$$

$$\forall n, p_n(t) : \sum_{i=1}^{I_n} w_{n,i}(p_n(t)) = 1. \tag{10}$$

Definition 2 *(NO/CNO, NOrmal type TP model): The TP model is NO (normal) type model if its weighting functions are Normal, that is if it satisfies (9), (10), and the largest value of all weighting functions is 1. The convex TP model is CNO (close to normal) if it satisfies (9), (10) and the largest value of all weighting functions is 1 or close to 1.*

TP model transformation is a numerical method allowing the transformation of qLPV models given as (7) to TP type polytopic model defined in (8) enabling the immediate application of LMI based control design. TP model transformation is also capable to find TP type approximations of the original model with arbitrary accuracy. qLPV models can be given as analytical equations based on physical considerations, as the result of soft-computing based identification techniques, or as an outcome of black-box identification. The transformation can be executed within a reasonable amount of time and can replace the analytical conversions by a tractable numerical operation carried out in a routine-like fashion.

3.2 Uncertainty structure

Based on the derivation presented in [28] it is assumed that the uncertain model takes the following structure:

$$\begin{aligned} \dot{x}(t) =& (\mathbf{A}(\mathbf{p}(t)) + \mathbf{D}_a(\mathbf{p}(t))\Delta_a(t)\mathbf{E}_a(\mathbf{p}(t)))x(t) \\ & (\mathbf{B}(\mathbf{p}(t)) + \mathbf{D}_b(\mathbf{p}(t))\Delta_b(t)\mathbf{E}_b(\mathbf{p}(t)))u(t), \end{aligned} \tag{11}$$

where the uncertain blocks $\Delta_a(t)$ and $\Delta_b(t)$ satisfy

$$\|\Delta_a(t)\| \leq \frac{1}{\gamma_a}, \quad \Delta_a(t) = \Delta_a^T(t), \tag{12}$$

$$\|\Delta_b(t)\| \leq \frac{1}{\gamma_b}, \quad \Delta_b(t) = \Delta_b^T(t) \tag{13}$$

and $\mathbf{D}_a(\mathbf{p}(t))$, $\mathbf{E}_a(\mathbf{p}(t))$, $\mathbf{D}_b(\mathbf{p}(t))$ and $\mathbf{E}_b(\mathbf{p}(t))$ are known scaling matrices.

3.3 Control structure

The implementation of full state feedback control is not always straightforward since in many cases the measurement of all states can lead to high sensor cost or measurement difficulties and in some cases the states do not correspond to physical values. In the present case it is assumed that only the pitch angle α of the NATA system is measured, therefore output feedback control structure is utilised. The observer has to be designed in such a way that it satisfies $\mathbf{x}(t) - \hat{\mathbf{x}}(t) \to 0 \quad \text{as} \quad t \to \infty$, where $\hat{\mathbf{x}}(t)$ denotes the state-vector estimated by the observer. Since parameter vector $\mathbf{p}(t)$ does not contain values from the estimated state-vector $\hat{\mathbf{x}}(t)$, the control design strategy presented in [29, 28] was utilised:

$$\begin{aligned} \dot{\hat{\mathbf{x}}}(t) &= \mathbf{A}(\mathbf{p}(t))\hat{\mathbf{x}}(t) + \mathbf{B}(\mathbf{p}(t))\mathbf{u}(t) + \mathbf{K}(\mathbf{p}(t))(\mathbf{y}(t) - \hat{\mathbf{y}}(t)) \\ \hat{\mathbf{y}}(t) &= \mathbf{C}(\mathbf{p}(t))\hat{\mathbf{x}}(t), \end{aligned}$$

where $\mathbf{u}(t) = -\mathbf{F}(\mathbf{p}(t))\mathbf{x}(t)$.

The current investigation applies a control design strategy that yields a controller and an observer, which have share the same polytopic structure of the model itself as:

$$\begin{aligned} \dot{\hat{\mathbf{x}}}(t) &= \mathscr{A} \boxtimes_{n=1}^{N} \mathbf{w}_n(p_n(t))\hat{\mathbf{x}}(t) + \mathscr{B} \boxtimes_{n=1}^{N} \mathbf{w}_n(p_n(t))\mathbf{u}(t) + \mathscr{K} \boxtimes_{n=1}^{N} \mathbf{w}_n(p_n(t))(\mathbf{y}(t) - \hat{\mathbf{y}}(t)) \\ \hat{\mathbf{y}}(t) &= \mathscr{C} \boxtimes_{n=1}^{N} \mathbf{w}_n(p_n(t))\hat{\mathbf{x}}(t) \\ \mathbf{u}(t) &= -\left(\mathscr{F} \boxtimes_{n=1}^{N} \mathbf{w}_n(p_n(t)) \right) \mathbf{x}(t). \end{aligned} \tag{14}$$

The control design aims in determining gains $\mathscr{F}$ and $\mathscr{K}$ that lead to stable output-feedback control structure. The LTI feedback gains $\mathbf{F}_{i_1,i_2,\ldots,i_N}$ and LTI observer gains $\mathbf{K}_{i_1,i_2,\ldots,i_N}$ are stored in tensor $\mathscr{F}$ and $\mathscr{K}$, which are called vertex feedback and observer gains.

3.4 Control performance specifications formulated in terms of LMIs

A large number LMIs guaranteeing various control performance specification has been developed for polytopic systems, which can be readily applied to design vertex controller and observer gains. The control performance objectives of the present investigation are the following:

- Asymptotically stable controller and observer;
- Robust stability of the controller for parameter uncertainties.
- Constrain on the control value.

LMI theorems derived in [28] are selected for the control design.

Theorem 1 *(Globally and asymptotically stable controller for uncertain qLPV systems) A controller stabilising the uncertain qLPV system (11) can be obtained by solving the following LMIs for* $\mathbf{P} > \mathbf{0}$ *and* $\mathbf{M}_r$ $(r = 1, \ldots, R)$

$$\mathbf{S}_{rr} < \mathbf{0},$$

$$\mathbf{T}_{rs} < \mathbf{0},$$

where

$$\mathbf{S}_{rr} = \begin{pmatrix} \left(\mathbf{P}\mathbf{A}_r^T + \mathbf{A}_r\mathbf{P} - \mathbf{B}_r\mathbf{M}_r - \mathbf{M}_r^T\mathbf{B}_r^T\right) & \mathbf{D}_{ar} & \mathbf{D}_{br} & \mathbf{P}\mathbf{E}_{ar}^T & -\mathbf{M}_r^T\mathbf{E}_{br}^T \\ \mathbf{D}_{ar}^T & -\mathbf{I} & 0 & 0 & 0 \\ \mathbf{D}_{br}^T & 0 & -\mathbf{I} & 0 & 0 \\ \mathbf{E}_{ar}\mathbf{P} & 0 & 0 & -\gamma_a^2\mathbf{I} & 0 \\ -\mathbf{E}_{br}\mathbf{M}_r & 0 & 0 & 0 & -\gamma_b^2\mathbf{I} \end{pmatrix},$$

and

$$\mathbf{T}_{rs} = \begin{pmatrix} \begin{pmatrix} \mathbf{P}\mathbf{A}_r^T \\ +\mathbf{A}_r\mathbf{P} \\ -\mathbf{B}_r\mathbf{M}_s \\ -\mathbf{M}_s^T\mathbf{B}_r^T \\ +\mathbf{P}\mathbf{A}_s^T \\ +\mathbf{A}_s\mathbf{P} \\ -\mathbf{B}_s\mathbf{M}_r \\ -\mathbf{M}_r^T\mathbf{B}_s^T \end{pmatrix} & \mathbf{D}_{ar} & \mathbf{D}_{br} & \mathbf{D}_{as} & \mathbf{D}_{bs} & \mathbf{P}\mathbf{E}_{ar}^T & -\mathbf{M}_s^T\mathbf{E}_{br}^T & \mathbf{P}\mathbf{E}_{as}^T & -\mathbf{M}_r^T\mathbf{E}_{bs}^T \\ \mathbf{D}_{ar}^T & -\mathbf{I} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{D}_{br}^T & 0 & -\mathbf{I} & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{D}_{as}^T & 0 & 0 & -\mathbf{I} & 0 & 0 & 0 & 0 & 0 \\ \mathbf{D}_{bs}^T & 0 & 0 & 0 & -\mathbf{I} & 0 & 0 & 0 & 0 \\ \mathbf{E}_{ar}\mathbf{P} & 0 & 0 & 0 & 0 & -\gamma_a^2\mathbf{I} & 0 & 0 & 0 \\ -\mathbf{E}_{br}\mathbf{M}_r & 0 & 0 & 0 & 0 & 0 & -\gamma_b^2\mathbf{I} & 0 & 0 \\ \mathbf{E}_{as}\mathbf{P} & 0 & 0 & 0 & 0 & 0 & 0 & -\gamma_a^2\mathbf{I} & 0 \\ -\mathbf{E}_{bs}\mathbf{A}_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\gamma_b^2\mathbf{I} \end{pmatrix}$$

for $r < s \leq R$, except the pairs (r,s) such that $\forall \mathbf{p}(t) : w_r(\mathbf{p}(t))w_s(\mathbf{p}(t)) = 0$ and where $\mathbf{M}_r = \mathbf{F}_r\mathbf{P}$.

The feedback gains can be obtained from the solution of the above LMIs as $\mathbf{F}_r = \mathbf{M}_r\mathbf{P}^{-1}$.

Theorem 2 *(Globally and asymptotically stable controller with constraint on the control value) The simultaneous solution of the LMIs of Theorem 1 and Theorem 2 in the form of:*

$$\phi^2\mathbf{I} \leq \mathbf{P}$$
$$\begin{pmatrix} \mathbf{P} & \mathbf{M}_r{}^T \\ \mathbf{M}_r & \mu^2\mathbf{I} \end{pmatrix} \geq 0$$

yields an asymptotically stable controller, where $\|\mathbf{u}(t)\|_2 \leq \mu$ is enforced at all time and $\|\mathbf{x}(0)\|_2 \leq \phi$.

Theorem 3 *(Globally and asymptotically stable observer) Assume the polytopic model (8) and a control structure as given by (14). An asymptotically stable observer can be obtained by solving the following LMIs for $\mathbf{P} > \mathbf{0}$ and $\mathbf{N}_r$ $(r = 1,\ldots,R)$:*

$$\begin{aligned} \mathbf{A}_r^T\mathbf{P} - \mathbf{C}_r^T\mathbf{N}_r^T + \mathbf{P}\mathbf{A}_r - \mathbf{N}_r\mathbf{C}_r &< \mathbf{0}, \\ \mathbf{A}_r^T\mathbf{P} - \mathbf{C}_s^T\mathbf{N}_r^T + \mathbf{P}\mathbf{A}_r - \mathbf{N}_r\mathbf{C}_s + \mathbf{A}_s^T\mathbf{P} - \mathbf{C}_r^T\mathbf{N}_2^T + \mathbf{P}\mathbf{A}_s - \mathbf{N}_s\mathbf{C}_r &< \mathbf{0} \end{aligned}$$

for $r < s \leq R$, except the pairs (r,s) such that $\forall \mathbf{p}(t) : w_r(\mathbf{p}(t))w_s(\mathbf{p}(t)) = 0$, and where $\mathbf{N}_r = \mathbf{P}\mathbf{K}_r$. The observer gains can derived from the solution of the above LMIs as $\mathbf{K}_r = \mathbf{P}^{-1}\mathbf{N}_r$.

