Basic equations of fluid dynamics treated by pseudo-analysis

Endre Pap*,,***, Doretta Vivona******

* Department of of Mathematics and Informatics, University of Novi Sad, Trg D. Obradovića 4, 21000 Novi Sad, Serbia, email:pap@dmi.uns.ac.rs

** Óbuda University, Becsi út 96/B, H-1034 Budapest, Hungary,

*** Educons University, Vojvode Putnika 87, 21208 Sremska Kamenica, Serbia,

**** Dipartimento di Scienze di Base e Applicate per l'Ingeneria, "Sapienza"- Universitá di Roma, Via A. Scarpa 10, 00161 Rome, Italy, e-mail:doretta.vivona@uniroma1.it

Abstract. There is given an application of pseudo-analysis in the theory of fluid mechanics. First, the monotonicity of the components of the velocity for the solutions of Euler equations is proven, which allows to obtain the pseudo-linear superposition principle for Euler equations. This principle is proven also for the Navier-Stokes equations but with respect to two different pairs of pseudo-operations. It is shown that Stokes equations satisfy the pseudo-linear superposition principle with respect to a pair of pseudo-operations which are generated with the same function of one variable.

Keywords: Fluid mechanics, Euler equations, Navier-Stokes equations, Stokes equations, pseudo-linear superposition principle, semiring.

.

1 Introduction

The motion of fluids was mathematically modeled in the period of more than two hundred years. The ordinary incompressible Newton Fluids are modeled by the Navier-Stokes equations and the related Euler equations. Some of the recent investigations are summarized in the two volumes of the Handbook of Mathematical Fluid Dynamics ([5, 6]).

We shall prove in this paper an important property of the three basic equations (Euler, Navier-Stokes, Stokes), the so called *pseudo-linear superposition principle*. To achieve this principle in full generality we shall neglect at this level the problem of the regularity of the solution, which is a very important part of the investigations in fluid dynamics, see ([1, 2, 3, 24]).

What we are doing, roughly speaking, is that we replace the usual field of real numbers by a semiring on a real interval $[a, b] \subset [-\infty, \infty]$ ([7, 8, 11, 12, 14]), where the corresponding operations are $\oplus$ (pseudo-addition) and $\odot$ (pseudo-multiplication). Based on the semiring structure there is developed in ([12, 13, 14, 15, 18, 19]) the so called pseudo-analysis, in an analogous way as classical analysis, introducing pseudo-measure, pseudo-integral, pseudo-convolution, pseudo-Laplace transform, etc.([15, 16, 17, 18, 20, 21, 22]). The advantage of the pseudo-analysis is that the problems (usually nonlinear) from many different fields (system theory, optimization, control theory, differential equations, difference equations, etc.) are covered with one theory, and so with unified methods. The pseudo-analysis is used for solving nonlinear equations (ODE,PDE, difference equations, etc.), based on pseudo-linear superposition principle, which means that if u_1 and u_2 are solutions of the considered nonlinear equation, then also $a_1 \odot u_1 \oplus a_2 \odot u_2$ is a solution for any numbers a_1 and a_2 from $[a, b]$. The important fact is that this approach gives also solutions in a new form, not achieved by other theories. In some cases it enables for the nonlinear equations to obtain exact solutions in a similar form as for the linear equations.

After some preliminaries in Section 2, and recalling some basic facts on the Euler equations in Section 3, we prove in Section 4 the monotonicity of the velocity for the solutions of the Euler equations. This help us to prove in Section 5 the pseudo-linear superposition principle for the Euler equations. This principle is achieved also for the Navier-Stokes with respect to two different pairs of pseudo-operations. In Section 6 it is shown that Stokes equations satisfy the pseudo-linear superposition principle but with respect to a pair of pseudo-operations which are generated with the same function of one variable.

2 Preliminary notions

We consider a fluid which occupies a 2-dimensional region, denoted by D, and we denote by ∂D the boundary of D. We denote by $\mathbf{x}$ the spatial coordinate $\mathbf{x} = (x, y)$, with t the time and with $\mathbf{u}$ the field of the velocity of each element: $\mathbf{u} = \mathbf{u}(\mathbf{x}, \mathrm{t}) = \mathbf{u}(u(x, y, t), v(x, y, t))$. Moreover we assume that the fluid has a well-defined mass density, indicated with $\rho = \rho(\mathbf{u}, t)$.

We shall use the following notations: grad $p = (p_x, p_y, p_z)$,

$$\partial_x = \frac{\partial}{\partial x} \ , \ \partial_t = \frac{\partial}{\partial t} \ ,$$

$$\text{div } \mathbf{u} = \nabla \mathbf{u} = \partial_x u + \partial_y v, \quad (\mathbf{u}\nabla)\cdot = u\partial_x \cdot + v\partial_y \cdot \quad .$$

The expression

$$\frac{D\cdot}{Dt} = \partial_t \cdot + (\mathbf{u}\nabla)\cdot \tag{1}$$

will be called the material derivative, and we have

$$\mathbf{a} = \frac{D\mathbf{u}}{Dt} = u\partial_x \mathbf{u} + v\partial_y \mathbf{u} + \partial_t \mathbf{u} = \partial_t \mathbf{u} + (\mathbf{u}\nabla)\mathbf{u} \ ,$$

$$\frac{D\rho}{Dt} + \rho \ \text{ div } \mathbf{u} = \frac{\partial \rho}{\partial t} \ + \ \text{div}(\rho \mathbf{u}) \ .$$

We consider two kinds of fluids:

– *incompressible fluid* if for any subregion W the volume is constant in t. This implies div $\mathbf{u} = 0$. From continuity equation and $\rho > 0$ it follows that the fluid is incompressible if and only if the mass density is constant: $\frac{D\rho}{Dt} = 0$;

– *homogeneous fluid* if the density ρ is constant in space.

The classical approach ([4]) is based on three assumptions:

1) *conservation of the mass* :
mass is neither created nor destroyed. The consequence of this principle is the so-called *continuity equation*:

$$\frac{\partial \rho}{\partial t} \ + \ \text{div}(\rho \mathbf{u}) \ = \ 0.$$

2) *balance of momentum* or *Newton's second law* :

$$\rho \frac{D\mathbf{u}}{Dt} = - \text{ grad p} + \mathbf{f} ,$$

where $\mathbf{f}$ are the forces.

3) *conservation of energy* :
energy is neither created nor destroyed.

3 Euler equations

In this paragraph we recall the equations of the motion of an incompressible fluid in 2-dimensional case. They are based on the Newton's second law, mass conservation and condition of incompressibility (*Euler equations*):

$$\rho \frac{D\mathbf{u}}{Dt} = - \text{ grad p} + \mathbf{f} \tag{2}$$

$$\frac{D\rho}{Dt} + \rho \text{ div } \mathbf{u} = 0$$

$$\text{div } \mathbf{u} = 0$$

$$\mathbf{u} \cdot \mathbf{n} = 0 \quad \text{ on } \partial D, \tag{3}$$

where $\mathbf{n}$ is the normal to the region D. (3) is the boundary condition.

The unknown functions of the system (2)-(3) are the components u, v of the velocity: $\mathbf{u} : I\!R \times I\!R \times I\!R^+ \to I\!R$, $\mathbf{u} = (u(\mathbf{x}, t), v(\mathbf{x}, t))$ and the pressure $p : I\!R \times I\!R \times I\!R^+ \to I\!R$. We denote by s the triple of the functions $u(\mathbf{x}, \text{t}), \text{v}(\mathbf{x}, \text{t}), \text{p}(\mathbf{x}, \text{t}) : \text{s} = (\text{u}(\mathbf{x}, \text{t}), \text{v}(\mathbf{x}, \text{t}), \text{p}(\mathbf{x}, \text{t}))$.

Without loss of generality we suppose that $\rho = 1$ and $\mathbf{f} = 0$.

Now we reformulate the equation (2) taking into account the definition of material derivative (1). We have

$$\partial_t \mathbf{u} + (\mathbf{u}\nabla)\mathbf{u} \ + \ \text{grad}\ p \ = 0 \tag{4}$$

$$\text{div}\ \mathbf{u} \ = \ 0 \tag{5}$$

$$v(x, y = 0, t) = 0 \quad . \tag{6}$$

As we have seen above, the velocity $\mathbf{u}$ depends on the variables x, y, t, in particular $y \in [0, \infty[$; the boundary condition (3) involves only the component of the velocity on the axe y, whose unit vector is $\mathbf{n}$.

Now, we project the first (vector) equation (4) on axes x and y:

$$\partial_t u + u\partial_x u + v\partial_y u + \partial_x p = 0 \tag{7}$$

$$\partial_t v + u\partial_x v + v\partial_y v + \partial_y p = 0 \quad . \tag{8}$$

We know ([2, 4, 10, 25, 26]) that the Euler equations are particular case of the Navier-Stokes equations when the viscosity ν of the fluid is zero. The solution of the Navier-Stokes equations can be well approximated by an Euler equation, when the viscosity is small, at least away from boundaries.

4 Monotonicity of the components of the velocity

Now we come back to the general discussion of the Euler equations (4)-(6). With the above notation, from the condition (5), we have for the Euler equations $\partial_x u + \partial_y v = 0$, i.e.,

$$v = -\int_0^y \partial_x u(x, y', t)\, dy'. \tag{9}$$

Proposition 4.1 . *Let* $\mathbf{u_1} = (u_1(\mathbf{x},t), v_1(\mathbf{x},t)), \mathbf{u_2} = (u_2(\mathbf{x},t), v_2(\mathbf{x},t))$ *be two velocities which satisfy the condition (5). If the function* $(u_2 - u_1)(\mathbf{x},t)$ *is either non-increasing or non-decresing with respect to* $\mathbf{x}$*, then the functions* $v_1(\mathbf{x},t)$ *and* $v_2(\mathbf{x},t)$ *satisfy either the following condition* $v_1 \leqslant v_2$ *or the condition* $v_2 \geqslant v_1$*, respectively, i.e., either*

$$\partial_x(u_2 - u_1) \leqslant 0 \Rightarrow v_1 \leqslant v_2$$

or

$$\partial_x(u_2 - u_1) \geqslant 0 \Rightarrow v_1 \geqslant v_2.$$

Proof. As the function $(u_2 - u_1)(\mathbf{x},t)$ is non-increasing with respect to $\mathbf{x}$, then $\partial_x(u_2 - u_1) \leqslant 0$. Therefore by the the condition (9) for v we have

$$\int_0^y \partial_x(u_2 - u_1)(x, y', t)\, dy' \leqslant 0,$$

and then

$$v_2 - v_1 = -\int_0^y \partial_x u_2(x, y', t)\, dy' - \left(-\int_0^y \partial_x u_1(x, y', t)\, dy'\right) \geqslant 0,$$

i.e., $v_2 \geqslant v_1$. □

From now on we consider the following sets of functions :

$$\mathcal{U}_{ni} = \{(u_1, u_2) \mid u_1 \leqslant u_2 \text{ and } \partial_x(u_2 - u_1) \leqslant 0\}$$

$$\mathcal{U}_{nd} = \{(u_1, u_2) \mid u_1 \geqslant u_2 \text{ and } \partial_x(u_2 - u_1) \geqslant 0\}.$$

As consequence of Proposition 4.1 we have the following:

Proposition 4.2 *If the couple of functions* (u_i, v_i) $i = 1, 2$ *satisfy the condition (9) and* u_i $i = 1, 2$ *are elements either of the set* $\mathcal{U}_{ni}$ *or the set* $\mathcal{U}_{nd}$*, then* $v_1 \leqslant v_2$ *and* $v_1 \geqslant v_2$*, respectively.*

5 Pseudo-linear superposition principle

5.1 Pseudo-analysis

We shall use the approach from ([14, 15, 18]). Let $[a, b]$ be a closed (in some cases semiclosed) subinterval of $[-\infty, \infty]$. We consider here a total order $\leq$

on $[a,b]$ (although it can be taken in the general case a partial order). The operation $\oplus$ (pseudo-addition) is a function $\oplus : [a,b] \times [a,b] \to [a,b]$ which is commutative, non-decreasing, associative and has a zero element, denoted by $\mathbf{0}$. Let $[a,b]_+ = \{x : x \in [a,b], \quad x \geqslant \mathbf{0}\}$. The operation $\odot$ (*pseudo-multiplication*) is a function $\odot : [a,b] \times [a,b] \to [a,b]$ which is commutative, positively non-decreasing, i.e. $x \leq y$ implies $x \odot z \leq y \odot z$, $z \in [a,b]_+$, associative and for which there exists a unit element $\mathbf{1} \in [a,b]$, i.e., for each $x \in [a,b]$, $\mathbf{1} \odot x = x$.

We suppose, further, $\mathbf{0} \odot x = \mathbf{0}$ and that $\odot$ is a distributive pseudo-multiplication with respect to $\oplus$, i.e.,

$$x \odot (y \oplus z) = (x \odot y) \oplus (x \odot z).$$

The structure $([a,b], \oplus, \odot)$ is called a *semiring*.

We shall use the following important cases (pairs):

$$\begin{aligned} \alpha \oplus \beta &= \min(\alpha, \beta), & \alpha \odot \beta &= \max(\alpha, \beta), \\ \alpha \oplus \beta &= \max(\alpha, \beta), & \alpha \odot \beta &= \min(\alpha, \beta), \\ \alpha \oplus \beta &= \min(\alpha, \beta), & \alpha \odot \beta &= \alpha + \beta, \\ \alpha \oplus \beta &= \max(\alpha, \beta), & \alpha \odot \beta &= \alpha + \beta. \end{aligned}$$

We translate the previous operations pointwise on functions.

We use the following notations:

$$\begin{aligned} \mathbf{u_1} = (u_1(\mathbf{x},t), v_1(\mathbf{x},t), t), \mathbf{u_2} = (u_2(\mathbf{x},t), v_2(\mathbf{x},t), t), \\ \mathbf{s_i} = (u_i(\mathbf{x},t), v_i(\mathbf{x},t), p_i(\mathbf{x},t)), \ i = 1,2, \end{aligned}$$

and specially for $p_1 = p_2 = p$ we take

$$\mathbf{s}_{i,p} = (u_i(\mathbf{x},t), v_i(\mathbf{x},t), p(\mathbf{x},t)), \ i = 1,2.$$

Given two triplets of solutions $\mathbf{s}_1$ and $\mathbf{s}_2$, we take

$$\mathbf{min}(\mathbf{s}_1, \mathbf{s}_2) := \Big(\min(u_1, u_2), \min(v_1, v_2), \min(p_1, p_2) \Big) \tag{10}$$

and

$$\mathbf{max}(\mathbf{s}_1, \mathbf{s}_2) := \Big(\max(u_1, u_2), \max(v_1, v_2), \max(p_1, p_2) \Big). \tag{11}$$

5.2 Superposition principle for the Euler equations

In this section prove the *pseudo-linear superposition principle* for the Euler equations.

Lemma 5.1 *Let* $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2$ *be two solutions of (7), (8), (5), such that both* $u_i, i = 1, 2$, *are either elements of* $\mathcal{U}_{ni}$ *or elements of* $\mathcal{U}_{nd}, i = 1, 2$.

Then the function

$$\mathbf{s}_{1,p} \oplus \mathbf{s}_{2,p} = \mathbf{min}(\mathbf{s}_{1,p}, \mathbf{s}_{2,p}),$$

where $\mathbf{min}(\mathbf{s}_{1,p}, \mathbf{s}_{2,p})$ *is defined by (10), is again solution of (7), (8), (5).*

Proof. We consider two solutions $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2$, of (7), (8), (5) . First we shall show that $\mathbf{s}_{1,p} \oplus \mathbf{s}_{2,p}$ satisfies (4), which is written in the form of the projection (7) and (8). So we shall prove that $\mathbf{s}_{1,p} \oplus \mathbf{s}_{2,p}$ satisfies (7). Using the notation $\mathbf{u}_1 = (u_1, v_1)$ and $\mathbf{u}_2 = (u_2, v_2)$, where $\mathbf{s}_{1,p} = (u_1, v_1)$ and $\mathbf{s}_{2,p} = (u_2, v_2)$ we have

$$\begin{aligned}
& \partial_t(u_1 \oplus u_2) + (u_1 \oplus u_2)\partial_x(u_1 \oplus u_2) + (v_1 \oplus v_2)\partial_y(u_1 \oplus u_2) + \partial_x(p \oplus p) \\
= \;& \partial_t(\min(u_1, u_2)) + (\min(u_1, u_2))\partial_x(\min(u_1, u_2)) \\
& +(\min(v_1, v_2))\partial_y(\min(u_1, u_2)) + \partial_x p \\
= \;& \begin{cases} \partial_t u_1 + u_1\partial_x u_1 + v_1\partial_y u_1 + \partial_x p & \text{as } (u_1, u_2) \in \mathcal{U}_{ni} \\ \partial_t u_2 + u_2\partial_x u_2 + v_2\partial_y u_2 + \partial_x p & \text{as } (u_1, u_2) \in \mathcal{U}_{nd} \end{cases} \\
= \;& 0 \,,
\end{aligned}$$

since by Proposition 4.2 we have for $i, j \in \{1, 2\}$ that $(u_1, u_2) \in \mathcal{U}_{ni}$, implies $v_i(x, y, t) \leq v_j(x, y, t)$, and $(u_1, u_2) \in \mathcal{U}_{nd}$ implies $v_i(x, y, t) \geq v_j(x, y, t)$. This means that $\mathbf{s}_{1,p} \oplus \mathbf{s}_{2,p}$ satisfies the equation (7).

In an analogous way we shall prove that $\mathbf{s}_{1,p} \oplus \mathbf{s}_{2,p}$ is solution of (8). Namely,

$$\partial_t(v_1 \oplus v_2) + (u_1 \oplus u_2)\partial_x(v_1 \oplus v_2) + (v_1 \oplus v_2)\partial_y(v_1 \oplus v_2) + \partial_x(p \oplus p)$$

$$= \begin{cases} \partial_t v_1 + u_1\partial_x v_1 + v_1\partial_y v_1 + \partial_x p & \text{as } (u_1, u_2) \in \mathcal{U}_{ni} \\ \partial_t u_2 + u_2\partial_x v_2 + v_2\partial_y u_2 + \partial_x p & \text{as } (u_1, u_2) \in \mathcal{U}_{nd} \end{cases}$$

$$= 0 .$$

Now we shall show that $\mathbf{s}_{1,p} \oplus \mathbf{s}_{2,p}$ is a solution of the equation (5). In fact

$$\begin{aligned} \text{div } (\mathbf{u_1} \oplus \mathbf{u_2}) &= \partial_x(\min(u_1, u_2)) + \partial_y(\min(v_1, v_2)) \\ &= \begin{cases} \partial_x u_1 + \partial_y v_1 & \text{as } (u_1, u_2) \in \mathcal{U}_{ni} \\ \partial_x u_2 + \partial_y v_2 & \text{as } (u_1, u_2) \in \mathcal{U}_{nd} \end{cases} \\ &= 0. \end{aligned}$$

So we have proved that $\mathbf{s}_{1,p} \oplus \mathbf{s}_{2,p}$ is solution of the system (7), (8) and (5). □

Lemma 5.2 *Under the same suppositions as in Lemma 5.1, we have that the function*

$$\mathbf{s}_{1,p} \oplus \mathbf{s}_{2,p} = \mathbf{max}(\mathbf{s}_{1,p}, \mathbf{s}_{2,p}),$$

defined by (11), is again a solution of the equations (7), (8) and (5).

As an immediate consequence of the previous Lemmas 5.1 and 5.2 we get the following theorems.

Theorem 5.3 *Let $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2,$ be two solutions of (7), (8), (5) such that (u_1, u_2) are elements either of $\mathcal{U}_{ni}$ or of $\mathcal{U}_{nd}$, and a_1, a_2 two real numbers. Then the pseudo-linear combination*

$$(a_1 \odot \mathbf{s}_{1,p}) \oplus (a_2 \odot \mathbf{s}_{2,p}) = \mathbf{min}(\mathbf{max}(a_1, \mathbf{s}_{1,p}), \mathbf{max}(a_2, \mathbf{s}_{2,p}))$$

with $\oplus, \odot$ given by (10) and (11), respectively, is again a solution of (7), (8) and (5) .

Theorem 5.4 *Let* $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2,$ *be two solutions of (7) ,(8) , (5) such that* (u_1, u_2) *are elements either of* $\mathcal{U}_{ni}$ *or of* $\mathcal{U}_{nd}$ *and* a_1, a_2 *two real numbers. Then the pseudo-linear combination*

$$(a_1 \odot \mathbf{s}_{1,p}) \oplus (a_2 \odot \mathbf{s}_{2,p}) = \mathbf{max}(\mathbf{min}(a_1, \mathbf{s}_{1,p}), \mathbf{min}(a_2, \mathbf{s}_{2,p}))$$

with $\oplus, \odot$ *given by (11) and (10), respectively, is again a solution of (7), (8) and (5) .*

We obtain, with an additional condition, the pseudo-linear superposition principle for another pair of pseudo-operations.

Theorem 5.5 *Let* $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2,$ *be two solutions of (7), (8) and (5) such that* (u_1, u_2) *are elements either of* $\mathcal{U}_{ni}$ *or of* $\mathcal{U}_{nd}$*. If* $(u_i, v_i), i = 1, 2$ *satisfy the condition*

$$\partial_y u_i = \partial_y v_i \qquad i = 1, 2. \tag{12}$$

then the pseudo-linear combination for two real numbers a_1, a_2

$$(a_1 \odot \mathbf{s}_{1,p}) \oplus (a_2 \odot \mathbf{s}_{2,p}),$$

where $\oplus$ *is given by (10) and* $\odot$ *is defined by*

$$\lambda \odot \mathbf{s} = \lambda \odot (u, v, p) = (\lambda + u, \lambda + v, \lambda + p), \tag{13}$$

is again a solution of (7), (8) and (5).

Proof. First, by Lemma 5.1 $\mathbf{min}(\mathbf{s}_{1,p}, \mathbf{s}_{2,p})$ is a solution of (7) and (8).

Now, it is easy to see that the trivial solution given by three constants $\lambda_i, i = 1, 2, 3$: $\mathbf{s}_c = (\lambda_1, \lambda_2, \lambda_3)$ is again a solution of (7),(8) and (5). We shall prove that for any real number λ , $\lambda \odot \mathbf{s}$ is a solution of (7). In fact,

$$\begin{aligned} & \partial_t(\lambda + u) + (\lambda + u)\ \partial_x(\lambda + u) + (\lambda + v)\ \partial_y(\lambda + u) + \partial_x(\lambda + p) \\ = \ & \partial_t u + (\lambda + u)\partial_x u + (\lambda + v)\partial_y u + \partial_x p \\ = \ & \partial_t u + u\partial_x u + v\partial_y u + \partial_x p + \lambda(\partial_x u + \partial_y u) \\ = \ & \lambda(\partial_x u + \partial_y u) = 0, \end{aligned}$$

where we have used the condition (12), which with (5) for $\mathbf{u}$, i.e., $\partial_x u + \partial_y v = 0$, implies

$$\partial_x u + \partial_y u = 0.$$

So, we have proved that $\lambda \odot \mathbf{s}$ is a solution of (7). In a analogous way we prove that it satisfies (8) and (5). □

In an analogous way we obtain the following theorem.

Theorem 5.6 *Let $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2$, be two solutions of (7), (8) and (5) such that (u_1, u_2) are elements either of $\mathcal{U}_{ni}$ or of $\mathcal{U}_{nd}$. If $(u_i, v_i), i = 1, 2$ satisfy the condition (12), then the pseudo-linear combination for two real numbers a_1, a_2*

$$(a_1 \odot \mathbf{s}_{1,p}) \oplus (a_2 \odot \mathbf{s}_{2,p}),$$

where $\oplus$ is given by (11) and $\odot$ is defined by (13) is again a solution of (7), (8) and (5). □

5.3 Superposition principle for Navier-Stokes equations

In this section we prove the *pseudo-linear superposition principle* to Navier-Stokes equations. We consider an incompressible homogeneous viscous flow: that means that div $\mathbf{u} = 0$, for the density $\rho = 1$, ν is the coefficient of viscosity, for the forces $\mathrm{f} = 0$. The equations of motion of this flow are the *Navier-Stokes equations*:

$$\rho \frac{D\mathbf{u}}{Dt} = - \text{ grad p} - \nu \Delta \mathbf{u} \tag{14}$$

$$\text{div } \mathbf{u} = 0$$

$$\mathbf{u} = 0 \quad \text{on } \partial D,$$

where $\Delta \mathbf{u}$ is the Laplacian of the velocity u, defined in this way:

$$\Delta \mathbf{u} = (\partial_{xx} + \partial_{yy})\mathbf{u} = (\partial_{xx} u + \partial_{yy} v),$$

as $\mathbf{u}(\mathbf{x}, t) = (u(x, y, t), v(x, y, t))$.