4 Control Design Results

4.1 TP model transformation of the NATA model

The first step of the control design is to obtain a polytopic form of the NATA model. This step was achieved by the execution of TP model transformation on the state matrix of the NATA model given by (5). Prior executing TP model transformation the transformation space Ω and the discretization grid M has to be defined. Ω was defined in the interval $U \in [8,20](m/s)$ and $\alpha \in [-0.3,0.3](rad)$ and the discretization grid is defined as $M_1 \times M_2$, where $M_1 = 137$ and $M_2 = 137$. The HOSVD-based canonical form for the discretized tensor $\mathscr{S}^D \in \mathbb{R}^{M_1 \times M_2 \times 7 \times 7}$ results in rank 2 in the first dimension and rank 3 in the second dimension. The exact CNO type convex representation of the NATA model can be given by 6 vertex LTI systems, the same number as in the case of the HOSVD-based canonical form. The weighting functions

$w_{1,i}(U)$, $i = 1..2$, and $w_{2,j}(\alpha)$, $j = 1..3$ of the HOSVD-based canonical form and the CNO type convex form are depicted in Figure 1.

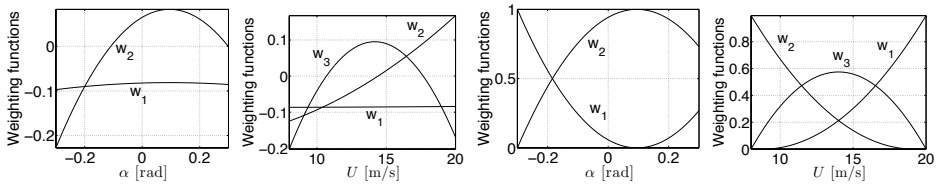


Figure 1
HOSVD-based canonical (left) and CNO type (right) weighting functions of the dimensions α and U.

4.2 LMI-based output feedback controller design

4.2.1 Defining the uncertainty in the trailing-edge servo-motor dynamics

The trailing-edge servo-motor was investigated in [17] resulting in dynamics as given in (3) with parameters $k_{\beta_{servo}}$ and $c_{\beta_{servo}}$. However, it can be assumed that the values of these parameters have some uncertainty, therefore the aim of this section is to define the uncertain structure of the trailing-edge servo-motor dynamics based on (11) in order to design a control system, which can asymptotically stabilize the uncertain qLPV system.

Parameter $k_{\beta_{servo}}$ appears in elements $A_{16}(\mathbf{p}(t))$, $A_{26}(\mathbf{p}(t))$ and $A_{36}(\mathbf{p}(t))$ of state matrix $\mathbf{A}(\mathbf{p}(t))$ and in elements B_{11}, B_{21} and B_{31} of input matrix $\mathbf{B}$ while parameter $c_{\beta_{servo}}$ appears in elements A_{13}, A_{23} and A_{33} of state matrix $\mathbf{A}(\mathbf{p}(t))$ based on which the uncertain blocks $\Delta_a(t)$ and $\Delta_b(t)$ can be defined as:

$$\Delta_a(t) = \begin{pmatrix} \Delta_{k_{\beta_{servo}}}(t) & 0 \\ 0 & \Delta_{c_{\beta_{servo}}}(t) \end{pmatrix} \tag{15}$$

and

$$\Delta_b(t) = \left(\Delta_{k_{\beta_{servo}}}(t) \right), \tag{16}$$

where functions $\Delta_{k_{\beta_{servo}}}(t)$ and $\Delta_{c_{\beta_{servo}}}(t)$ are bounded functions representing the discrepancy between the actual and nominal values of parameters $k_{\beta_{servo}}$ and $c_{\beta_{servo}}$ respectively.

In order to match the uncertain parameters with the corresponding elements of the system matrix $\mathbf{S}(\mathbf{p}(t))$ scaling matrices $\mathbf{D}_a(\mathbf{p}(t))$, $\mathbf{E}_a(\mathbf{p}(t))$, $\mathbf{D}_b(\mathbf{p}(t))$ and $\mathbf{E}_b(\mathbf{p}(t))$ have to be defined.

Proposition 1 *There are two basic possibilities in constructing the scaling matrices which results in the same overall uncertain structure of (11), however, the different structures can* ***highly*** *influence the LMI feasibility tests (see results in the following section).*

Uncertainty structure 1

$$\mathbf{D}_a = \begin{pmatrix} -\mathbf{M}_{eom_{13}}^{-1} c_{\beta_{servo}} & -\mathbf{M}_{eom_{13}}^{-1} k_{\beta_{servo}} \\ -\mathbf{M}_{eom_{23}}^{-1} c_{\beta_{servo}} & -\mathbf{M}_{eom_{23}}^{-1} k_{\beta_{servo}} \\ -\mathbf{M}_{eom_{33}}^{-1} c_{\beta_{servo}} & -\mathbf{M}_{eom_{33}}^{-1} k_{\beta_{servo}} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{E}_a = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{17}$$

and

$$\mathbf{D}_b^T = \begin{pmatrix} \mathbf{M}_{eom_{13}}^{-1} k_{\beta_{servo}} & \mathbf{M}_{eom_{23}}^{-1} k_{\beta_{servo}} & \mathbf{M}_{eom_{33}}^{-1} k_{\beta_{servo}} & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_b = 1. \tag{18}$$

Uncertainty structure 2

$$\mathbf{D}_a = \begin{pmatrix} -\mathbf{M}_{eom_{13}}^{-1} & -\mathbf{M}_{eom_{13}}^{-1} \\ -\mathbf{M}_{eom_{23}}^{-1} & -\mathbf{M}_{eom_{23}}^{-1} \\ -\mathbf{M}_{eom_{33}}^{-1} & -\mathbf{M}_{eom_{33}}^{-1} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{E}_a = \begin{pmatrix} 0 & 0 & c_{\beta_{servo}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & k_{\beta_{servo}} \end{pmatrix} \tag{19}$$

and

$$\mathbf{D}_b^T = \begin{pmatrix} \mathbf{M}_{eom_{13}}^{-1} & \mathbf{M}_{eom_{23}}^{-1} & \mathbf{M}_{eom_{33}}^{-1} & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_b = k_{\beta_{servo}}. \tag{20}$$

4.2.2 Control design results

The CNO type convex polytopic representation on the NATA model can be immediately applied for LMI-based control design. The following controllers and observers were designed based on various control performance specifications for both **Uncertainty structure 1** and **Uncertainty structure 2**.

Proposition 2 *Defining the maximal allowable difference between the nominal and actual values of parameters $k_{\beta_{servo}}$ and $c_{\beta_{servo}}$ for a given control solution can be done in the following way:*

Recall that the uncertain blocks $\Delta_a(t)$ and $\Delta_b(t)$ were defined as

$$\Delta_a(t) = \begin{pmatrix} \Delta_{k_{\beta_{servo}}}(t) & 0 \\ 0 & \Delta_{c_{\beta_{servo}}}(t) \end{pmatrix}, \quad \Delta_b(t) = \left(\Delta_{k_{\beta_{servo}}}(t)\right)$$

and they satisfy

$$\|\Delta_a(t)\| \leq \frac{1}{\gamma_a}, \quad \|\Delta_b(t)\| \leq \frac{1}{\gamma_b}.$$

Since $\Delta_a(t)$ is a diagonal matrix it has a norm that equals the absolute value of its largest element; $\Delta_b(t)$ is a scalar value having a norm equal to its absolute value. Matrix $\Delta_a(t)$ contains $\Delta_b(t)$, therefore γ_a can be set to equal γ_b. The maximal discrepancy of the two parameters are given as

$$\Delta^{max}_{k_{\beta_{servo}}} = \frac{1}{\gamma_a} \geq |\Delta_{k_{\beta_{servo}}}(t)|;$$
$$\Delta^{max}_{c_{\beta_{servo}}} = \frac{1}{\gamma_a} \geq |\Delta_{c_{\beta_{servo}}}(t)|,$$

where superscript "max" stands for the maximal allowable difference.

In the following the differentiation for Control solution $n.1$ and $n.2$ stands for Uncertainty structure 1 and 2 respectively.

Control solution 1.1 and 1.2

Control solutions 1.1 and 1.2 were designed with the aim to find the minimal value for $\gamma_a = \gamma_b$ allowing the maximal uncertainties in parameters $k_{\beta_{servo}}$ and $c_{\beta_{servo}}$ the controller can asymptotically stabilize. The feedback gains were designed by applying the LMIs from Theorem 1 and the observer gains by applying LMIs from Theorem 3.

Control solution 1.1 achieved $\gamma_{a_{min}} = \gamma_{b_{min}} = 1.44$ while control solution 1.2 resulted in $\gamma_{a_{min}} = \gamma_{b_{min}} = 1.77$.

Control solution 2.1 and 2.2

The aim in designing **Control solutions 2.1, 2.2, 3.1, 3.2, 4.1 and 4.2** was to find a trade-off between the maximal allowable parameter discrepancy and keeping the

control signal as low as possible. The feedback gains were derived by applying the LMIs from Theorems 1 and 2 with the initial state condition bound set as $\phi = 0.002$ for each design. The observer gains for each solution were derived by applying LMIs from Theorem 3.

In case of **Control solution 2.1 and 2.2** the maximal allowable discrepancy between the nominal and actual parameter was set to 50% resulting in $\gamma_{a_{min}} = \gamma_{b_{min}} = 2$. The minimal values for the control constrain μ that leads to feasible controller was $\mu_{min} = 44$ for **Control solution 2.1** and $\mu_{min} = 13286$ for **Control solution 2.2**.

Control solution 3.1 and 3.2

$\gamma_{a_{min}} = \gamma_{b_{min}} = 5$ was set for **Control solution 3.1 and 3.2** yielding feasible solutions with $\mu_{min} = 26$ for **Control solution 3.1** and $\mu_{min} = 3672$ for **Control solution 3.2**.

Control solution 4.1 and 4.2

The minimal value of $\mu_{min} = 22$ for **Control solution 4.1** and $\mu_{min} = 3595$ for **Control solution 4.2** was achieved by setting $\gamma_{a_{min}} = \gamma_{b_{min}} = 10$.