We consider two-dimensional incompressible flow in the upper half plane $y > 0$; so the projections of the Navier-Stokes equations (14) on axes x and y are the following:

$$\partial_t u + u\partial_x u + v\partial_y u + \partial_x p + \nu(\partial_{xx} u + \partial_{yy} u) = 0 \tag{15}$$

$$\partial_t v + u\partial_x v + v\partial_y v + \partial_y p + \nu(\partial_{xx} v + \partial_{yy} v) = 0 \tag{16}$$

$$\partial_x u + \partial_y v = 0 \tag{17}$$

$$u = v = 0 \qquad \text{on} \qquad \partial D. \tag{18}$$

In analogous way as in section 5.2 of the Euler equations, we obtain the following theorems.

Theorem 5.7 *Let $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2$, be two solutions of (15) - (18) and a_1, a_2 two real numbers, such that (u_1, u_2) are elements either of $\mathcal{U}_{ni}$ or of $\mathcal{U}_{nd}$. Then the pseudo-linear combination*

$$(a_1 \odot \mathbf{s}_{1,p}) \oplus (a_2 \odot \mathbf{s}_{2,p}) = \mathbf{min}(\mathbf{max}(a_1, \mathbf{s}_{1,p)}, \mathbf{max}(a_2, \mathbf{s}_{2,p}))$$

with $\oplus, \odot$ given by (10) and (11), respectively, is again a solution of (15) - (18). □

Theorem 5.8 *Let $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2$, be two solutions of (15) - (18) and a_1, a_2 two real numbers, such that (u_1, u_2) are elements either of $\mathcal{U}_{ni}$ or of $\mathcal{U}_{nd}$. Then the pseudo-linear combination*

$$(a_1 \odot \mathbf{s}_{1,p}) \oplus (a_2 \odot \mathbf{s}_{2,p}) = \mathbf{max}(\mathbf{min}(a_1, \mathbf{s}_{1,p)}, \mathbf{min}(a_2, \mathbf{s}_{2,p}))$$

with $\oplus, \odot$ given by (10) and (11), respectively, is again a solution of (15) - (18). □

Theorem 5.9 *Let $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2$, be two solutions of (15) - (18) such that (u_1, u_2) are elements either of $\mathcal{U}_{ni}$ or of $\mathcal{U}_{nd}$. which satisfy the condition (12). Then the pseudo-linear combination for two real numbers a_1, a_2*

$$(a_1 \odot \mathbf{s}_{1,p}) \oplus (a_2 \odot \mathbf{s}_{2,p}),$$

where $\oplus$ and $\odot$ are given by (10) and (13), respectively, is again a solution of (15) - (18). □

Theorem 5.10 *Let* $\mathbf{s}_{i,p} = (u_i, v_i, p), i = 1, 2,$ *be two solutions of (15) - (18) such that* (u_1, u_2) *are elements either of* $\mathcal{U}_{ni}$ *or of* $\mathcal{U}_{nd}$*. If the solutions satisfy the conditions (12) , then the pseudo-linear combination for two real numbers* a_1, a_2

$$(a_1 \odot \mathbf{s}_{1,p}) \oplus (a_2 \odot \mathbf{s}_{2,p}),$$

where $\oplus$ *and* $\odot$ *are given by (11) and (13), respectively, is again a solution of (15) - (18).* □

6 Superposition principle for Stokes equations

We know ([2]) that the *Stokes equations* are approximate equations for incompressible flow:

$$\partial_t \mathbf{u} + \text{ grad p} + \nu \Delta \mathbf{u} = 0$$

$$\text{div } \mathbf{u} = 0.$$

The projections on axes x and y of the equations above are:

$$\partial_t u + \partial_x p + \nu(\partial_{xx} u + \partial_{yy} u) = 0 \tag{19}$$

$$\partial_t v + \partial_y p + \nu(\partial_{xx} v + \partial_{yy} v) = 0. \tag{20}$$

$$\partial_x u \ + \ \partial_y v \ = \ 0. \tag{21}$$

In this section we prove the *pseudo-linear superposition principle* for Stokes equations: we shall consider the application to solutions of (19)-(21), which depend only of time t.

Theorem 6.1 *Let* $\mathbf{s}_i(t) = (u_i(t), v_i(t), p_i(t))$*,* $i = 1, 2$ *be solutions of (19)-(21). Then the pseudo-linear combination for two real numbers* a_1, a_2

$$(a_1 \odot \mathbf{s}_1) \oplus (a_2 \odot \mathbf{s}_2),$$

where $\oplus$ *and* $\odot$ *are given with generator* g *defined by*

$$g(a) = e^{-\ c\ a},\ c > 0, \quad \textit{and then}\ \ g^{-1}(b) = -\frac{1}{c}\ \log\ b,$$

$$\mathbf{s}_1 \oplus \mathbf{s}_2 = (g^{-1}(g(u_1) + g(u_2)), g^{-1}(g(v_1) + (g(v_2)), g^{-1}(g(p_1) + g(p_2))),$$

and

$$\begin{aligned} a \odot \mathbf{s} &= (g^{-1}(g(a) \cdot g(u)), g^{-1}(g(a) \cdot g(v)), (g^{-1}(g(a) \cdot g(p))) \\ &= (a + u, a + v, a + p) \end{aligned}$$

is again solution of (19)-(21).

Proof. Let $\mathbf{s}_i(t) = (u_i(t), v_i(t), p_i(t))$ be solutions of the equation (19)-(21).

First we shall prove that $(u_1 \oplus u_2, p_1 \oplus p_2)$ is solution of (19), i.e.,

$$\partial_t(u_1 \oplus u_2) + \partial_x(p_1 \oplus p_2) + \nu(\partial_{xx}(u_1 \oplus u_2) + \partial_{yy}(u_1 \oplus u_2)) = 0. \tag{22}$$

Put

$$U = e^{-cu_1} + e^{-cu_2}\ ,\ P = e^{-cp_1} + e^{-cp_2}\ , \tag{23}$$

we have

$$\partial_t(u_1 \oplus u_2) = \frac{\partial_t u_1 e^{-cu_1}}{U} + \frac{\partial_t u_2 e^{-cu_2}}{U}, \partial_x(p_1 \oplus p_2) = \frac{\partial_x p_1 e^{-cp_1}}{P} + \frac{\partial_x p_2 e^{-cp_2}}{P}; \tag{24}$$

moreover

$$\partial_{xx}(u_1 \oplus u_2)$$

$$= \frac{1}{U^2}\left(\partial_{xx}u_1 e^{-cu_1}U\ +\ \partial_{xx}u_2 e^{-cu_2}U\ -\ c(\partial_x u_1 - \partial_x u_2)^2 e^{-c(u_1+u_2)}\right), \tag{25}$$

$$\partial_{yy}(u_1 \oplus u_2)$$

$$= \frac{1}{U^2}\left(\partial_{yy}u_1 e^{-cu_1}U\ +\ \partial_{yy}u_2 e^{-cu_2}U\ -\ c(\partial_y u_1 - \partial_y u_2)^2 e^{-c(u_1+u_2)}\right). \tag{26}$$

Therefore, the left side of the equation (22) is the following:

$$\frac{\partial_t u_1 e^{-cu_1}}{U}\ +\ \frac{\partial_t u_2 e^{-cu_2}}{U}\ +\ \frac{\partial_x p_1 e^{-cp_1}}{P}\ +\ \frac{\partial_x p_2 e^{-cp_2}}{P} + \tag{27}$$

$$\frac{\nu}{U^2}\left(\partial_{xx}u_1e^{-cu_1}U \;+\; \partial_{xx}u_2e^{-cu_2}U \;-\; c(\partial_x u_1-\partial_x u_2)^2e^{-c(u_1+u_2)}\right)+$$

$$\frac{\nu}{U^2}\left(\partial_{yy}u_1e^{-cu_1}U \;+\; \partial_{yy}u_2e^{-cu_2}U \;-\; c(\partial_y u_1-\partial_y u_2)^2e^{-c(u_1+u_2)}\right).$$

Moreover, in (27) we sum the terms which contain the function u_1 and its derivatives:

$$\frac{\partial_t u_1e^{-cu_1}}{U} \;+\; \frac{\partial_x p_1e^{-cp_1}}{P} \;+\; \nu\left(\frac{1}{U^2}\left(\partial_{xx}u_1 \;+\; \partial_{yy}u_1\right)e^{-cu_1}U\right), \tag{28}$$

the same for the function u_2

$$\frac{\partial_t u_2e^{-cu_2}}{U} \;+\; \frac{\partial_x p_2e^{-cp_2}}{P} \;+\; \nu\left(\frac{1}{U^2}\left(\partial_{xx}u_2 \;+\; \partial_{yy}u_2\right)e^{-cu_2}U\right). \tag{29}$$

Setting: $E_{ij}=e^{-c(u_i+p_j)}, i,j=1,2$, we have $e^{-cu_1}P=E_{11}+E_{12},\ e^{-cp_1}U=E_{11}+E_{21}$ $(E_{12}\neq E_{21})$ and then (28) =

$$\frac{1}{U\ P}\left(\partial_t u_1e^{-cu_1}P+\partial_x p_1e^{-cp_1}U+\nu P\left(\left(\partial_{xx}u_1+\partial_{yy}u_1\right)e^{-cu_1}\right)\right), \tag{30}$$

from which

$$(30)=\frac{1}{U\ P}\left(\left(\partial_t u_1+p_{1x}+\nu\left(\partial_{xx}u_1+\partial_{yy}u_1\right)\right)E_{11}\right)\;+$$

$$\frac{1}{U\ P}\left(\left(\partial_t u_1+\nu\left(\partial_{xx}u_1+\partial_{yy}u_1\right)\right)E_{12}+\partial_x p_1E_{21}\right); \tag{31}$$

similarly (29) =

$$\frac{1}{U\ P}\left(\partial_t u_2e^{-cu_2}P+\partial_x p_2e^{-cp_2}U+\nu P\left(\left(\partial_{xx}u_2+\partial_{yy}u_2\right)e^{-cu_2}\right)\right) \tag{32}$$

from which

$$(32)=\frac{1}{U\ P}\left(\left(\partial_t u_2+\partial_x p_2+\nu\left(\partial_{xx}u_2+\partial_{yy}u_2\right)\right)E_{22}\right)\;+$$

$$\frac{1}{U\ P}\left(\left(\partial_t u_2+\nu\left(\partial_{xx}u_2+\partial_{yy}u_2\right)\right)E_{21}+\partial_x p_2E_{12}\right). \tag{33}$$

First, in (31) and (33) the coefficients of E_{11} and E_{22} are zero as u_1 and u_2 are solutions of (19). Now we sum the other terms :

$$(31)+(33)=\frac{1}{U\ P}\Big(\ (\partial_t u_1+\nu\,(\partial_{xx}u_1+\partial_{yy}u_1))\,E_{12}\ +\ \partial_x p_1 E_{21}\Big)+$$

$$\frac{1}{U\ P}\Big(\ (\partial_t u_2+\nu\,(\partial_{xx}u_2+\partial_{yy}u_2))\,E_{21}\ +\ \partial_x p_2 E_{12}\Big). \tag{34}$$

As $u_i, i=1,2$ are solutions of (19), we get

$$(34)=\frac{1}{U\ P}\Big(\ -\partial_x p_1 E_{12}+\partial_x p_1 E_{21}-\partial_x p_2 E_{21}+\partial_x p_2 E_{12}\Big)\ =$$

$$\frac{1}{U\ P}\Big(\partial_x(p_1-p_2)(E_{21}-E_{12})\Big)\ =0,$$

since by the supposition the functions p_i depends only on time, $\partial_x(p_1-p_2)=\partial_x p_1-\partial_x p_2=0$. In (27) it remains:

$$\frac{\nu}{U^2}\Big(\ -c(\partial_x u_1-\partial_x u_2)^2 e^{-c(u_1+u_2)}-\ c(\partial_y u_1-\partial_y u_2)^2 e^{-c(u_1+u_2)}\Big)\ =$$

$$\frac{-c\nu}{U^2}\Big((\partial_x u_1-\partial_x u_2)^2+(\partial_y u_1-\partial_y u_2)^2\Big)\,e^{-c(u_1+u_2)}=0,$$

since by the supposition the functions u_i depends only on time, $\partial_x(u_1-u_2)=\partial_x u_1-\partial_x u_2=0$. So, we have shown that $(u_1\oplus u_2, p_1\oplus p_2)$ is a solution of the equation (19).