5 Simulation Results and Evaluation

5.1 Simulation

The responses of the control solutions were verified by numerical simulations. The base free stream velocity is chosen to equal $U = 14.1 m/s$ for two reasons; first, it belongs to the critical free stream velocity range where the NATA model exhibits limit cycle oscillations, second, to be comparable with the results of several previous papers, which used the same speed in their measurements or simulations. The controller was turned off for five seconds at each simulation to let the oscillations develop, but the figures bellow show only that part of the responses where the controller is turned on.

Each controller was tested in two simulation cases:

- **Simulation case 1** represents the response without any perturbations. In order to fully test the allowable uncertainty ranges, functions $\Delta_{k_{\beta_{servo}}}(t)$ and $\Delta_{c_{\beta_{servo}}}(t)$ were set as:
  - $\Delta_{k_{\beta_{servo}}}(t) = \frac{1}{\gamma_{a_{min}}} sin\left(6\pi t + \frac{\pi}{2}\right)$;
  - $\Delta_{c_{\beta_{servo}}}(t) = \frac{1}{\gamma_{a_{min}}} sin(10\pi t)$;

where $\gamma_{a_{min}}$ takes the minimal value corresponding to each control solution.

- **Simulation case 2** represents system that has several perturbations that can occur during physical implementation. Simualtion case 1 was extended with the following perturbations:
  - the computational delay is represented by 1 *ms* constant time delay;
  - the control signal is saturated at $\pm 2[rad/s]$;
  - the sensor noise is represented by normally distributed random noise with 10% variance;
  - the free stream velocity varies as $U(t) = 14.1 + 5sin(2\pi t)$;
  - input disturbance $u_d(t) = \frac{30}{180}\pi$ is added to the control signal 1 second after the controller is turned on;

Figures 2 and 3 show the time response of the closed loop system for Control solution 2.1 and 2.2 for Simulation case 1, Control solution 2.1 and 2.2 for Simulation case 2, Control solution 3.1 and 3.2 for Simulation case 1 and Control solution 3.1 and 3.2 for Simulation case 1 respectively.

5.2 Evaluation

Each control solution can asymptotically stabilise the NATA model, however, it is important to note that the control solutions guarantee stability within the domain Ω. The control performance of each control solution is evaluated based on the maximal allowable parameter uncertainty, maximal control signal and settling time.

- **Control solution 1.1 and 1.2** - maximal allowable parameter uncertainty (70% for solution 1.1 and 56.5% for solution 1.2); very high control signal (u_{max} is in the order of magnitude of 10^5); settling time approximately 1*s* for Simulation case 1. Control solutions 1.1 and 1.2 were not able to stabilize the NATA system as the high feedback gains lost stability due to the time delay in Simulation case 2. *Difficult for practical implementation due to high control signals.*

- **Control solution 2.1 and 2.2** - high allowable parameter uncertainty (50%), high control signal magnitude ($u_{max} = 150$ for Control solution 2.1 and $u_{max} = 250$ for Control solution 2.2); settling time approximately 1*s* for Simulation case 1 and approximately 1.5*s* for Simulation case 2. *High allowable parameter uncertainty for acceptable control signal magnitude.*

- **Control solution 3.1 and 3.2** - acceptable allowable parameter uncertainty (20%), low control signal magnitude ($u_{max} = 35$ for Control solution 3.1 and

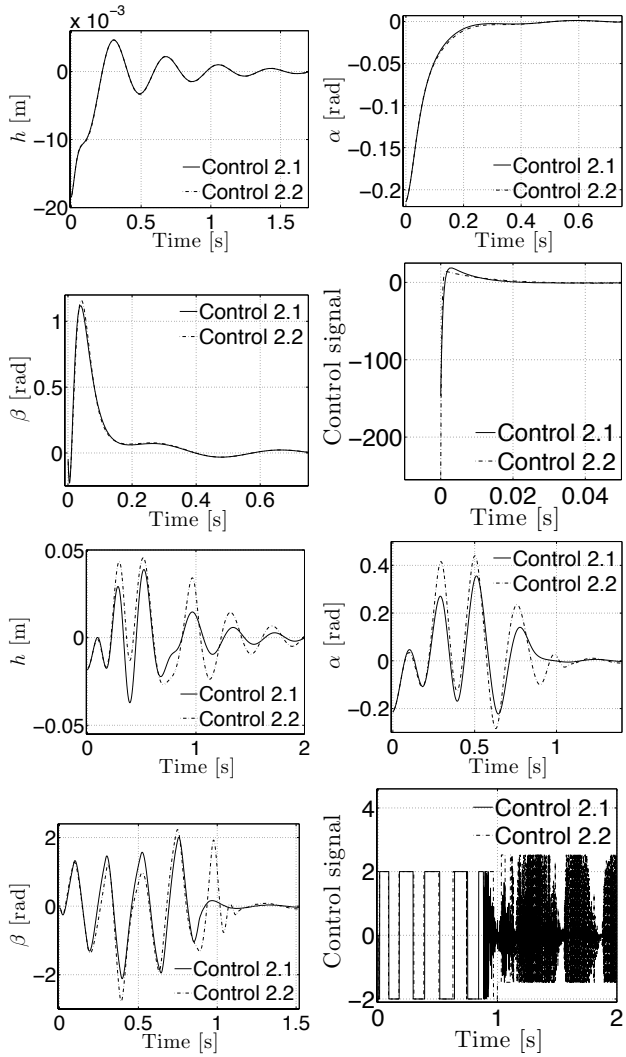


Figure 2
Time response of Control solution 2.1 and 2.2 for Simulation case 1 (first row) and Simulation case 2 (second row).

$u_{max} = 7.8$ for Control solution 3.2); settling time approximately 1s for Simulation case 1 and approximately 1.5s for Simulation case 2. *Low control signal magnitude for acceptable parameter uncertainty.*

- **Control solution 4.1 and 4.2** - minimal allowable parameter uncertainty (10%), smallest control signal ($u_{max} = 15$ for Control solution 4.1 and $u_{max} = 7$ for Control solution 4.2); settling time approximately 1s for Simulation case 1 and

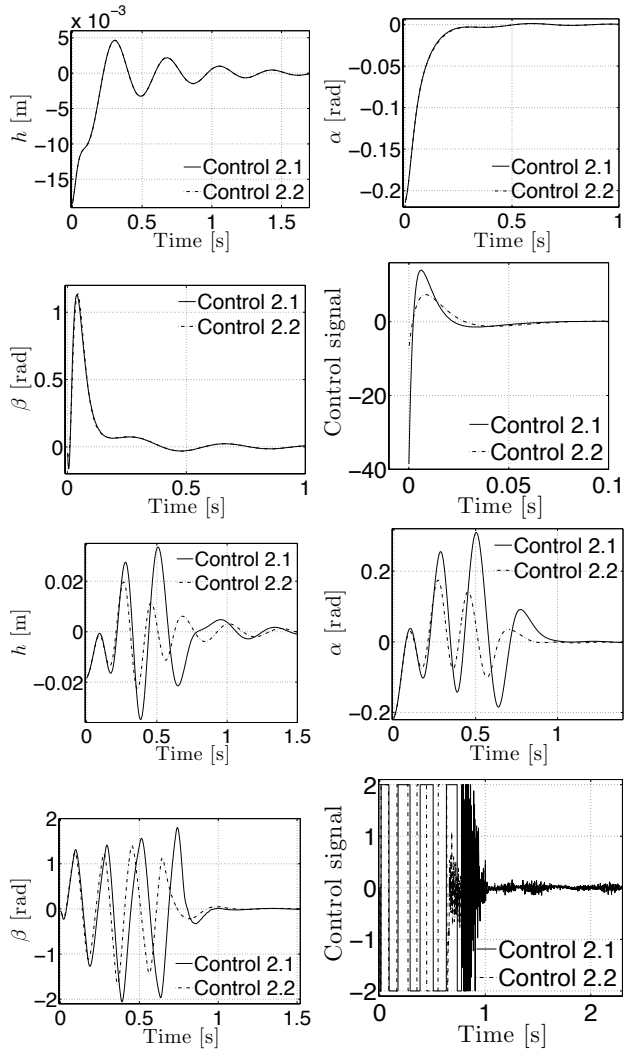


Figure 3
Time response of Control solution 3.1 and 3.2 for Simulation case 1 (1. and 2. row) and Simulation case 2 (3. and 4. row).

approximately 1.5s for Simulation case 2. *Further decrease in the allowable uncertainty does not decrease the control signal significantly.*

Generally, it can be concluded that maximising the allowable parameter uncertainties without a constrain on the control value leads to unacceptably high control signals. On the other hand, it is possible to find a trade-off between the maximal acceptable

uncertainty and the limit of the control signal magnitude. In this term, Control solution 3.2 achieved the best results in the simulations. Control solution 2.1 can be a chosen in case higher uncertainty is a requiremed. Decreasing the allowable uncertainty to very low levels however, does not results in significant decrease in the control signal magnitude.

It can be noted that it is worth to test various uncertainty structures (Uncertainty structure 1 and 2 in the present case), as it highly influences the LMI feasibility results. Uncertainty structure 1 led to better control performance at higher allowable uncertainties, while Uncertainty structure 2 was more favourable at lower parameter uncertainties.

There is no significant difference in the settling time in any control solution.

The LMIs defining the constrain on the control signal lead to large differences between the smallest control signal bounds in case of Uncertainty structure 1 and 2, which imply high conservativity of bound μ, however, within the same uncertainty structure it can indicate the control signal magnitude effectively.

The control performance of the derived solutions can be compared to results found in other publications dealing with the same NATA model. LQR controller was designed in [17] with somewhat longer settling time and lower control signals, however, the model has nonlinearity only in the dimension of U and full state feedback is utilised instead of output feedback. Similar control performance was achieved in [23] and in [18], which was expected as the same control design methodology was utilised. However, [18] designed controller for the 2 DoF NATA model, while there is no robustness involved in the control design of [23]. LQR based output feedback controller is designed in [30] with similar control performance as Simulation case 1 in the present investigation.

6 Conclusions

The proposed control design strategy based on Tensor Product model transformation can be executed systematically in a routine-like fashion and can include various forms control performance specifications formulated in terms of Linear Matrix Inequalities. The paper gives a robust stabilising output feedback control solution for the three degrees-of-freedom Nonlinear Aeroelastic Test Apparatus that can involve parameter uncertainties. Finding an acceptable trade-off between maximal allowable uncertainties and the upper bound of the control signal value was straightforward based on the systematic execution of the numerical control design. It was shown that varying the structure of the same uncertainty of quasi-Linear Parameter Varying systems can highly influence the feasibility tests of Linear Matrix Inequalities and thus the resulting control performance. Both, out of the two possible uncertainty structures available, can lead to good control performance depending the actual specifications, meaning it is worthwhile to investigate both structures for the given task.