Changing u_i with v_i, $i=1,2$ in the previous proof we can prove that $(v_1\oplus v_2, p_1\oplus p_2)$ is solution of the equation (20), and then $\mathbf{s}_1\oplus\mathbf{s}_2$ is a solution of (19).

Now we shall prove that $\mathbf{u}_1\oplus\mathbf{u}_2$ is a solution of the equation (21). In fact, from (23) and (24), we get

$$\mathrm{div}(\mathbf{u}_1\oplus\mathbf{u_2})=\partial_x(u_1\oplus u_2)+\partial_y(v_1\oplus v_2)=$$

$$\frac{\partial_x u_1 e^{-cu_1}}{U}+\frac{\partial_x u_2 e^{-cu_2}}{U}+\frac{\partial_y v_1 e^{-cu_1}}{U}+\frac{\partial_y v_2 e^{-cu_2}}{U}=0.$$

As regards the product $\odot$, we note that $\partial_x(a+u)=\partial_x u$, and so on, so also $a\odot u$ is solution of (19) and (21). □

7 Conclusion

In this paper it was proven the pseudo-linear superposition principle for the Euler, Navier-Stokes and Stokes equations. In order to achieve this principle for the first two equations we used the monotonicity of the velocity.

The obtained results will serve in the future for different applications, e.g., [23], and as a base for the construction of the general weak solutions as in [8, 11, 14, 17], which are in a wider class than previously considered class of monotone functions, and allow movement also in harder structures than fluid, see [9].

8 Acknowledgment

The first author was partially supported by the project MPNS-174009 and "Mathematical models of intelligent systems and their applications" by Provincial Secretariat for Science and Technological Development of Vojvodina. The second author thanks Basileus project (Erasmus Mundus) for academic years 2009-2010, 2010-2011 and the University of Novi Sad for the invitation and the kind hospitality.

References

[1] L. Caffarelli, R. Kohn, L. Nirenberg: Partial regularity of suitable weak solutions of Navier-Stokes equations, Comm. Pure and Appl. Math. 35, 771-831, 1982.

[2] A.J. Chorin, J. Marsen, A Mathamatical Intoduction to Fluid Mechanics, Springer-Verlag, 1993.

[3] P. Constantin: Open problems and Research Directions in Mathematical Study of Fluid Dynamics , in Mathematicacs Unlimited - 2001 and Beyond (Eds. B. Engquist, W. Schmid), Springer, 2001, 353-360.

[4] R. Deutray, J-L. Lions: Mathematical Analysis and Numerical Methods for Science and Tecnology, Vol.4,6, Springer-Verlag, 2000.

[5] S. Friedlander, D. Serre: Handbook of Mathematical Fluid Dynamics, Vol.1, Elsevier, 2002.

[6] S. Friedlander, D. Serre: Handbook of Mathematical Fluid Dynamics, vol.2, Elsevier, 2003.

[7] M. Grabisch, J. L. Marichal, R. Mesiar, E. Pap: Monograph: Aggregation Functions, Acta Polytechnica Hungarica 6,1 (2009), 79-94.

[8] V. Kolokoltsov, P. Maslov: Idempotent Analysis and Its Applications, Kluwer Academic Publishers, Dordrecht, 1997.

[9] M. Kuffova, P. Nečas: Fracture Mechanics Prevention: Comprehensive Approach-based Modelling?, Acta Polytechnica Hungarica 7, 5, 2010, 5-17.

[10] O. A. Ladyzhenskaya, The Mathematical Theory of Viscous Incompressible Flows, (2nd ed.), Gordon and Breach, 1969.

[11] V.P. Maslov, S. N. Samborskij (Eds.): Idempotent Analysis, Advances in Soviet Mathematics 13, Amer. Math. Soc., Providence, Rhode Island, 1992.

[12] E. Pap, Integral generated by decomposable measure, Univ. u Novom Sadu Zb. Rad. Prirod.-Mat. Fak. Ser. Mat. 20,1, 135-144, 1990.

[13] E. Pap: g-calculus, Univ. u Novom Sadu Zb. Rad. Prirod.-Mat. Fak. Ser.Mat., 23,1, 1993, 145-150.

[14] E. Pap: Null-Additive Set Functions, Dordrecht-Boston-London, Kluwer Academic Publishers, 1995.

[15] E. Pap: Decomposable measures and nonlinear equations, Fuzzy Sets and Systems, 92, 1997, 205-222.

[16] E. Pap: Solving nonlinear equations by non-additive measures, Nonlinear Analysis, 30, 1997, 31-40.

[17] E. Pap: Applications of decomposable measures, in Handbook Mathematics of Fuzzy Sets-Logic, Topology and Measure Theory (Eds. U. Höhle, S.R. Rodabaugh), Kluwer Academic Publishers, 1999, 675-700.

[18] E. Pap: Pseudo-Additive Measures and Their Applications, in Handbook of Measure Theory,Chap.35, (Ed. E. Pap), Elsevier, North-Holland, 2002, 1403-1468.

[19] E. Pap: Applications of the generated pseudo-analysis on nonlinear partial differential equations, Proceedings of the Conference on Idempotent Mathematics and Mathematical Physics (Eds. G. L. Litvinov, V. P. Maslov), Contemporary Mathematics 377, American Mathematical Society, 2005, 239-259.

[20] E. Pap: Applications of pseudo-analysis on models with nonlinear partial differential equations, (Eds. E. Pap, J. Fodor) Proceedings of the 5th International Symposium on Intelligent Systems and Informatics, Subotica, 2007, 7-12.

[21] E. Pap, N. Ralević: Pseudo-Laplace Transform, Nonlinear Analysis 33, 1998, 533-550.

[22] E. Pap, D. Vivona: Non-commutative and non-associative pseudo-analysis and its applications on nonlinear partial differential equations, J. Math. Anal. Appl. 246, 2000, 390-408.

[23] R. M. Patel, G. Deheri, H. C. Patel: Effect of Surface Roughness on the Behavior of a Magnetic Fluid-based Squeeze Film between Circular Plates with Porous Matrix of Variable Thickness, Acta Polytechnica Hungarica 8, 5, 2011, 145-164.

[24] A. Shnirelman: On the nonuniqueness of weak solutions of the Euler equation, Comm.Pure and Appl.Math. 50, 1997, 1260-1286.

[25] M. Sammartino, R.E. Cafflisch: Zero Viscosity Limit for Analitic Solutions of the Navier-Stokes Equation on a Half-Space. I.Existence for Euler and Prandtl Equations, Commun. Math. Phys. 192, 1998, 433-461.

[26] M. Sammartino, R. E. Cafflisch: Zero Viscosity Limit for Analitic Solutions of the Navier-Stokes Equation on a Half-Space. II.Construction of the Navier-Stokes Solution, Commun. Math. Phys. 192, 1998, 465-491.

A Template-based Model Transformation Approach for Deriving Multi-Tenant SaaS Applications

Kun Ma[1], **Bo Yang**[2], **Ajith Abraham**[3, 4]

[1] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, 250022 Jinan, China
e-mail: ise_mak@ujn.edu.cn

[2] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, 250022 Jinan, China
E-mail: yangbo@ujn.edu.cn

[3] Machine Intelligence Research Labs, Scientific Network for Innovation and Research Excellence, 98071 Auburn, USA
e-mail: ajith.abraham@ieee.org

[4] IT For Innovations, VSB - Technical University of Ostrava, Ostrava - Poruba, Czech Republic

Abstract: Software-as-a-Service (SaaS) and Model-Driven Engineering (MDE) are two of the most dominant software engineering paradigms nowadays. Multi-tenancy is the key to successful SaaS. In this paper, we introduce a data middleware to customize the multi-tenant database first. In addition, with the help of model transformation, it is possible to generate SaaS applications from the models. However, most of the current model transformation approaches do not fully support the requirements for model synchronization, and they do not cater for the specific problems faced in the multi-tenancy. Therefore, an effective and simple template-based model transformation and model synchronization approach based on model evolution of MDE paradigms is fully integrated for the development of SaaS multi-tenant applications. The proposed framework uses a novel extensible business component model (xBC) to sufficiently describe both the structural and behavioral properties of SaaS applications. The distribution and uninterrupted running of the generated SaaS applications proves that our approach is feasible and correct in practice.

Keywords: Software-as-a-Service; multi-tenancy; textual template evolution; model transformation; model synchronization

1 Introduction

Model-Driven Engineering (MDE) is becoming the dominant software engineering paradigm to specify, develop and maintain software systems, mainly because it can raise the level of abstraction and automation in software construction [1]. Some findings on experiences from using model-based development in industry from the EA-MDE project indicate that 83% of our questionnaire respondents think that MDE improved productivity and maintainability [2]. The use of MDE has the following consequences for a software development process [3]: 1) More time can be devoted to analyzing the business; 2) The time needed to perform coding tasks is reduced; 3) Productivity is improved as the time necessary for coding is reduced.

Software-as-a-Service (SaaS) is a software delivery on-demand model in which software and its associated data are hosted centrally in the cloud. According to International Data Corporation's (IDC) latest market report, SaaS will grow at a 26.4 percent compound annual growth rate (CAGR) through 2015[4]. As SaaS of the cloud infrastructures is the future tendency of the IT industry, it is urgent to research on the generation approach of SaaS applications. A recent survey of organizations with experience using cloud applications and platforms reveals that the most urgent need is how to tightly integrate it with other applications and how to convert the legacy systems into SaaS applications [5].

Therefore, it is natural that we wonder how both paradigms, MDE and SaaS, can be integrated and benefit from each other. Bruneliere et al. discuss two different collaboration scenarios between MDE and SaaS [6]: 1) MDE for the cloud refers to the use of MDE techniques to facilitate and (semi)automate the development of SaaS applications. 2) MDE in the cloud involves using cloud infrastructure to enable MDE in new and novel ways, corresponding to on-demand Modeling as a Service (MaaS) initiative. Similar to SaaS, MaaS would allow the deployment and on-demand execution of modeling and model-driven services over the Internet. In accordance with scenario 1, we aim to identify opportunities for using MDE to support the development of cloud-based SaaS multi-tenant applications. Therefore, this paper proposes a transparent SaaS multi-tenant data middleware, which is fully integrated with template-based model transformation approach and model synchronization based on model evolution of MDE paradigms for the development of SaaS multi-tenant applications.

The rest of the paper is organized as follows. Section 2 discusses the background and related work. In Section 3, an extensible business component model (*xBC*) is presented to describe SaaS multi-tenant business and database to the fullest. The architecture of multi-tenant data middleware and *xBC* is discussed in detail. In Section 4, a template-based model transformation approach that supports model synchronization is presented to generate the SaaS application. Conclusions are provided in the last Section.

2 Related Works

The major problem of SaaS modeling lies in the customization of the data and business.

2.1 Multi-tenant Data Model

An important requirement for SaaS applications is the support of multiple tenants [7]. Data architecture is an area in which the optimal degree of isolation for a SaaS application can vary significantly depending on technical and business considerations. An overview of approaches for data management in a multi-tenant deployment can be found in [8] and [9]. The paper categorizes existing approaches of shared applications and briefly explains them in Figure 1, each of which lies at a different location in the continuum between isolation and sharing. 1) **Separate schema with shared application** involves housing multiple tenants in the same database, with each tenant having its own set of tables, which are grouped into a schema created specifically for the tenant. Unfortunately, this approach tends to lead to higher costs for maintaining equipment, backing up tenant data and restoring data in the event of a failure. The number of tenants that can be housed on a given database server is limited by the number of schemas that the server can support. 2) **Shared schema with shared application** involves using the same database and the same set of tables to host multiple tenants' data. A given table can include records from multiple tenants stored in any order; a Tenant ID column associates every record with the appropriate tenant. The shared schema approach has the lowest hardware and backup costs. However, this approach may incur additional development effort in the area of security, to ensure that tenants can never access other tenants' data. Compared with the two approaches, we lead to improvements in the relational database, and propose multi-tenant data middleware.