Acknowledgements

The research was supported by the Hungarian National Development Agency, (ERC-HU-09-1-2009-0004, MTASZTAK) (OMFB-01677/2009).

The research was part of the Zoltán Magyary Postdoctoral Scholarship.

This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/1-11-1-2012-0001 'National Excellence Program'.

Takarics Béla publikációt megalapozó kutatása a TÁMOP 4.2.4.A/1-11-1-2012-0001 azonosító számú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése országos program című kiemelt projekt keretében zajlott. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

References

[1] V. Mukhopadhyay, "Historical perspective on analysis and control of aeroelastic responses," *Journal of Guidance, Control, and Dynamics*, vol. 26, pp. 673–684, Sep. 2003. [Online]. Available: http://doi.aiaa.org/10.2514/2.5108

[2] J. J. Block, *Active control of an aeroelastic structure*. Texas A&M University, 1996.

[3] J. J. Block and T. W. Strganac, "Applied active control for a nonlinear aeroelastic structure," *Journal of Guidance, Control, and Dynamics*, vol. 21, pp. 838–845, Nov. 1998. [Online]. Available: http://doi.aiaa.org/10.2514/2.4346

[4] T. W. Strganac, J. Ko, and D. E. Thompson, "Identification and control of limit cycle oscillations in aeroelastic systems," *Journal of Guidance, Control, and Dynamics*, vol. 23, pp. 1127–1133, Nov. 2000.

[5] J. Ko, A. J. Kurdila, and T. W. Strganac, "Stability and control of a structurally nonlinear aeroelastic system," *Journal of Guidance, Control, and Dynamics*, vol. 21, pp. 718–725, Sep. 1998. [Online]. Available: http://doi.aiaa.org/10.2514/2.4317

[6] J. Ko, T. W. Strganac, and A. J. Kurdila, "Adaptive feedback linearization for the control of a typical wing section with structural nonlinearity," *Nonlinear Dynamics*, vol. 18, no. 3, pp. 289–301, 1999, 10.1023/A:1008323629064. [Online]. Available: http://dx.doi.org/10.1023/A:1008323629064

[7] G. Platanitis and T. Strganac, "Control of a nonlinear wing section using leading- and Trailing-Edge surfaces," *Journal of Guidance, Control, and Dynamics*, vol. 27, pp. 52–58, Jan. 2004. [Online]. Available: http://doi.aiaa.org/10.2514/1.9284

[8] ——, "Suppression of control reversal using leading- and Trailing-Edge control surfaces," *Journal of Guidance, Control, and Dynamics*, vol. 28, pp. 452–460, May 2005. [Online]. Available: http://doi.aiaa.org/10.2514/1.6692

[9] W. Xing and S. N. Singh, "Adaptive output feedback control of a nonlinear aeroelastic structure," *Journal of Guidance, Control, and Dynamics*, vol. 23, pp. 1109–1116, Nov. 2000. [Online]. Available: http://doi.aiaa.org/10.2514/2.4662

[10] S. N. Singh and L. Wang, "Output feedback form and adaptive stabilization of a nonlinear aeroelastic system," *Journal of Guidance, Control, and Dynamics*, vol. 25, pp. 725–732, Jul. 2002. [Online]. Available: http://doi.aiaa.org/10.2514/2.4939

[11] S. Singh, "State feedback control of an aeroelastic system with structural nonlinearity," *Aerospace Science and Technology*, vol. 7, pp. 23–31, Jan. 2003.

[12] K. K. Reddy, J. Chen, A. Behal, and P. Marzocca, "Multi-Input/Multi-Output adaptive output feedback control design for aeroelastic vibration suppression," *Journal of Guidance, Control, and Dynamics*, vol. 30, pp. 1040–1048, Jul. 2007. [Online]. Available: http://doi.aiaa.org/10.2514/1.27684

[13] K. W. Lee and S. N. Singh, "Multi-Input Noncertainty-Equivalent adaptive control of an aeroelastic system," *Journal of Guidance, Control, and Dynamics*, vol. 33, pp. 1451–1460, Sep. 2010. [Online]. Available: http://doi.aiaa.org/10.2514/1.48302

[14] R. C. Scott and L. E. Pado, "Active control of wind-tunnel model aeroelastic response using neural netwroks," *Journal of Guidance, Control, and Dynamics*, vol. 23, no. 6, pp. 1100–1108, 2000.

[15] Z. Wang, A. Behal, and P. Marzocca, "Model-Free control design for Multi-Input Multi-Output aeroelastic system subject to external disturbance," *Journal of Guidance, Control, and Dynamics*, vol. 34, pp. 446–458, Mar. 2011. [Online]. Available: http://doi.aiaa.org/10.2514/1.51403

[16] Z. Prime, B. Cazzolato, and C. Doolan, "A mixed H2/Hx scheduling control scheme for a two degree-of-freedom aeroelastic system under varying airspeed and gust conditions," in *Proceedings of the AIAA Guidance, Navigation and Control Conference and Exhibit*, Honolulu, Hawaii, 2008, pp. 1–16.

[17] Z. Prime, B. Cazzolato, C. Doolan, and T. Strganac, "Linear-Parameter-Varying control of an improved Three-Degree-of-Freedom aeroelastic model," *Journal of Guidance, Control, and Dynamics*, vol. 33, Mar. 2010. [Online]. Available: http://doi.aiaa.org/10.2514/1.45657

[18] P. Baranyi, "Tensor product Model-Based control of Two-Dimensional aeroelastic system," *Journal of Guidance, Control, and Dynamics*, vol. 29, pp. 391–400, Mar. 2006. [Online]. Available: http://doi.aiaa.org/10.2514/1.9462

[19] ——, "Output feedback control of Two-Dimensional aeroelastic system," *Journal of Guidance, Control, and Dynamics*, vol. 29, pp. 762–767, May 2006. [Online]. Available: http://doi.aiaa.org/10.2514/1.14981

[20] P. Grof, P. Baranyi, and P. Korondi, "Convex hull manipulation based control performance optimisation," *WSEAS Transactions on Systems and Control*, vol. 5, no. 8, pp. 691–700, Aug. 2010, stevens Point, Wisconsin, USA.

[21] J. J. Block and H. Gilliat, "Active control of an aeroelastic structure," in *AIAA Meeting Papers on Disc*. Reno, NV: American Institute of Aeronautics and Astronautics, Inc., Jan. 1997, pp. 1–11.

[22] B. Takarics, P. Grof, P. Baranyi, and P. Korondi, "Friction compensation of an aeroelastic wing - a TP model transformation based approach." IEEE, Sep. 2010, pp. 527–533.

[23] B. Takarics and P. Baranyi, "Tensor-product-model-based control of a three degrees-of-freedom aeroelastic model," *Journal of Guidance, Control, and Dynamics*, vol. 36, no. 5, pp. 1527–1533, Sep. 2013.

[24] P. Baranyi, L. Szeidl, P. Várlaki, and Y. Yam, "Definition of the HOSVD-based canonical form of polytopic dynamic models," in *3rd International Conference on Mechatronics (ICM 2006)*, Budapest, Hungary, July 3-5 2006, pp. 660–665.

[25] P. Baranyi, "TP model transformation as a way to LMI based controller design," *IEEE Transaction on Industrial Electronics*, vol. 51, no. 2, pp. 387–400, April 2004.

[26] P. Baranyi, L. Szeidl, P. Várlaki, and Y. Yam, "Numerical reconstruction of the HOSVD-based canonical form of polytopic dynamic models," in *10th International Conference on Intelligent Engineering Systems*, London, UK, June 26-28 2006, pp. 196–201.

[27] P. Galambos and P. Baranyi, "Trepresenting the model of impedance controlled robot interaction with feedback delay in polytopic lpv form: Tp model transformation based approach," *Acta Polytechnica Hungarica*, vol. 10, no. 1, pp. 139–157, Jan. 2013.

[28] K. Tanaka and H. O. Wang, *Fuzzy Control Systems Design and Analysis: A Linear Matrix Inequality Approach*. John Wiley & Sons, Inc., 2001.

[29] C. W. Scherer and S. Weiland, *Linear Matrix Inequalities in Control*. DISC course lecture notes, 2000, http://w3.ele.tue.nl/fileadmin/ele/MBS/CS/Files/Courses/DISClmi/lmis1.pdf, retrieved on 05.02.2012.

[30] N. Bhoir, "Output feedback nonlinear control of an aeroelastic system with unsteady aerodynamics," *Aerospace Science and Technology*, vol. 8, pp. 195–205, Apr. 2004.

Modelling the Composite Competitiveness Index of the Knowledge-based Society

Andrea Katić[1], Tibor Kiš[2], Ilija Ćosić[3], Simonida Vukadinović[4], Tinde Dobrodolac Šeregelj[5]

[1]University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad/ EDUCONS University, Vojvode Putnika 87, 21208 Sremska Kamenica, Serbia, e-mail: andrea.katic@educons.edu.rs

[2]University of Novi Sad, Faculty of Economics in Subotica, Segedinski put 9-11, 24000 Subotica, Serbia, e-mail: tkis@ef.uns.ac.rs

[3]University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, e-mail: ilijac@uns.ac.rs

[4]EDUCONS University, Vojvode Putnika 87, 21208 Sremska Kamenica, Serbia, e-mail: simonida.vukadinovic@educons.edu.rs

[5]University of Novi Sad, Faculty of Economics in Subotica, Segedinski put 9-11, 24000 Subotica, Serbia, e-mail: tinde@ef.uns.ac.rs

Abstract: In the era of globalization of world markets, movement of the European Union toward the knowledge-based society, and Serbia's EU accession process, competition has become one of the most important characteristics of successful development and achievement of these objectives. The existing competitiveness indices are discussed in this study, and it has been found that they do not reflect the position of transition countries and Serbia appropriately. A great impact of the quality indicators on Serbian ranking was determined based on the analysis of the existing studies. A new, original index is therefore proposed – the Competitiveness Index of the Knowledge-Based Society. The index is applied to 18 territorial units, including the EU and the Western Balkan countries, Serbia, and Vojvodina as a European region. The new model set out in this article provides a more realistic and objective picture of the state of Serbia and the Western Balkans as regards the competitiveness of knowledge. In other words, the new competitiveness model of knowledge-based society provides a better monitoring of the development of the Republic of Serbia and the Western Balkans on their way towards development of knowledge society. The results are analyzed and discussed.

Keywords: competitiveness indices; knowledge society; knowledge as a criterion for competitiveness; Serbia

1 Introduction

The European Union has set the development towards the knowledge-based society as its central objective to achieve competitive advantage in global competition. The Western Balkan countries, including Serbia, have set the EU accession as their strategic goal, and in this regard they have adjusted their objectives to global EU strategy Europe 2020, which aspires to knowledge society.

A great number of indicators for monitoring the model of competitiveness, i.e. competitiveness indices, have appeared to monitor the degree of achievement the objectives of these strategies, the level of market development, and the level of competitiveness of national and regional economies at the end of the 20th Century. However, while strategies define clear objectives, the basic problem is the selection of appropriate indicators, which should show the degree of achievement of the set strategies, as well as monitoring and controlling the set objectives.

The fundamental objective of this article is to create a new model for evaluation of knowledge competitiveness of nations. It is particularly important to identify those indicators in the model functioning that contribute, as well as those that reduce the quality of monitoring in the area of knowledge. The position of selected countries has been analyzed, and it represents one of the main results of our research. The study also includes regional level, involving Vojvodina in the analyses.

According to the analysis of the existing models for monitoring national competitiveness, the basic hypothesis of work is set: A new model of competitiveness, based on knowledge and predominantly quantitatively expressed parameters, provides a more realistic evaluation of the competitiveness of a country. The sub-hypothesis of the research is set as follows: We can identify certain groups of parameters (subindices) where the differences between the best and the worst ranked country are small, as well as those groups of parameters where these differences are significant. Testing each of the assumptions requires examination of the model structure, relations of individual parts and their functioning. Mathematical and statistical methods were used for quantitative relations within the model structure and in relations of model with other defined phenomena.

2 Conceptual Background

It could be argued that economic development has always been based on knowledge. However, the scope and the importance of knowledge for economic processes have fundamentally changed over the past few years [20]. What has changed in comparison to the old, traditional economics is that productivity growth, driven by technological and organizational innovation, has become a key source of economic growth. With concern about the environment, the restrictions

on the use of natural resources are becoming more obvious. The source that enables overcoming that is the knowledge and the creation and linking of knowledge that supports the development of new commercial products and services [16].

"Knowledge society" as a term was first used by Drucker in 1969 [8], and its current meaning was accepted in the last decade of the 20th Century. Drucker describes the knowledge society as a society of mobility and considers it to be the most competitive society in the history of mankind [9]. The OECD (1996) defines a knowledge-based economy as one in which the production, distribution and use of knowledge are the main drivers of growth, wealth creation, and employment for all industries [22].

The criteria of knowledge society are a high percentage of highly educated population, large government investments in education, science and research, encouragement of lifelong learning, high-quality and accessible information and communication infrastructure and services, propulsive and competitive economy, sustainable technological development, wide availability of information and easy access to them. Comparison of development of different national economies based on knowledge is made by evaluations of international organizations, state institutions, statistical departments and other institutions in collaboration with scientists [9]. Different authors and schools that have defined a number of indices related to competitiveness, which merged groups of various sets of criteria, emerged at the end of the 20th and beginning of the 21st Century. Organizations such as the World Economic Forum (WEF) and the Institute for Management Development publish rankings of national competitiveness among countries every year. These rankings are used as benchmarks for national policy makers and other interested parties in the evaluation of their countries in various fields [21]. There are several studies examining competitiveness indices including innovations, knowledge and technological progress [23, 22, 15, 20].

The number of indices that describe the competitiveness of knowledge today is considerably higher. We examined 23 composite indices, which define competitiveness of an economy, including the parameters of knowledge. It was noticed that they can be classified into the following four categories [17, 18]:

- competitiveness indices
- knowledge competitiveness indices
- innovation competitiveness indices
- information and communication technologies competitiveness indices

In order to closely follow the progress of European countries in transition, i.e. degree of fulfilment of the objectives set in the strategy of development until 2020 and on, it is necessary to set up a new, revised model, which will better indicate specific problems, the so-called bottlenecks in the development towards achieving the status of knowledge society. The key parameters of this new model are knowledge, innovation, R&D, education, the use of IT technology, the development of knowledge-intensive jobs and sustainable development.

3 Methodology

This research used calculating the average values, processing of time series, regression and correlation analysis. Results of the research are presented in analytical tables and charts. Secondary data, mainly official statistical reports and publications of prominent institutions, were also used as inputs in this phase of the study. The model used in this article is based on the manual establishment of thematic indicators. Constructing the composite indicator includes several stages. This model consists of nine steps:

1. Development of the conceptual framework;
2. Selection of data (indicators) and the sample size;
3. Transformation of irreversible data and replacing missing data;
4. Classification of indicators by thematic groups;
5. Standardization of individual indicators and assigning weight coefficients;
6. Aggregation and formation of thematic indicators;
7. Weighting of thematic indicators;
8. Aggregation and formation of competitiveness index;
9. Testing the competitiveness index.

It is important to mention that this process should not necessarily be seen as a sequential, and in many cases these steps are simultaneously taken [4].

4 Data and Results

4.1 Development of the Conceptual Framework

The framework for establishing a composite indicator as a summary of the phenomenon should provide a clear definition of what is to be measured and demonstrate which individual indicators should be sought and weighted [20]. The model presented and used in this article describes the state of an economy according to parameters of knowledge society in which general economic preconditions, the use of information and communication technologies, education, research and development, innovation and sustainable developmentare included. In [10] it is indicated that successful knowledge economies include factors such as long-term investments into education, sufficient innovation capacity, adequate information infrastructure and favourable economic surroundings.

4.2 Selection of Data (Indicators) and the Sample Size

In the ideal case, variables should be selected on the basis of their analytical validity, measurability and relevance to the emergence of the indexation, rather than solely on the basis of data availability. In practice, the lack of necessary data is a frequent situation due to the fact that certain phenomena cannot be measured, or because no one has tried to measure them. Proxy measures can be used in this case, as a solution that should be adopted when there are problems of interstate comparability [16] However, using proxy measures means measuring something that is related to the phenomenon, but it is not the same as the phenomenon which is analysed. The selection of variables requires a balance between simplifying and complexity [4]. Scaling of variables with an adequate measure of size (e.g., population, income, land area, etc) is necessary in order to have an objective comparison between countries of different sizes.

4.2.1 Selection of Indicators and Data Sources

The model presented in the article contains 65 indicators which may be considered to represent the standards of knowledge-based society. It includes information about the state of the economy, the use of information and communication technologies, education, R&D, innovativeness and sustainable development. Three out of 65 selected indicators represent mixed indicators (already measured composite indicators), the remaining are quantitative. By analyzing the existing models, it was found that the choice of larger number of parameters with the use of quantitative indicators makes composite index more objective. Qualitative parameters are subject to manipulation because they depend on assessors' subjective opinions. For this reason they are not included in the model developed in the article. However, a great deal of potentially very impactful parameters are not used because of this limitation.

1. **Political stability and absence of violence** is a composite indicator measuring probability that the government will be destabilized or overthrown, in an unconstitutional or violent way, including politically motivated violence and terrorism. Source: World Bank, The Worldwide Governance Indicators, (http://info.worldbank.org/governance/wgi/index.asp, 2012);
2. **GDP per capita** is gross domestic product divided by the number of inhabitants of middle age, in $ 1000. Source: World Bank, (http://data. worldbank.org/indicator/NY.BDP.PCAP.CD?display=default, 2012);
3. **Time required to start a business** is the average duration in days required to complete all the procedures with a minimum subsequent additional obligations and payments. Source: World Bank, (www.doingbusiness.org, 2012);
4. **Time needed for export** is the period in days required to complete all the necessary procedures for export of product. Source: World Bank, (www.doingbusiness.org, 2012);

5. **Time needed for import** is the period in days required to complete all the necessary procedures for import of product. Source: World Bank, (www.doingbusiness.org, 2012);
6. **Percentage of households that use the internet**. Source: Eurostat survey on ICT use by households or individuals, (http://epp.eurostat.ec.europa.eu/cache /ITY_SDDS/FR/isoc_bde15c_esms.htm, 2012), ITU, UN specialized agency for ICT, (www.itu.int/ITU-D/ict/statistics/, 2012);
7. **Percentage of population that use the internet**;
8. **Percentage of households with high speed internet** (with a flow rate not less than 100 Mb/s);
9. **Percentage of population with high speed internet**. Source for 6, 7, 8, and 9: Eurostat survey on ICT use by households or individuals, (http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/FR/isoc_bde15c_esms.htm ITU, UN specialized agency for ICT, www.itu.int/ITU-D/ict/statistics/, 2012);
10. **Percentage of population that uses internet every day**;
11. **Percentage of population that uses internet once a week**. Source for 10 and 11: Eurostat survey on ICT use by households or individuals, (http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/FR/isoc_bde15c_esms.htm);
12. **The use of Facebook**, as the number of users as percentage of population;
13. **The use of Facebook as the number of users as on-line users** is measured by the Socialbakers Company, (Source for 12. and 13: http://www.socialbakers.com/facebook-statistics/);
14. **Video uploads on YouTube** – number of video clips on YouTube measured for a population of 15 to 69 years of age. Source: Global Innovation Index, 2012;
15. **Wikipedia, monthly editing on 100 internet users** is measured on 100 Internet users. Source: Wikimedia Analysis on Information Flow, (stats.wikimedia.org, 2012);
16. **The number of mobile telephony** subscribers in relation to the population – the number of subscriptions to a public mobile telephone service, which provides access to a public fixed network (PSTN) using mobile technology. Source: ITU, UN specialized agency for ICT, (www.itu.int/ITU-D/ict/statistics/, 2012);
17. **Sophistication of service – citizens**;
18. **Sophistication of service – companies** – online sophistication is a measure of the level of development of government services. Source: European Commission, Digitizing of public services in Europe, 9 benchmark measurement 2009, European Commission, 2010;

19. **Active mobile broadband internet users per 100 inhabitants**. Source: ITU, UN specialized agency for ICT, (www.itu.int/ITU-D/ict/statistics/, 2012);
20. **Wikipedia – the percentage of share in total monthly postedcontent.** Source: Wikimedia Traffic Analysis Report, (stats.wikimedia.org, 2012);
21. **E-government** – the percentage of citizens aged from 16 to 74 who use public services available online – the percentage of citizens who use e-government services (within last three months);
22. **E-government** – the percentage of companies who use public services available online. Source: Eurostat survey on the use of ICT and e-commerce in companies, Eurostat, 2010;
23. **E-commerce** – the percentage of population that orders goods or services via the Internet, Source: Eurostat survey on the use of ICT by households and individuals, (http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/FR/isoc_bde 15c_esms.htm);
24. **Mobile phone services** – the average cost per minute of different types of mobile calls (in PPP $) are measured;
25. **Rates for fast internet** – the payment of monthly subscriptions for fixed broadband Internet service is measured (PPP (Purchasing Power Parity) in $). Source for 24 and 25: United Nations specialized agency for ICT, (http://www.itu.int/ITU-D/IKT/statistics/, 2012);
26. **Percentage of highly educated population (30-34)**. Source: European Commission, Innovative List, Innovation Union Scoreboard, UNU-MERIT, 2010;
27. **No. of students per 100 000 inhabitants**, ISCED Classification 5 and 6;
28. **Percentage of graduate students in engineering, manufacturing and construction**, in relation to total no. of graduate students. Source for 27 and 28: UNESCO Institute, within online statistical report, (www.stats.uis.unesco.org, 2012);
29. **Faculty enrolment**, percentage of the total number of secondary school graduates – ISCED Classification 5 and 6. Source: World Bank, World Development Indicators Online, (http://data.worldbank.org/indicator /SE.TER.ENRR, 2012);
30. **Percentage of students who study abroad**;
31. **Percentage of enrolled PhD students** within the total number of enrolled students, ISCED Classification, 6. Source for 30 and 31: UNESCO Institute, statistical report,(www.stats.uis.unesco.org, 2012);
32. **Percentage of employees with ICT skills**. Source: Eurostat Labour Force Survey, Eurostat, 2011a;