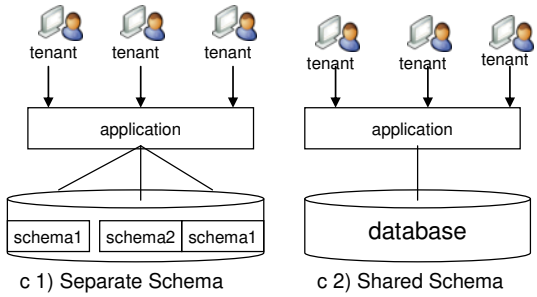


Figure 1
The common SaaS multi-tenant data models

2.2 SaaS Application Modeling

Several researchers have proposed using variability modeling techniques from software product line engineering in the context of service-based systems. Chang et al. address the problem that the variability of business processes and services is not explicitly modeled, which hinders implementing adaptive service-based systems [10]. They extend the XML schemas of service description languages in order to cater for variability. Liu et al. propose a new modeling method for constructing SaaS Service using extended Web Services Conversation Language (WSCL) [11]. Wang et al. propose a service community model based on eXtensible Markup Language (XML) for a bilateral SaaS mode which is abstracted from a real project of the nationwide service network for sharing science and technology information [12]. Although these approaches propose explicitly documenting variability, they do not cater for the specific problems faced in the SaaS context (e.g., multi-tenancy). Motivated by these problems, the extensible business component model (*xBC*), based on the extension of the Unified Modeling Language (UML) profiles, is abstracted from SaaS applications to describe the multi-tenant business to the utmost.

2.3 Model Transformation Approach in Support of Model Synchronization

Model transformations are essential in the process of MDE [13]. The development of a software system is an iterative process with frequent modifications to the involved models according to the user requirements [14]. As a consequence, an effective and simple model transformation methodology that supports model synchronization is needed urgently. However, most of the current model transformation approaches have some limits, such as fully incremental support for model synchronization. Additionally, the behavior of the SaaS application cannot be modeled in detail, in which case it is easier to write source code manually. This means that the mixture often leads to a wide range of inconsistencies [15]. Therefore, this information should be kept during the model transformation, and several possibilities exist to develop model transformations for the sake of model synchronization. An overview of model transformation and synchronization systems can be found in [16]. As outlined in the introduction, MDE requires a bidirectional solution which preserves model contents when synchronizing as much as possible. However, many available model transformation approaches only support classical one-way batch-oriented transformations [17]. This basic feature updating existing target models based on changes in the source models is also referred to as *change propagation* in the Query/View/Transformation (QVT) final adopted specification [17]. The QVT implementation [18] is only unidirectional but partly incremental. Other existing TGG-based approaches also do not provide a comparable automatic and computational incremental solution

(for a detailed discussion see [19] and [20]). Compared to these approaches, our approach of model transformation that supports model synchronization is based on model evolution. This approach of model synchronization will only take the storage space of model repositories rather than extra space. Only the models with changed version number need a subsequent model transformation. This method is named *source incrementality*, which is simple and useful for working with large scale source models. In this way, model synchronization is a special and partial model transformation. This is a good way to minimize the amount of source that needs to be reexamined by a transformation when the source is changed.

3 Extensible Business Component Model in Support of Multi-Tenancy

The important features of SaaS applications are multi-tenant data and business customization.

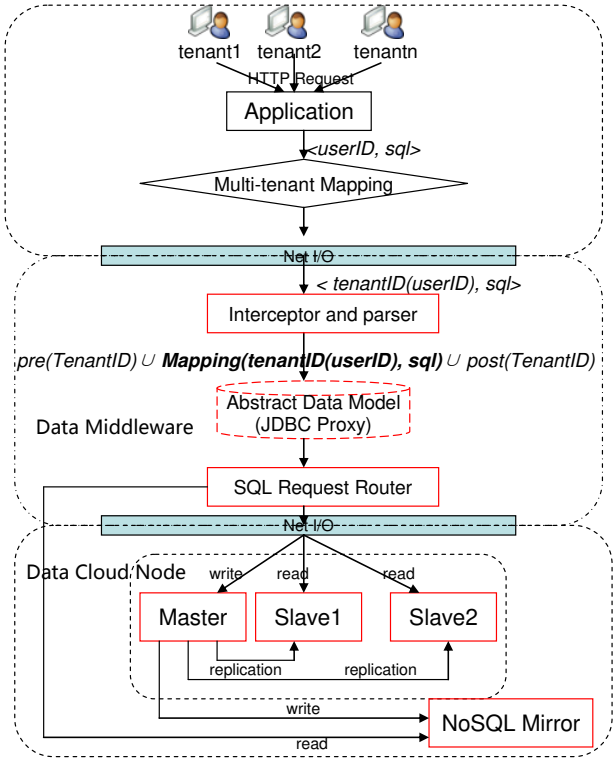


Figure 2
Multi-tenant Data Middleware

3.1 A Transparent Multi-Tenant Data Model

In this section, we propose a data middleware to customize the multi-tenant database. The architecture of data middleware shown in Figure 2 is comprised of the Abstract Data Model, the JDBC Interceptor and parser, the SQL Request Router and the Data Cloud Node.

3.1.1 Abstract Data Model

The Abstract Data Model acts as database JDBC proxy without the storage of any data for the sake of smooth transition. It is transparent to the application, owning the same set of tables and views as the physical database. Therefore, the application can connect to the abstract data model without any modification. All the application lifecycle management procedures (upgrade or patch) may remain as they are. The abstract data model provides the logical data isolation for the tenants with higher demand on security.

3.1.2 The JDBC Interceptor and Parser

The JDBC Interceptor and parser is used to intercept the SQL and formulate the new SQL to the Abstract Data Model. The new SQL is transformed from the original SQL and tenant information. The interceptor process is denoted as $sql(u) \rightarrow pre(TenantID) \cup Mapping(tenantID(userID), sql) \cup post(TenantID)$, where *pre(TenantID)* is the pre personalized operation, *post(TenantID)* means the post personalized operation, and *Mapping* is the transformation function. The current tenant account is added to the Request Session with some minor modifications.

3.1.3 The SQL Request Router

The SQL Request Router sends the SQL request to different nodes of data cloud on average. One of the more powerful features is the ability to do "Read/Write Splitting". The *read* request is assigned to the *slave* node, while the *write* request is assigned to the *master* node. Database replication enables data from the master to be replicated to one or more slaves.

Replication is based on the master server keeping track of all changes to its databases (change of structure, updates, deletes, and so on) in its binary log. The binary log serves as a written record of all events that modify the database structure or content (data) from the moment the server is started. Typically, SELECT statements are not recorded because they modify neither the database structure nor content. The binary log is the collection of SQL statements after the dump operation.

3.2 Tenant Expression

Expression is a dynamic value, which is substituted when running. Common types of expressions are constants, session variables, the return value of static function and the requested variables. The tenant expression is also provided to obtain the current tenant information from the context of Web request. The syntax of tenant expression is shown in Table 1, which is used in extensible business component models to describe the personalized business.

Table 1
Tenant expression

Name	Definition
$T{tenant.tenantID}	ID of current tenant
$T{tenant.userID}	ID of current user
$T{tenant.loginName}	Login name of current tenant

3.3 Extensible Business Component Model-supported Multi-Tenancy

A **model** is a 2-tuple: *model:=(name, attributes)*, where *attributes* is a set of properties of this model, denoted as *attributes={x|x Attribute}*. The property of a model is defined as *Attribute :=(name, type, default)*, which includes a lifetime identifier, its type and the default value. We use *m(s)/f* to denote a model *m* of the system *s* in the formalism *f*. The formalism of a model is usually called a **metamodel**. The instance of a model is called an object. ***Meta****(o,m)=true* means that model *o* is the instance of metamodel *m*.

Extensible business component model (*xBC*) is proposed to describe the SaaS business to the greatest extent. The metamodel of *xBC* is divided into three different layers, shown in Figure 3: business process, business object and business presentation. A separation of design concerns into distinct model layers has several advantages, such as ease of maintenance, orientation to the viewpoint, and the ability to select specialized tools and techniques for specific concerns.

3.3.1 The Business Process Model

The business process model describes the basic business logic of an SaaS application, including *create*, *read*, *update* and *delete* (CRUD) business, compound CRUD business and user defined special business. Generally, clicking the button or hyperlink of SaaS applications in the user interface triggers the specific business process. The input of the SaaS application is often a user's form. The submission of a Web form is always triggered by a button [21]. The derived models of business process contain database-related manipulation, Uniform Resource Locator (URL), code blocks and so on. Database-related manipulation is

a direct operation of the database, such as Structured Query Language (SQL) statements and stored procedure; URL means a navigation of a Web page, such as an HTML page and JSP. Not all the business behavior can be represented in models. Some business processes are easy to describe by the source codes. Therefore, we propose a novel derived code model named *CodeBlock* which uses *dependency injection* [22] and *method interception* [22] techniques to embed source codes into models.

Business logic *BP* is defined as the instance of metamodel *BusinessProcessLogic*, satisfying ***Meta****(BP, BusinessProcessLogic)=true*. *BP* is denoted as *BP := ((name, String, ""), {(parameters, String[0..*], null), (returntype, String, "")})*.

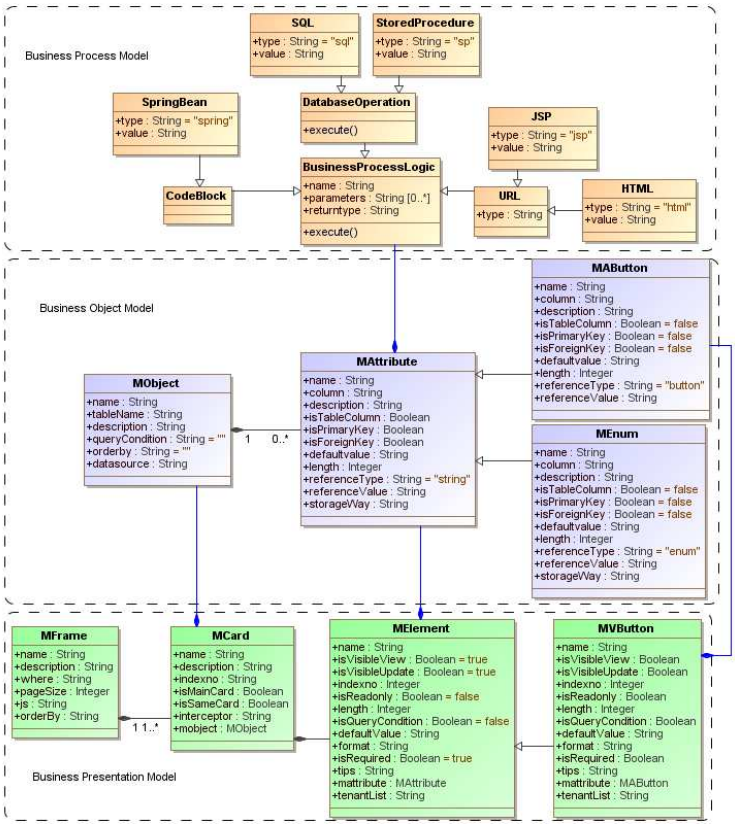


Figure 3
Metamodel of Extensible Business Component Model (xBC)

3.3.2 The Business Object Model

The business object model describes the organization of the business concepts managed by the SaaS application, which include *MObject*, *MAttribute*, *MAButton*, *Reference*, and so on. In order to refine the details of business objects, it is divided

into business object model *MObject* and the attribute model of the business object *MAttribute*. In the context of SaaS modeling, *MObject* defines the name, the description of a business object, table mappings (i.e. corresponding to the table of the relational database), and the query condition (i.e. the value range of business data represented by the instance of *MObject*). *MAttribute* describes the property of the business object, including the name, description, column (i.e. corresponding to the key of the table of the relational database), and so on. The most important property of *MAttribute* is the *reference* and *storage way*. *Reference* is made up of *reference type* and *reference value*. *Reference type* can be further broken into *primitive data type* and *special reference type*. *Primitive data type* is the data type identified by the system, such as *string*, *integer*, *Universally Unique Identifiers* (UUID) and *stringdate*. While the value of *special reference type* can be *button* (user-defined button), or *enum* (enumerated data type). These references require *reference value,* which is additional information for the reference type. *Reference value* is a series of concrete enumerated values or a list of data for the data type *enum*, and *reference value* is the name of the business process model for the data type *button*. The derived model *MAButton* is the bridge between the business process model and the business object model, which represents a special *MAttribute*. The property *storageway* of *MAttribute* presents the data storing, whether in sparse table so as to solve the SaaS "*null schema*" problem.