33. **No. of researchers on 1000 inhabitants** – total., Source: UNESCO, (www.stats.uis.unesco.org, 2012);
34. **PISA scale – reading**;
35. **PISA scale – mathematics**;
36. **PISA scale – science** is based on generally accepted PISA testing programs of elementary school students. Source: OECD, (http://www.oecd.org/pisa/, 2012);
37. **Percentage of rural population**, Source: World Bank, (http://data. worldbank.org/indicator/SP.RUR.TOTL.ZS/countries, 2012);
38. **Number of doctorates on 1000 inhabitants** aged from 25 to 34, source: European Commission, Innovative List, Innovation Union Scoreboard, UNU-MERIT, 2010;
39. **Implementation phase of the first and second cycle of Bologna** is an indicator of educational development in European countries;
40. **Phase of external quality system** – the qualification framework has been introduced in the Bologna agenda between 2001 and the 2003;
41. **Implementation phase of te ECTS system** – the European Credit Transfer System (ECTS) is student credit system, which is a measure of required students' work necessary to achieve certain outcomes;
42. **Implementation phase of the diploma supplement.** Source for 39, 40, 41, 42: Eurostat, The European Higher Education Area in 2012: Bologna Process Implementation Report, (Eurostat, 2012);
43. **Participation of the ICT sector** (manufacturing and services) **in GDP**;
44. **Participation of the ICT sector** (manufacturing and services) **in total employment**. Source: Eurostat evaluation based on Structural Business Statistics and national accounts statistics (http://epp.eurostat.ec.europa.eu, 2012);
45. **ICT export services** (percentage of total exports services). Source: International Monetary Fund, The Statistical Yearbook of Balance of Payments and Data Files, published by the World Bank, (http://data.worldbank.org/indicator/BX.GSR.CCIS.ZS/countries, 2012);
46. **ICT export of products** (percentage of total export of products). Source: UNCTAD database of the UN Conference on Trade and Development, published by the World Bank. (http://data.worldbank.org/indicator/ TX.VAL.IKTGODINEZS.UN?cid=GPD_31, 2012);
47. **Percentage of employees with university education**, aged from 15 to 64 years, compared to the total number of employees. Source: Eurostat, Labour market statistics, Eurostat, 2011a;

48. **Investment in research and development** - percentage of GDP. Source: UNESCO Institute, www.stats.uis.unesco.org, 2012;
49. **Export of knowledge - intensive services** (percentage of total export of services). Source: UN Statistics Division, unstats.un.org, 2012;
50. **Spending on tertiary (higher) education** per student, percentage of GDP. Source: UNESCO Institute, regular online statistical report, (www.stats.uis.unesco.org, 2012);
51. **Spending on education**, as a percentage of GDP, Source for 49 and 50: UNESCO Institute, regular online statistical report, (www.stats.uis.unesco.org, 2012);
52. **Percentage of creativity export of total percentage of services** – sum of credits in EBOPS (classification of the extended payment balance services). Source: UNCTAD, Creative Economy Report, 2010;
53. **Percentage of creativity export of total percentage of goods** measures the technological competitiveness of the EU, i.e. ability to commercialize the results of research and development and innovation in the international markets. It also reflects the specialization of production by countries. Source for 51 and 52: UNCTAD, Creative Economy Report, 2010;
54. **Percentage of export of services related to computers** (percentage of commercial services). Source: The International Monetary Fund, Statistical Yearbook of Payment Balance and Data Files, World Bank, http://data.worldbank.org/indicator/TX.VAL.OTHR.ZS.WT, 2012);
55. **High technology export** (percentage of total export of goods). Source: World Bank (http://data.worldbank.org/indicator/TX.VAL.TECH.MF.ZS/countries ?display=default, 2012);
56. **Employment in knowledge-intensive professions.** Source: Eurostat, Labour Market Statistics, 2011;
57. **Number of scientific publications per million inhabitants - SCI list**. Source: Thomson Reuters (Scientific) Inc. Web of Science, Science Citation Index Expanded;
58. **Number of academic and professional articles in journals**;
59. **Number of academic and professional articles in per million inhabitants**. Source: The National Science Foundation, Science and Engineering Indicators, World Bank, (http://data.worldbank.org/indicator/IP.JRN.ARTC.SC/ countries?display=default, 2012);
60. **Number of patent applications per million inhabitants**. Source: World Intellectual Property Organization, World Intellectual Property Indicators, WIPO, 2011;

61. **The use of pure and nuclear energy in total consumption** in percentages, Source: World Bank, (http://data.worldbank.org/indicator/EG.USE. COMM.CL.ZS, 2012);
62. **The price of electricity – households e/100kWh**;
63. **The price of electricity - industry e/100kWh**. Source: Eurostat, Indicators of Energy, Transport and Environment (Eurostat, 2011b);.
64. **Greenhouse gas emissions in CO_2 per capita**, measured by the Information Centre Analysis of Carbon Dioxide, the Department of Ecological Sciences, Oak Ridge National Laboratory, Tennessee, USA, World Bank (http://data.worldbank.org/indicator/EN.ATM.CO2E.PC?display=default);
65. **Energy consumption in 1000 kWh per capita**, measured by The International Energy Agency and Eurostat, Indicators of energy, transport and environment (Eurostat, 2011b).

4.2.2 Selection, Size and Construction of the Sample

Sampling in this article was conducted according to the data of statistical yearbooks of the analysed countries, the database of Eurostat, the European Commission, the World Bank, the ITU, the UNECO, as well as on the basis of other relevant studies dealing with the measurement of competitiveness. Research and data analysis was performed for the calendar year 2010. The new model of Competitiveness Index of the Knowledge Based Society has been created and applied to the Republic of Serbia, the other Western Balkan countries and selected countries in Europe. The Autonomous Province Vojvodina of the Republic of Serbia is listed in the comparison on the regional level. The selected territorial units are: Sweden (SE), Finland (FI), Switzerland (CH), Denmark (DK), Norway (NO), Germany (DE), Austria (AT), Slovenia (SI), Montenegro (MN), Hungary (HU), Croatia (HR), Romania (RO), Bulgaria (BG), Macedonia (MK), Albania (AL), Bosnia and Herzegovina (BA), Serbia (RS) and Vojvodina (VO).

This choice was made in order to test the index for countries with different levels of development and status in the EU (member states, accessing countries, candidate countries, potential candidate countries for EU membership, the member states of the European Economic Community), and all of them should fit into the future knowledge-based society of Europe.

4.2.3 Transformation of Irreversible Data and Replacing of Missing Data

For indicators of irreversible character (3, 4, 24, 25, 37 62, 63 and 64), in the sense that lower value indicates a higher level of development, it is necessary to make a transformation:

$$Xtrans = 2 * (Xmax – Xmin) - Xi \quad (1)$$

One of the fundamental problems in the selection of variables was lack of available and comparable data. In this article, we used the nearest neighbour

method, on which basis values are complement according to estimate in respect of the most similar case.

4.2.4 Classification of Indicators into Thematic Groups

The 65 selected indicators are classified into six thematic subindices: General Preconditions, Using Advanced Technologies, Education, Research and Development, Innovation and Sustainable Development.

The General Preconditions subindex represents a conditional element of the knowledge-based society is the economic impact, consisting of items 1 to 5 from the list of indicators.

The Use of Advanced Technologies subindex. Effective communication, distribution, assimilation and development of ideas and knowledge are facilitated by providing a modern and adequate infrastructure. ICT are essential factor of the knowledge-based society. This subindex consists of items No. 6 through 25 form the list of indicators.

The Education subindex. Human capital refers to the well-educated and skilled workforce [19, 5, 2]. This subindex consists of indicators listed under numbers 26 to 42.

The Research and Development subindex refers to the development of an effective innovation system in firms, research centres and other relevant organizations and institutions, which results in new goods, new processes and new knowledge. This subindex encompasses indices from 43 to 56.

The Innovation subindex consists of indices 57 to 60.

The Sustainable Development subindex consists of indices 61 to 65.

4.3 Standardization of Individual Indicators and Weighting

In order to avoid problems of mixing different measuring units they should be normalized or standardized. Different techniques can be used in this way, and each has its advantages and disadvantages and can produce different results [20].

To obtain average equal to 100 for all variables, the following conversion was applied:

$$s_{ij} = \frac{x_{ij}}{\bar{x}_j} \cdot 100 \qquad (2)$$

where:

x_{ij} is value of the j-th indicator of indicator of the i-th state;

s_{ij} is standardized value of the j-th indicator of indicator i-th of state; and

$\bar{x}_j$ is1 average value of the j-th indicator.

Variables that are used for construction of the competitiveness index must be weighted to reflect the importance, reliability and other characteristics of the basic data. A way to identify the appropriate weights is through empirical analysis, especially using methods based on correlations among the used variables (e.g. regional analysis, principal component analysis, factor analysis, etc.) [23]. However, it is not certain that the correlations will correspond to real connections between the measured phenomena [15]. Alternatively, the weights can be set up in cooperation with various interested parties (e.g., experts, creator of policies, etc.), on condition that they understand the strengths, weaknesses and peculiarities of data within a given theoretical framework or the weights can be assigned according to the quality and availability of data. Since different weighting techniques can produce quite different results, no weighting approach is safe tool of obtaining credible results. For this reason, in [3] it is argued that the same weighting should be the norm. In [4] this attitude is accepted on the basis of simplicity, in terms of composite construction and interpretation [20].

Compression of standard values using weights with the total sum of 1 was used to form corresponding subindices. Composite subindices were further weighted to form a common composite index. The structure of the composite index with weighting factors of its subindices is shown in Figure 1. This approach is different from [1, 4], because authors assigned different weight values based on their contribution. The highest weights were given to Using Advanced Technologies (25%) and Education (25%) because authors believe that these are groups of parameters that have the greatest influence on the development of the knowledge society and also these subindices contain the largest number of individual parametres. The General Precondition subindex consists of 5 indicators, and is weighted 20%. The Political Stability parameter is weighted 30%, while the remaining 3 got the value of weighting factor of 10%.

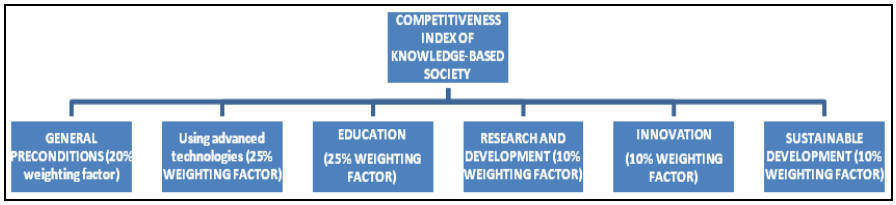


Figure 1

Competitiveness Index of Knowledge-Based Society

Table 1 shows the cumulative standardized and weighted values of parameters summarized in composite subindices. Assigning weight coefficients to subindices, as well as aggregation, i.e. adding the values of the composite subindices and the formation of the final composite index and ranking of countries and regions were carried out in the Table 2.

Table 1
Composite subindices – standardized and weighted values and the composite index

Composite subindices	General Preconditions (GP)	Using Advanced Technologies (AT)	Education (ED)	Research and Development (RD)	Innovation (IN)	Sustainable Development (SD)
SE	149.5	140.9	136.6	150.7	189.5	153.2
DE	129.3	119.9	114.3	124.1	365.4	75.0
FI	145.7	139.3	145.5	146.7	173.0	124.1
AT	142.1	120.7	105.1	107.0	131.0	106.2
CH	176.5	116.6	125.5	128.8	216.0	101.3
DK	160.9	142.1	126.9	134.6	170.9	88.1
NO	206.1	144.8	128.0	116.2	153.8	181.4
SI	95.2	114.0	102.2	94.1	107.3	88.3
ME	68.8	72.9	88.0	60.7	6.1	87.1
HU	78.8	99.5	92.7	136.3	52.5	81.7
HR	77.5	86.9	85.1	80.3	44.7	85.0
RO	67.1	67.5	89.7	101.3	24.5	98.6
BG	56.9	78.5	91.8	83.9	22.6	88.0
MK	52.4	78.1	69.9	66.4	36.4	85.3
AL	48.9	62.3	69.8	41.1	17.4	113.1
RS	52.1	78.6	80.7	93.2	37.6	88.6
VO	51.8	80.4	88.0	91.8	45.0	78.3
BA	40.4	57.1	60.1	43.0	6.3	76.8
Sum:	1800.0	1800.0	1800.0	1800.0	1800.0	1800.0

Table 2
Composite index – assigning weight coefficients to subindices and aggregation and formation of the competitiveness index

Composite subindices	GP	AT	ED	RD	IN	SD	COMPO-SITE INDEX	RANK
Weight	**0.2**	**0.25**	**0.25**	**0.1**	**0.1**	**0.1**		
SE	29.9	35.2	34.2	15.1	18.9	15.3	148.6	**2**
DE	25.9	30.0	28.6	12.4	36.5	7.5	140.9	**4**
FI	29.1	34.8	36.4	14.7	17.3	12.4	144.7	**3**
AT	28.4	30.2	26.3	10.7	13.1	10.6	119.3	**7**
CH	35.3	29.1	31.4	12.9	21.6	10.1	140.4	**5**
DK	32.2	35.5	31.7	13.5	17.1	8.8	138.8	**6**
NO	41.2	36.2	32.0	11.6	15.4	18.1	154.6	**1**
SI	19.0	28.5	25.6	9.4	10.7	8.8	102.1	**8**
ME	13.8	18.2	22.0	6.1	0.6	8.7	69.4	**15**
HU	15.8	24.9	23.2	13.6	5.2	8.2	90.9	**9**

HR	15.5	21.7	21.3	8.0	4.5	8.5	79.5	**10**
RO	13.4	16.9	22.4	10.1	2.4	9.9	75.2	**11**
BG	11.4	19.6	22.9	8.4	2.3	8.8	73.4	**13**
MK	10.5	19.5	17.5	6.6	3.6	8.5	66.3	**16**
AL	9.8	15.6	17.4	4.1	1.7	11.3	59.9	**17**
RS	10.4	19.7	20.2	9.3	3.8	8.9	72.2	**14**
VO	10.4	20.1	22.0	9.2	4.5	7.8	73.9	**12**
BA	8.1	14.3	15.0	4.3	0.6	7.7	50.0	**18**
Sum:	360.0	450.0	450.0	180.0	180.0	180.0	1800.0	

According to the developed composite index of competitiveness of the knowledge-based society, Norway stands out as the first on the ranking list with 153.2 index points. It is followed by Sweden (149.3), Germany (147.7), Finland (144.9), Switzerland (141.3) and Denmark (138.4). Although located on the 7th place, Austria has an above average number of points (117.7), while the eighth placed Slovenia represents an average ranking of countries with 95.5 index points. Hungary occupies the ninth place with 89.8 points. The Western Balkan countries (and the region of Vojvodina) have a similar number of scored points, and are ranked in the following order after Hungary: Croatia (78.6), Vojvodina (74.3), Romania (74.2), Serbia (72.9), Bulgaria (72.7) and Montenegro (69.3). Lower-ranked countries in the Western Balkans are Macedonia (66.0), Albania (59.9) and Bosnia and Herzegovina (50.1).

4.4 Testing the Composite Indicator

As it has been already mentioned, there are several possibilities regarding the selection, standardization and aggregation of variables into one composite indicator. The results depend on the chosen approach. For this reason, sensitivity tests are conducted to analyse the impact of the inclusion or exclusion of different variables, change in weights, the use of different techniques of standardization, etc., on the results of the composite indicator. The combination of uncertainty and sensitivity analysis can be used to estimate the robustness of the composite indicator as well as for quality improvement. The uncertainty analysis examines how uncertainty is propagated within the input data through the structure of the composite indicator and affects its value, while sensitivity analysis evaluates the contribution of individual source of uncertainty to deviation of the final result. Composite indicators usually measure phenomena that are related to the well-known and measurable concept (e.g. economic growth). These connections can be used for testing the strength of composite explanation. Common cross-plot method provides a good way to illustrate such connections. Correlation analysis is equally useful for testing, whereby high correlation indicates a composite indicator of high quality [20]. This article uses the methods of correlation, regression and variance.

4.4.1 Regression Analysis Based on the Indicators of Economic Dynamism

The indicator of economic dynamism shows how GDP per capita affects the final result. The indicator is obtained in such a way that standardization of values of the composite index using the "minimum-maximum method" is first performed. This method transforms real values to values between zero (minimum value) and one (leader of the maximum value). This gives a picture of the distance of some country from the best and worst ranked state, i.e. common composite index compared to the difference between maximum and minimum.

The common composite index compared to the difference between maximum and minimum (yi) is calculated according to the formula:

$$yi = (Xi - Xmin) / (Xmax - Xmin) \quad (3)$$

where yi is the standardized value, Xi is the actual value, Xmax is the maximum value and Xmin is the minimum value.

Standardization as a method does not affect the ranking of countries for individual indicators. Indicators of economic dynamism (ECi) are:

$$(ECi) = GDPi\ (1 + yi) \quad (4)$$

where yi is a common composite index in relation to the difference between maximum and minimum, and GDPi per capita in USD thousands.

The obtained Indicator of Economic Dynamism is shown in Table 3 together with the Competitiveness Index of the Knowledge-Based Society. Ranks are assigned to the countries/regions in both situations. Rank type I refers to the classification of economies according to the Competitiveness Index of the Knowledge-Based Society and Rank type II on the results of Indicator of Economic Dynamism.

Table 3

Composite Competitiveness Index of the Knowledge-Based Society and indicator of Economic Dynamism – ranking of countries

COUNTRY	The common Composite Competitiveness Index of the Knowledge-Based Society	RANK type I	The common Composite competitiveness index in relation to the difference between maximum and minimum	Gross domestic product per capita in USD thousands	Indicator of Economic Dynamism	RANK type II
SE	149.3	**2**	0.9623	48.9	96.0	**4**
DE	147.7	**3**	0.9468	40.2	78.2	**6**
FI	144.9	**4**	0.9197	45.1	86.5	**5**
AT	117.7	**7**	0.6558	45.2	74.8	**7**
CH	141.3	**5**	0.8849	67.5	127.2	**2**
DK	138.4	**6**	0.8561	55.9	103.7	**3**
NO	153.2	**1**	1.0000	84.5	169.1	**1**
SI	99.5	**8**	0.4786	22.9	33.8	**8**

ME	69.3	**15**	0.1861	6.5	7.7	**12**
HU	89.8	**9**	0.3849	12.9	17.8	**9**
HR	78.6	**10**	0.2765	13.8	17.6	**10**
RO	74.2	**12**	0.2340	7.5	9.3	**11**
BG	72.7	**14**	0.2189	6.3	7.7	**13**
MK	66.0	**16**	0.1539	4.5	5.1	**16**
AL	59.9	**17**	0.0948	3.7	4.0	**18**
RS	72.9	**13**	0.2208	5.3	6.4	**14**
VO	74.3	**11**	0.2345	5.0	6.2	**15**
BA	50.1	**18**	0.0000	4.4	4.4	**17**
Sum:	1800.0		8.7087			

4.4.2 Analysis of the Range of Variation and Variance

This analysis shows how big are differences between the top-ranked and lowest-ranked economy according to composite subindices. The standard deviation represents the average deviation from the average value. The coefficient of variation represents the quotient of the standard deviation and average value. Analysis results are presented in Table 4. Some groups of parameters have a greater range of variations, and other less. The difference between the Innovation and Education subindices is particularly significant. Namely, the difference between the best and worst ranked countries of 359.2 points within the subindex Innovation is measured. This result can be explained by conspicuous differences between Western countries and the Western Balkan countries in terms of the parameters that make this subindex. However, the analysis showed that these differences are not so significant within the Education and Using Advanced Technologies subindices. More than four times lower difference than within the Innovation subindex is measured within Education subindex, amounting to 85.4 points. Similar is with the Using Advanced Technologies subindex, for which this difference is 92 index points. The analysis of the coefficient of variance shows similar results. The coefficient of variation of 96.5%, which expresses great differences within the analyzed countries in terms of this subindex measured within the Innovation subindex. On the other hand, the coefficient of variation of 24.65% was measured within the Education subindex, which shows significantly smaller differences between countries when it comes to this subindex. There is a long tradition in education in the Western Balkans territory, and this is the reason why these differences are less pronounced compared to Western European countries.