Business object is defined as 2-tuple, which is the instance of business object model. It is denoted as *BO :=(mobject, mattributes)*, where *mobject* is an instance of *MObject*, and *mattribute*s is a set of instances of *MAttribute*s.

3.3.3 The Business Presentation Model

The business presentation models contain the details of the graphic appearance of SaaS applications. It is composed of *MFrame*, *MCard*, *MElement* and *MVButton*. *MFrame* is the entrance to present business data for users. The instance of *MFrame* is related only to a main *MCard* and some other detail *MCard*s. Users navigate the business data represented with *MFrame* after clicking the link of the system menu. The property *where* of *MFrame* means the value range of business data in the Web User Interface (UI). *MCard* is the thinning of *MFrame* for the sake of the maintenance of a business object. The instance of *MCard* is related to several *MElement*s. The business of *MCard* is often the CRUD and other compound database business. *MElement* is the smallest unit of business presentation models, which may be the presentation of the business data. The important property of *MElement* is *isVisibleUpdate*, *isVisibleView* and *isQueryCondition*. When *isVisibleUpdate* is true, the *MElement* is a storage element. And the business data represented by *MElement* can be modified in the maintenance interface; when *isVisibleView* is true, the *MElement* is a presentation element. The business data represented by *MElement* can be only displayed in the Web UI; when *isQueryCondition* is true, it is as a query condition in the query area. These are known as *storage MElement*, *presentation MElement* and *query*

MElement, respectively. User-defined button *MVButton* is also a kind of *MElement*, and its specific business is defined in the property *referenceValue* of related *MAButton*. In order to support tenant customization, the property *tenantList* of *MElement* means the tenant list which allows the displaying of this element. Only the tenancy in the list can see the impression of *MElement*.

The business presentation object is defined as 2-tuple, which is the instance of business presentation model. It is denoted as *VO :=(mcard, melements)*, where ***Meta****(mcard, MCard)=true*, and *melements* is a set of instances of *MElement*. The Web UI object is denoted as *UI :=(mframe, vos)*, where *mframe* is the instance of *MFrame* and *vos* is a set of *VO*s.

The business presentation object is used to define the graphic UI of the business data represented by the business object model. Therefore, several basic properties of *MCard* and *MElement* of *VO*s are generated from the properties of *MObject* and *MAtttibute* of *BO*s, denoted as $m_1(s)/BO \rightarrow m_2(s)/VO$, where $BO \subset xBC$, $VO \subset xBC$. This generation is an assistant tool for modeling the details of the business presentation models, which are described in binary relation. As mentioned before, a binary relation that is specified by using a set comprehension predicate *P*, e.g., in $R = \{\langle a, b\rangle | P(a, b)\}$. The values of the properties of *VO*s are generated from *MO*s according to the transformation rule r_1 and r_2, shown in Figure 4, which is further defined in the first-order predicate logic of binary relation. The rule r_1 generates *MCard*s, while the rule r_2 generates *MElement*s. The assistant tool executes the mapping rules in order implicitly.

dom $r_1=\{bo|bo \in BO\}$, where $BO \subset xBC$
ran $r_1=\{vo|vo \in VO\}$, where $VO \subset xBC$
$r_1=\{< bo, vo >|$
$(\forall bo \in BO \wedge \exists vo \in VO \wedge vo.mcard.attributes.name="C\_"+bo.mobject.attributes.name \wedge$
$vo.mcard.attributes.description=bo.mobject.attributes.description)\}$

dom $r_2=\{ma|ma \in bo.mattributes\}$, where $bo \in BO$, $BO \subset xBC$
ran $r_2=\{e|e \in vo.melements\}$, where $vo \in VO$, $VO \subset xBC$
$r_2=\{<ma, e>| (\forall ma \in BO\ bo.MA \wedge bo \in BO \wedge \exists e \in vo.elements \wedge vo \in VO$
$e.attributes.name="E\_"+ma.attributes.name \wedge e.attributes.tips=ma.attributes.description \wedge$
$e.attributes.length=ma.attributes.length \wedge e.attributes.defaultValue=ma.attributes.defaultValue \wedge$
$(\forall ma.attributes.referenceType=button \wedge \exists e \in vo.elements \wedge e.attributes.defaultValue='' \wedge$
$e.attributes.isQueryCondition=false \wedge e.attributes.format='') \wedge$
$(\forall ma.attributes.referenceType=integer \wedge \exists$
$e \in vo.elements \wedge e.attributes.defaultValue=0 \wedge e.attributes.format='^-?\d+\$') \wedge$
$(\forall ma.attributes.referenceType=string \wedge \exists e \in vo.elements \wedge e.attributes.defaultValue='') \wedge$
$(\forall ma.attributes.referenceType=stringdate \wedge \exists$
$e \in vo.elements \wedge e.attributes.defaultValue='' \wedge e.attributes.format='yyyyMMdd'))\}$

Figure 4
Model transformation rule from business object model to business presentation model

3.4 Version Control in Extensible Business Component Model

As mentioned, models are the primary artifact of the software development process in MDE. These models are typically developed by distributed environments consisting of teams at different organizations and locations. These teams usually build multiple overlapping models which represent different aspects of the same systems. In addition, models undergo a complex evolution during their life cycles. As a consequence, one of the techniques used to support model management activities is the version control of models. However, present-day MDE tools offer only limited support for the version control of models. Traditional version control systems are based on the copy-modify-merge approach [23], which is not fully exploited in MDE since current implementations lack model-orientation.

In contrast, we use Java Content Repository (JCR) [24] as the storage of models. A content repository, shown in Figure 5, consists of one or more workspaces, each of which contains a tree of items. An item is either a *node* or a *property*. Each node may have zero or more child nodes and zero or more child properties. There is a single root node per workspace which has no parent. All other nodes have one parent. The model may be considered as a node, and the property of the model may be considered as a property. The JCR 2.1 (JSR-333) [24] specification provides simple and independent versioning or full versioning of a node in the repository. A versioning repository has, in addition to one or more workspaces, a special version storage area. A new version is added to the version history of a versionable node when one of its workspace instances is checked-in. The model stored in the repository can be restored to a previous version, which is useful when developers have made some fatal mistake in modeling the system. There are two basic operations of nodes. To create a new version of a versionable node, the application calls *checkin*. In order to alter a versionable node, the node must be *checked out*. There are some open source tools fully conforming to the implementation of the JCR specification, such as Jackrabbit and ModeShape[1].

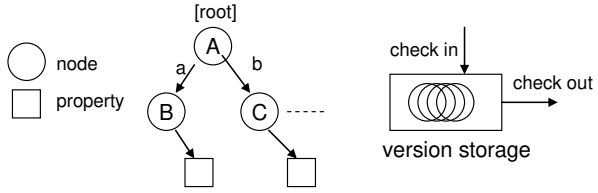


Figure 5
Java Content Repository

[1] Apache Jackrabbit and ModeShape are a JCR implementation that provides access to content stored in many different kinds of systems, which can be downloaded from http://jackrabbit.apache.org and http://www.jboss.org/modeshape respectively.

4 Template-based Model Transformation in Support of Model Synchronization

4.1 Template Engine

4.1.1 Template Data Model

The basic structure of **template data model** is a tree shown in Figure 6. The root node is the Web UI object. All the data models of SaaS applications save in the model repository.

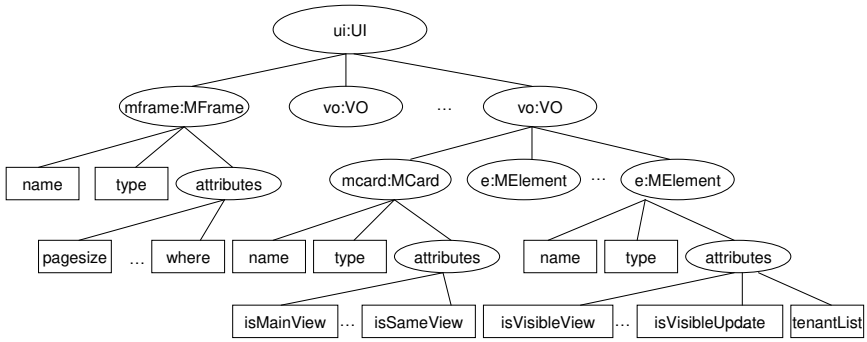


Figure 6
Template data model of xBC

4.1.2 Syntax and Semantics of Template

A **template** is a series of template statements. The set of transformation rules from *xBC* to codes is denoted as $F=\sum template_i$. The model transformation rule of the textual template evolution is based on all the template statements.

A *template statement* is defined as 4-tuple: *TemplateStatement:=<Text, Interpolation, Tag, Comment>*. The **text** is static text and it will keep constant after model transformation; the **interpolation** is used to insert the value of the expression converted to text, which is the dynamic content of templates. The format of interpolations is *${expression}*; the **tag** introduces some evolution mechanism to satisfy the requirements for the specific application field, such as macro, iteration, condition and function statements. Also the tag can execute some directives. In fact there are two types of directives: predefined directives and user-defined directives. User-defined directives are extensions of directives. Some directives have been implemented, such as Macro, Conditional directives, List directives and Function; the **comment** will be ignored and not be written to the output.

4.1.3 Template Component

The main ideas of the template reuse are to divide the templates into parts of the components, and each component can generate the relatively independent target framework. Therefore, a pluggable template called plugin, which is a series of templates, generates code based on some open source libraries (Such as SQL, Spring, Hibernate, MyBatis, Struts, JSF, Web Service, etc.).

4.2 Model Transformation from Extensible Business Component Models to Codes

Aiming to obtain software corresponding to this business system, we can use the architecture composed of Business Component (BC) and Business Process (BP). The system is an integration of composition with many business processes, one of which is connected with a series of business actions. All the BCs and BPs in the runnable system can be generated from the templated-base model transformation approach.

From the viewpoint of model transformation, the model mapping from *xBC*s to codes is denoted as $m_1(s)/xBC \rightarrow m_2(s)/Code$ shown in Figure 7. From the viewpoint of function, the model mapping is denoted as *codes=models+textual template*.

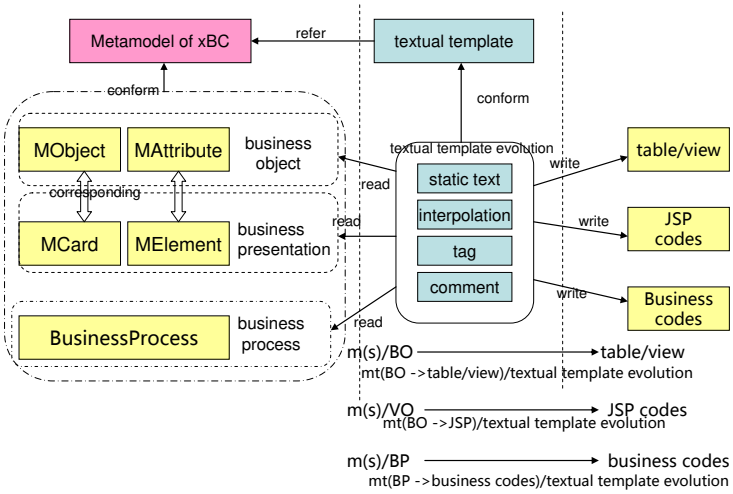


Figure 7
Transformation process between xBCs and codes

4.3 Model Synchronization Based on Model Evolution

Source models involved in model synchronization may face with the modifications shown in Table 2. The modifications in the three circumstances are identified based on the version number. All the version numbers of models involved in model synchronization are recorded. In the next model synchronization, the version of involved models is needed to compare with the last recorded version. If the model is not in the last recorded models, it is **addition**; if the new version is greater than the past and it is not a new model, it indicates that the model is **updated** after the last model synchronization; if one of the last recorded models is not involved in the next model synchronization, it indicates that the source model has been **deleted**. The model synchronization algorithm *PSM2CodeSync* from *xBC* to Web JSP codes is shown in Figure 8.