Table 4

Analysis of the range of variation and variance

Composite subindex	**GP**	**AT**	**ED**	**RD**	**IN**	**SD**
max	40.36	57.73	60.10	41.14	6.12	75.03
min	206.13	147.29	145.48	150.67	365.37	181.37
range of variation	165.78	89.56	85.38	109.53	359.24	106.34

variance	2701.58	1064.48	601.11	1112.18	9326.86	1080.04
standard deviation	51.98	32.63	24.52	33.35	96.58	32.86
average value	100.00	100.00	100.00	100.00	100.00	100.00
coefficient of variation	51.98%	32.63%	24.52%	33.35%	96.58%	32.86%

4.4.3 Correlation Matrix of Composite Indicators

The matrix (Table 5) shows the correlation of economic dynamism (ECD) with other subindices.

Table 5
Correlation matrix of composite indicators

	ECD	**GP**	**AT**	**ED**	**RD**	**IN**	**SD**
ECD	1.000	0.986	0.872	0.855	0.686	0.752	0.685
GP	0.986	1.000	0.898	0.895	0.741	0.764	0.647
AT	0.872	0.898	1.000	0.921	0.843	0.892	0.489
ED	0.855	0.895	0.921	1.000	0.880	0.758	0.575
RD	0.686	0.741	0.843	0.880	1.000	0.700	0.369
IN	0.752	0.764	0.892	0.758	0.700	1.000	0.251
SD	0.685	0.647	0.489	0.575	0.369	0.251	1.000

All relations that have a value above 0.7 are significant. According to the overview in Table 5, it can be concluded that the economic dynamism is associated with all the subindices except the Research and Development and Sustainable Development subindices, with which it is least consistent. The analysis of correlation is equally useful for testing, where high correlation indicates a composite indicator of high quality [20].

5 Dilemmas and Reflections

The Index of Competitiveness of the Knowledge-Based Society was calculated for 17 countries and one region, through standardization, weighting and aggregation. This choice was made in order to test the index for countries with different levels of development and status in the EU. Results showed that Serbia and Vojvodina are rated at about 75% of the average of the selected countries, and at the level of neighbouring EU countries and Croatia (which joined the EU in July 2013).

Testing of the composite index was conducted through the indicators of economic dynamism that show how Competitiveness Index of the Knowledge-Based Society depends on the parameters of GDP per capita. It was found that the differences between countries of Western Europe and the Western Balkans

are dramatically higher when they are reflected through the GDP, as a representative indicator of economic development, than when this difference is measured in the parameters of knowledge.

The range of variations of different parameters is various. The difference between the *Innovation* and *Education* subindices is particularly significant. This result can be explained by prominent differences between Western countries and the Western Balkan countries in terms of scientific achievements which are reflected in the number highly qualified articles and in the patent applications. On the other hand, the difference within the *Education* subindex is measured more than four times lower compared to the subindex *Innovation* which amounts 85.4 points. It shows significantly smaller differences between countries when it comes to this subindex. Obtained results can be explained by the existence of a long tradition of education in the Western Balkan countries, and this is the reason why these differences are less pronounced compared to Western European countries. The sub-hypotheses of the article was accepted in this manner: **Certain groups of subindices can be distinguished, where in some cases the difference between the best and worst ranked countries is small, as well as those groups of subindices where these differences are significant.**

The authors of this article were in dilemma, that is, how to substitute for the missing data. They have chosen not to use qualitative data, but at the expense that there are some non-described fields due to lack of data. For that reason, some proxy data are used. Although a large number of quantitative data was used, as many as 65, the dilemma was whether it was enough, that is, whether using more data it would provide even better results. The authors have chosen this number for practical reasons, i.e. availability of data. This dilemma leads to the appointment of a hypothesis for future research: with improved data availability, it is possible to describe knowledge-based society.

It is also very important to follow the trends of knowledge-based society and constantly adapt methodologies in this direction, by creating new and rejecting unnecessary parameters that describe the knowledge-based society. So, it would be suitable, for example, to include into future research the indicator of Percentage population who uses smartphones or the indicator Percentage of smart TVs in households. Besides, it is always necessary to analyze the importance of each and every group of parameters (subindices), in order to perform an adequate allocation of weighting factors. Also, during the standardization and transformation of data, it is possible to apply several methods, and it is necessary to consider whether the selected one is appropriate.

Conclusion

The main goal of this article was to create a new model for estimating the national knowledge competitiveness, whose implementation would help to achieve improvement the quality of monitoring in this area. It was found that Serbia and other countries in transition are not analysed adequately when in terms of the competitiveness of knowledge, that the existing indices analysing

Serbia do not provide enough information about the state of development of knowledge in Serbia, and that Serbia is at the very bottom when compared to European and countries worldwide. It was also noticed that the existing models of knowledge contain a large number of qualitative indicators, which are subject to manipulative influences of experts, while models based on quantitative indicators consists of a small number of parameters. This is why it was concluded that the existing models of competitiveness that contain parameters of knowledge are not suitable for countries in transition like Serbia.

The results of this paper can be used primarily for adjustment of existing and development of new strategies, which accompany the European, national and regional documents. The results are also applicable for the identification of negative trends in regional development and its balancing. Paper also contributes to easier control of the set goals, discover reasons continuous low ranking of Serbia, and draw conclusions for strategic development planning.

It was suggested that a new, revised model should be set for successful monitoring of progress and the degree of achieving the goals set in strategies of development of European countries and Serbia until 2020. The basic hypothesis of the work is accepted: **The new competitiveness model based on knowledge and predominantly expressed quantitative parameters gives a more realistic evaluation of competitiveness of a country.**

Developing models of competitiveness often entails various difficulties. In the efforts to include as many indicators that describe the desired phenomenon, many authors face the problems of data collection and the lack of data for individual countries and regions. In that case, some authors try to insert qualitative indicators or assessments, which could lead to manipulation with the results by subjective evaluations.

Hopefully this paper has special importance for scientific and research workers studying fields related to knowledge and knowledge competitiveness. The range of this article can be expanded in further research. Taking a sample that would include all the European countries, more relevant comparison and ranking of economies could be made. Besides, future multi-annual monitoring of the competitiveness, according to the model developed in the article, would allow tracking of progress and growth rate of overall results and the individual parameters and groups of parameters for selected countries. The possibility that also occurs for future researchers is calculation of the Competitiveness Index of the Knowledge-Based Society for European regions. Significant results could also be obtained by sensitivity analysis of individual parameters and groups of parameters on the overall index result, which would contribute to making adequate conclusions and guidelines from the research.

References

[1] Babbie E. (1995), The Practice of Social Research. Wadsworth, Belmont

[2] Barro, R., Sala-i-Martin, X. (1995), Economic Growth. McGraw-Hill, New

York. Chs. 1-4

[3] Baylis J., Smith S., The Globalization of World Politics, An Introduction to International Relations (New York: Oxford University Press, 1999), p. 15

[4] Booysen F. (2002), An Overview and Evaluation of Composite Indices of Development. Soc Indic Res 59:115-115

[5] Brunetti A. (1997), Political Variables in Cross-Country Growth Analysis. J of Econ Surv 11(2):163-190

[6] Chen D. H. C., Dahlman C.J. (2005), The Knowledge Economy, the KAM Methodology and World Bank Operations. World Bank Institute Working Paper no. 37256

[7] Daugėlienė R. (2006), Towards Knowledge-Based Economy: Modelling Knowledge Expression Assessment. European Union Enlargement of 2004 and Beyond: Responding to the Political, Legal and Socio-Economic Challenges, 451-469

[8] Drucker P. (1969), The Age of Discontinuity. Guidelines to Our Changing Society. Harper and Row: New York

[9] Drucker P. (2006), My View of Management - Ideas that Have Improved Management. Novi Sad: Adizes (In Serbian)

[10] European Commission (2010), Digitizing Public Services in Europe: Putting Ambition into Action, 9th Benchmark Measurement, Capgemini, IDC, Rand Europe, Sogeti and DTi, European Commission, Directorate General for Information Society and Media

[11] Eurostat (2010), Internet Use in Households and by Individuals in 2010, Eurostat

[12] Eurostat (2011 b), Energy, Transport and Environment Indicators, Eurostat

[13] Eurostat (2011a), European Union Labour Force Survey, Eurostat

[14] Eurostat (2012), The European Higher Education Area in 2012: Bologna Process Implementation Report, ISBN 978-92-9201-256-4, Eurostat

[15] Freudenberg M. (2003), Composite Indicators of Country Performance: a Critical Assessment. OECD Science, Technology and Industry Working Papers, 2003/26, OECD Publishing, Paris

[16] Huggins R. and Izushi H. (2007), Competing for Knowledge**,** Creating, Connecting, and Growing, Routledge, New York

[17] Katić A., Ćosić I., Anđelić G., Raletić S. (2012), Review of Competitiveness Indices that Use Knowledge as a Criterion, Acta Polytechnica Hungarica 9(5):25-45, ISSN 1785-8860

[18] Mankiw N., Romer D., Weil D. (1992), A Contribution to the Empirics of Economic Growth. Q J Econ 107(2):407-437

[19] Nardo M., Saisana M., Saltelli A., Tarantola S., Hoffman A., Giovannini E. (2005), Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Statistics Working Paper CTD/DOC(2005)3, OECD, Paris

[20] Nijkamp P., Siedschlag I. (2011), Innovation, Growth and Competitiveness, Dynamic Regions in the Knowledge-Based World Economy, Springer Heidelberg Dordrecht London New York

[21] Önsela S., Ülengina F., Ulusoyb G., Aktaşc E., Kabakc O., Topcuc Y. (2008), A new Perspective on the Competitiveness of Nations. Socio-Economic Planning Sciences. 42(4): 221-246

[22] Organisation for Economic Cooperation and Development (OECD). (1996), The Knowledge-Based Economy. Paris: OECD

[23] Porter M., Stern S. (1999), The New Challenge to America's Prosperity: Findings from the Innovation Index. Council of Competitiveness, Washington DC

[24] Saisana M. et al. (2005), State of the Art Report on Composite Indicators for the Knowledge-Based Economy. Deliverable 5.1 of the WP5 of the KEI project

[25] UNU-MERIT (2010) Innovation Union Scoreboard, The Innovation Union's Performance Scoreboard for Research and Innovation, Maastricht Economic and Social Research and Training Centre on Innovation and Technology (UNU-MERIT). www.proinno-europe.eu/metrics

[26] www.itu.int (2012), International Telecommunication Union

[27] www.oecd.org/pisa/ (2012) OECD Programme for International Student Assessment (PISA)

[28] www.sr.wikipedia.org (2012), Wikipedia

[29] www.stats.uis.unesco.org (2012), UNESCO Institute for Statistics

[30] www.stats.wikimedia.org (2012), Wikimedia Statistics

[31] www.tempus.ac.rs (2012), Kancelarija Tempus fondacije

[32] www.theatlantic.com (2012), Magazine Atlantic

[33] www.unstats.un.org (2012) United Nations Statistics Division