Table 2
Classification of modifications

Name	Definition
add	Source model is added
delete	Source model is deleted
update	The property of source model is changed

```
Function PSM2CodeSync
Input: ui:UI
Output: codes
// justify whether need code generation
if(isNewModel(ui)){
   PSM2Code(ui);
}elseif (isDeletedOperation(ui)){
   deleteGeneratedCodes(ui);
}else if(modelDetection(ui)){
    PSM2Code(ui);
}

Function PSM2Code
Input: mframe:MFrame
Output: codes
generateCode("query", ui);
generateCode("insert", ui);
generateCode("update", ui);
generateCode("detail", ui);
recordCurrentVersion (ui);//record the last version of models

Function generateCode
Input: templateName, ui:UI
Output: codes
helper.processText();
helper.processInterpolation();
helper.processTag();
helper.processComment();//textual template evolution
```

```
Function modelDetection
Input: ui: UI
Output: true or false
if(versionChange(ui.mframe)) return true;
for each VO vo in ui.vos {
  if(versionChange(vo.mcard)) return true;
  if(versionChange(mcard.mobject)) return true;
  for each MElement melement in vo.melements {
    if(versionChange(melement)) return true;
    if(versionChange(melement.mattribute)) return true;
    if(versionChange(melement.mattribute.referencevalue)) return true;
}}
return false;
```

Figure 8

Model synchronization algorithm between xBCs and codes

Conclusions

This study was aimed at investigating the model transformation approach to generate SaaS applications. The main contributions of the study are outlined below:

1) A data middleware of a multi-tenant database is presented. As this approach is transparent to the application, the applications of tenants can share this data model without any modification. The abstract data model provides the logic isolation of different tenants. The master/slave database in the data cloud is a kind of horizontal scalability to improve the performance of data.

2) In this paper, a novel Extensible business Component model named *xBC* is proposed for describing both the structural and behavioral properties of generic SaaS applications. The tenant expression, the property *storageWay* of *MAttribute*, and the property *tenantList* of *MElement* are presented to support multi-tenancy of SaaS applications. Its architecture of metamodel and extension mechanism is discussed in detail. In addition, we use versioning nodes of JCR as the storage of models. The model stored in the repository can be restored to a previous version according to the version number.

3) Additionally, our approach for model transformation that supports model synchronization based on model evolution is presented. This model transformation approach is based on the textual template evolution, and this model synchronization approach will only take up the storage space of model repositories rather than some extra space. Only the changed models need a subsequent model transformation. That is a good way to minimize the amount of source that needs to be reexamined by a transformation when the source is changed.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Contract Numbers 60873089 and 60903176, the Provincial Natural Science

Foundation for Outstanding Young Scholars of Shandong under Contract Numbers JQ200820, the Program for New Century Excellent Talents in University under Contract Numbers NCET-10-0863 and the Science and Technology Development Program of Shandong Province under Contract Number 2011GGX10116.

References

[1] Sánchez, P., Moreira, A., Fuentes, L., Araújo, J., Magno, J.: Model-driven Development for Early Aspects, Information and Software Technology 52 (2010) 249-273

[2] Gao, X., Li, Z.: Business Process Modeling and Analysis Using UML and Polychromatic Sets, Production Planning and Control 17 (2006) 780-791

[3] Whittle, J., Service Model-Driven Development: A Practical Approach. London: Chapman & Hall; 2012

[4] Mahowald, R P.: Worldwide Software as a Service 2011–2015 Forecast and 2010 Vendor Shares, International Data Corporation, USA, 2011

[5] Narasimhan, B., Nichols, R.: State of Cloud Applications and Platforms: The Cloud Adopters' View, Computer 44 (2011) 24-28

[6] Brunelière, H., Cabot, J., Frédéric, J.: Combining Model-driven Engineering and Cloud Computing, in Proceedings of 6th European Conference on Modelling Foundations and Applications, Paris, France, June 15-18, 2010, pp. 1-2

[7] Guo, J., Sun, W., Huang, Y., Wang, Z., Gao, B.: A Framework for Native Multi-Tenancy Application Development and Management, in Proceedings of The 9th IEEE International Conference on E-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing, Tokyo, Japan, July 23-26, 2007, pp. 551-558

[8] Jacobs, D., Aulbach, S.: Ruminations on Multi-Tenant Databases, BTW 103 (2007) 514-521

[9] Ma K., Chen, Z., Abraham, A., Yang, B., Sun, R.: A Transparent Data Middleware in Support of Multi-Tenancy, in Proceedings of the 7th International Conference on Next Generation Web Services Practices, Salamanca, Spain, October 19-21, 2011, pp. 1-5

[10] Chang, S. H., Kim, S. D.: A Variability Modeling Method for Adaptable Services in Service-Oriented Computing, in Proceedings of the 11th International Software Product Line Conference, Kyoto, Japan, September 10-14, 2007, pp. 261-268

[11] Liu, Y., Zhang, B., Liu, G., Wang, D., Gao, Y.: Personalized Modeling for SaaS Based on Extended WSCL, in Proceedings of the 2010 IEEE Asia-Pacific Services Computing Conference, Hang Zhou, China, December 06-10, 2010, pp. 355-362

[12] Wang, Z., Zhao, Z., Fang, J., Wang X.: A SaaS-Friendly Service Community Model and Its Application in the Nationwide Service Network for Sharing Science and Technology Information, Chinese Journal of Computers 33 (2010) 2033-2043

[13] Mukerji, J., Miller, J.: The MDA Guide Version 1.0.1, Object Management Group, USA, 2003

[14] Subramanyam, R., Weisstein, F. L., Krishnan, M. S.: User Participation in Software Development Projects, Communications of the ACM 53 (2010) 137-141

[15] Egyed, A.: Automatically Detecting and Tracking Inconsistencies in Software Design Models, IEEE Transactions on Software Engineering 37 (2011) 188-204

[16] Czarnecki, K., Helsen, S.: Feature-based Survey of Model Transformation Approaches, IBM System Journal 45 (2006) 621-645

[17] Object Management Group: Meta Object Facility (MOF) 2.0 Query/View/Transformation Final Adopted Specification 1.1, Object Management Group, USA, 2011

[18] ikv++ technologies ag: medini QVT 1.7.0 (2011), http://projects.ikv.de/qvt/

[19] Giese, H., Wagner, R.: From Model Transformation to Incremental Bidirectional Model Synchronization, Software and Systems Modeling 8 (2009) 21-43

[20] Ma, K., Yang, B., Chen, Z., Abraham, A.: A Relational Approach to Model Transformation with QVT Relations Supporting Model Synchronization, Journal of Universal Computer Science 17 (2011) 1863-1883

[21] Duggan, D., Service Oriented Architecture: Entities, Services, and Resources. NJ.: Wiley-IEEE Computer Society; 2012

[22] Tanter, É., Toledo, R., Pothier, G., Noyéb, J.: Flexible Metaprogramming and AOP in Java, Software and Systems Modeling 72 (2008) 22-30

[23] Collins-Sussman, B., Fitzpatrick, B. W., Pilato, C. M., Version Control with Subversion for Subversion 1.6: The Official Guide And Reference Manual. NY.: Soho Press; 2010

[24] Nuescheler, D.: JSR 333: Content Repository for Java Technology API Version 2.1 Early Draft Review, Java Community Process, USA, 2011

Compressive Behaviour of Metal Matrix Syntactic Foams

Imre Norbert Orbulov

Depatment of Materials Science and Engineering, Budapest University of Technology and Economics, Bertalan Lajos utca 7, H-1111, Budapest, Hungary, orbulov@eik.bme.hu

János Ginsztler

Research Group for Metals Technology of the Hungarian Academy of Sciences, Bertalan Lajos utca 7, H-1111, Budapest, Hungary, jginsztler@mti.bme.hu

Abstract: The compressive behaviour of three different metal matrix syntactic foams (MMSFs) was investigated. The results showed that the engineering factors such as the size of the used hollow spheres, the aspect ratio (height / diameter ratio) of the specimens and the temperature of the tests have significant effects on the compressive strength and properties. The smaller microballoons with thinner wall ensured higher compressive strength due to their more flawless microstructure and better mechanical stability. The higher aspect ratio of the specimens resulted in worse compressive characteristics (lower strength, lower specific energy consuming capacity). The elevated temperature tests revealed ~30% drop in the compressive strength. However, the strength remained high enough for structural applications; therefore MMSFs are good choices for light structural parts working at elevated or room temperature. The proper size selection of the reinforcing hollow spheres ensures potential for tailoring the compressive characteristics of MMSFs.

Keywords: metal matrix composite; syntactic foams; metallic foams; compressive properties

1 Introduction

Nowadays metallic foams have become more and more important, and this is confirmed by the increasing number of papers published on this topic. The 'conventional' metallic foams, which consist of a metal structure, a gas phase and stabilising particles, have been written about widely in the literature thanks to their potential application possibilities as automotive parts, energy absorbers or blast and collision damping elements in buildings or vehicles, etc. However, there are

still existing problems, for example in the foaming process [1, 2]. The metallic foams have a special class, namely metal matrix syntactic foams (MMSFs). In MMSFs, the porosity is ensured by the incorporation of ceramic microballoons [3, 4]. The microballoons are commercially available and they contain mainly various oxide ceramics [5, 6]. The quality of the microballoons (uniform wall thickness and flawless wall) has a strong effect on the mechanical and other properties of the foams. The MMSFs have numerous perspective applications (covers, hulls, walls, castings, or in the automotive industry sectors) because of their high energy absorbing and damping capability and due to their low density [7].

The MMSFs can be produced by pressure infiltration or by stir casting; both ways are common in the literature. The main mechanical load mode of MMSFs is compression; therefore, the compression characteristics have been investigated in some aspects. Palmer et al. studied the pressure infiltration process and mechanical behaviour of various microballoon and metal matrix combinations. Compressive stress-strain data were gathered for foams prepared from combinations of Al1350, Al5083 and Al6061 alloys for both 45 μm and 270 μm spheres [8]. Balch et al. fabricated aluminium matrix MMSFs by liquid metal infiltration of commercially pure (cp-Al) and Al7075 aluminium. The cp-Al foam exhibited peak strengths in compression of over 100 MPa, but the Al7075 matrix foams had significantly higher peak strengths, up to 230 MPa [9, 10]. Rohatgi et al. investigated the pressure infiltration technique of nickel coated and uncoated microballoons. In their other work, loose beds of microballoons (cenospheres) were pressure infiltrated with A356 alloy melt to fabricate MMSFs. The processing variables included melt temperature, gas pressure and the size of microballoons. The effect of these processing variables on the microstructure and compressive properties of the synthesized composites was characterized [11, 12]. Kiser et al. performed investigations on the mechanical response of MMSFs under both uniaxial compression and constrained die compression loadings. The key material parameters that varied were the matrix strength and the ratio of the wall thickness to radius of the microballoons. They observed that the energy absorption capacity was extremely high in comparison with values that are typical of ‘conventional’ metal foams [13]. Wu et al. established a new method to predict the compressive strength of MMSFs, showing the relation between the relative wall thickness of the microballoons and the compressive strength of such foams. The deformation mechanisms of syntactic foams was also discussed [14]. Tao et al. investigated the mechanical properties of MMSFs with monomodal and bimodal distribution of microballoons. The bimodal foams have the advantages of a flat plateau regime, high plateau stress, lower density and good ductility. In the next step, Al matrix MMSFs with additional Al particles embedded were fabricated by pressure infiltration. With the introduction of Al particles, the ductility of the syntactic foams was significantly increased, and the compressive strength also increased by up to 30% [15, 16]. Zhang et al. manufactured aluminium matrix MMSFs with low-cost porous ceramic spheres of diameters between 0.25 and 4 mm by pressure infiltration casting. The mechanical response of the syntactic

foams with different sphere sizes and densities under static and dynamic conditions was investigated. They found that the plateau strength, and thus the amount of energy absorption of the syntactic foam, was largely determined by the volume fraction of Al and to a lesser extent by the mechanical properties of the ceramic spheres in the foam [17]. In the works of Mondal et al., microballoons in the range of 30–50 vol% were used as space holders for making syntactic aluminium foam using a stir-casting technique. The synthesized MMSF was characterized in terms of microstructures, hardness and compressive deformation behaviour. The plateau stress of these MMSFs is considerably higher than those of conventional aluminium foams. The dry sliding wear behaviour of MMSFs has also been studied using a pin-on-disc apparatus [18, 19]. Rabiei and O'Neill produced MMSFs using gravity casting techniques. The foam was comprised of steel hollow spheres packed into a random dense arrangement, with the interstitial space between spheres infiltrated with a casting aluminium alloy. The composite foam developed in the study displayed superior compressive strength and energy absorption capacity [20]. Ramachandra and Radhakrishna synthesized aluminium based MMSFs containing up to 15 wt% of microballoons by the stir casting method. Properties like density, hardness, microhardness, ductility and ultimate tensile strength were investigated. The MMSFs produced were also subjected to corrosion, dry sliding wear and slurry erosive wear tests. The addition of microballoons reduced the density of composites while increasing some of their mechanical properties. The results of wear studies have shown that the resistance to wear increased with an increase in the percentage of microballoons [21, 22]. In the work of Couteau and Dunand, aluminium MMSFs with densities of 1.2-1.5 g/cm^3 were deformed at 500°C under constant uniaxial compressive stresses ranging from 5 to 14 MPa. The foam's creep behaviour was characterized by a short primary stage and a long secondary stage where the strain rate was constant and minimum [23]. The microballoons in MMSFs work as stress concentrators and have influence on the crack propagation in materials [24].

Most of the MMSFs are produced by pressure infiltration; therefore, the infiltration parameters (like required threshold pressure) have also been studied. Trumble presented an analysis of spontaneous infiltration to model non-cylindrical pores. [25]. Bárczy and Kaptay developed a new infiltration model for 'closely packed equal sphere - CPES' structure. In their study the threshold pressure, the threshold contact angle and the equilibrium height of penetration were determined. All these parameters are significantly different from those obtained from the traditional capillary penetration model, but similar to the Carman model. The experiments demonstrated the reliability of the theoretical results [26]. Asthana et al. also overviewed some fundamental materials phenomena relevant to the infiltration processing of metal-matrix composites. They stated that the lack of a comprehensive theoretical framework required further research to be done [27].

The aim of this paper is to extend the knowledge regarding MMSFs by analysing the effects of the microballoon size, the elevated test temperature and the effects of the aspect ratio (height / diameter ratio) of compression specimens.

2 Investigated Materials

Overall three types of MMSFs were produced by pressure infiltration from the combination of technical purity aluminium (Al99.5) and three ceramic microballoons (designated by SL150, SL300 and Globocer). The chemical composition of the matrix material is listed in Table 1.

Table 1
Chemical composition of Al99.5 material

Matrix	Composition (wt%)								
	Al	Si	Fe	Cu	Mn	Mg	Zn	V	Ti
Al99.5	99.5	0.25	0.4	0.05	0.05	0.05	0.05	0.05	0.03

The SL150 and SL300 microballoons were provided by Envirospheres Pty Ltd [6], while the Globocer type hollow spheres were shipped by Hollomet GmbH [28]. The main differences between the three types are in the average diameter, density and wall thickness. Their main parameters are summarized in Table 2.

Table 2
Morphological properties and phase constitution of the applied hollow ceramic spheres

Type	Density at 64 vol% (gcm^{-3})	Average		Al_2O_3	SiO_2	Mullite	Quartz
		diameter (μm)	thickness (μm)	(wt%)			
SL150	0.576	100	3.69	30-35	45-50	15-20	0-5
SL300	0.691	150	6.75				
Globocer	0.816	1450	58				

The MMSFs were produced by pressure infiltration in a special infiltration chamber (Fig. 1). In the first step, the microballoons were poured into a mould to the half and they were densified by gentle tapping and knocking to get a randomly closed pack structure (RCPS). The maximal volume fraction that can be reached with quasi-equal diameter spheres is 64 vol%, as is published in [29]. After this, a layer of alumina mat separator was placed on the top of the microballoons and a block of matrix material was put on the mat. The mould was situated into the infiltration chamber, the chamber was closed and the whole system was evacuated by a vacuum pump (rough vacuum). Proper heating was ensured by three heating zones and the temperatures of the matrix block and the microballoons were

continuously monitored by two thermocouples. After the melting of the matrix, the vacuum pump was switched off and argon gas was let into the chamber at a previously set pressure. Due to this, a pressure difference was built up between the mould (vacuum) and the chamber (argon pressure). This pressure difference forced the molten metal to infiltrate the space between the microballoons. After solidification the mould was removed and cooled to room temperature.

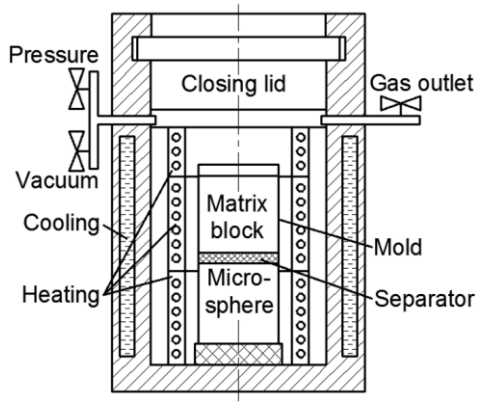


Figure 1
Schematic sketch of the infiltration chamber

Then the complete MMSF block (~40×60×180 mm) could be removed from the mould. For further details about the production process, please refer to [4]. The blocks were designated by their constituents: for example, Al99.5-SL150 stands for an MMSF block with Al99.5 matrix and with ~64 vol% SL150 microballoons. The main physical properties, such as density, porosity, are presented in Table 3.

Table 3
Density and porosity values of the prepared MMSFs

Specimen	Density (g/cm^3)		Porosity (%)		
	Theoretical	Measured	Microballoon	Matrix	Total
Al99.5-SL150	1.34	1.43	50.9	-6.3	44.7
Al99.5-SL300	1.41	1.52	48.2	-7.2	41.0
Al99.5-Globocer	1.53	1.49	45.0	2.6	47.6

The theoretical density and microballoon-porosity (the porosity ensured by the hollow spheres) were calculated from the average geometrical parameters of the microballoons. The matrix porosities (the volume of the pores in the matrix material divided by the volume of the whole specimen) were calculated as the difference between the theoretical and measured density divided by the theoretical density. The negative matrix porosity refers to the infiltrated microballoons (the microballoon-porosity should be decreased). The values of the matrix porosity always remained below 7.2%, so the infiltration can be qualified as a suitable one.

3 Experiments

The main loading mode of foam materials is the compression; therefore compression tests were performed on the cylindrical specimens. The diameter (D) of the specimens was 14 mm and the height (H) of the specimens was 14, 21 and 28 mm (aspect ratio (H/D) 1, 1.5 and 2 respectively). The compression tests were performed on a MTS 810 type universal testing machine in a four column tool at room temperature. The surfaces of the tool were grinded and polished. The specimens and the tool were lubricated with anti-seize material. The test speed was 0.15 mm/s, which ensured quasi-static compression. Six specimens were compressed at room temperature and at elevated (220 °C) temperature until 50% engineering strain from each MMSF type to get representative results. Overall 36 specimens were compressed (at room temperature: 6 pcs Al99.5-SL150, 6 pcs Al99.5-SL300 and 6 pcs Al99.5-Globocer; at elevated temperature: 6 pcs Al99.5-SL150 and 6 pcs Al99.5-SL300). The aim of these tests was to determine how the MMSFs would perform as structural elements at higher temperature. The tests were performed and evaluated in accordance with the ruling standard about the compression tests of cellular materials [30].

4 Results and Discussion

The load bearing capacity of MMSFs depends on many parameters, such as the type and size of the microballoons, the test temperature, etc. In order to characterize these effects, we have done numerous compression tests as described in the previous section. During the tests, the engineering stress–engineering strain curves were plotted. The size of the hollow spheres has a detrimental effect on the compressive behaviour of MMSFs. The smaller hollow spheres (SL150 and SL300) ensured high compressive strength, but the first appearance of the fracture was sudden and quite rigid. For example, Fig. 2 shows the graph for an MMSF specimen containing small microballoons (Ø100 µm) tested at room temperature. Gupta et al. [31] and Bunn and Mottram [32] have investigated polymer matrix syntactic foams with similar stress–strain diagrams. According to the results of Gupta et al. the general stress–strain curves were divided into three parts [31]. Based on their idea, the diagram of MMSFs can be divided into three main parts containing overall five sections. In the first section (from point A to B) the specimens were deformed elastically only. In this section, the microballoons remained unharmed, as can be observed in Fig. 2a; there are no cracks at all. The overall deformation is related to the elastic deformation of the composite. The slope of the first part is defined as structural stiffness (S (MPa); see [30] about the standardized compression test of cellular materials). The stiffness is one of the characterizing properties of the MMSFs. In the vicinity of point A, the deviation from the fitted dashed line can be caused by the internal friction of the material or

by the springs of the tool and the natural friction of the sliding parts of the tool. Due to this, it should be distracted from the measured strain.

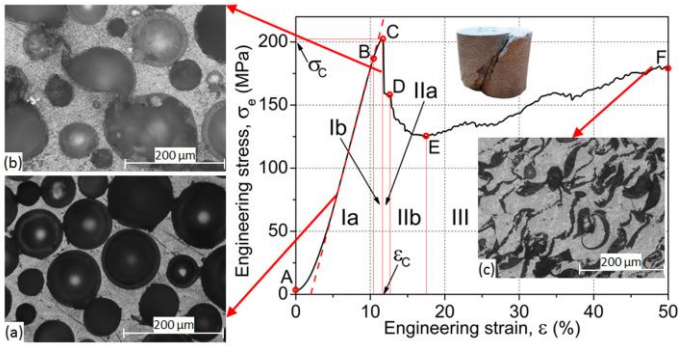


Figure 2
Typical compressive diagram of Al99.5-SL150-H/D=1.5 MMSF specimen

In the second section, from point B to C, the plastic deformation of the matrix began. The load transfer between the matrix and the microballoons increased to its maximum, but the microspheres remained still unharmed. At the end of this section, at point C, the stress reached the compressive strength (σ_c (MPa)) at the fracture strain (ε_c (%)). These parameters are also important characterizing properties, because they show the load bearing capacity of the MMSFs directly. At point C, the first crack appeared in the specimen. This first rupture was very thin and very sharp, and only one row of the microballoons was cracked, as is shown in Fig. 2b. The plane of the crack closed ~45° with the load direction, because in the case of uniaxial loading, the maximum shear load appears in this direction. The stress suddenly dropped to point D due to the reduced load bearing capacity caused by the fracture of the microballoons and the movement of the recently formed specimen halves. From point D to E, the fracture band expanded and the crack became thicker and thicker. The neighbouring microballoons broke and the load bearing capacity decreased further, but more slowly due to the friction between the specimen halves. This deformation phenomenon consumed significant strain and mechanical energy due to the fracture of the ceramic microballoons and due to the plastic deformation of the matrix. The absorbed specific mechanical energy (W (J/cm^3)) is the fourth main characterizing parameter of the MMSFs, as it indicates the damping and protecting capability of the MMSFs against a blast, collision or simple vibration. The absorbed specific energy is equal to the area under the recorded stress-strain curve and can be integrated numerically. From point E, the complete densification of the specimens took place. At the end of this process the cavities of the broken microballoons were filled by the matrix material due to its plastic deformation (Fig. 2c). This part, the plateau region, absorbs a lot of energy, because it is relatively long and has high stress value. The shape of the diagrams after point E can be ascending or constant (usually ascending because the densifying material needs higher force to

be deformed). It may contain larger drops or local peaks due to secondary cracks. The process ended at 50% engineering strain when the test stopped (F in Fig. 2).

Larger hollow spheres (Globocers with Ø1450 µm) caused somewhat different behaviour during room temperature compressive tests (Fig. 3). The tracked properties were the same (compressive strength, fracture strain, structural stiffness and the absorbed specific energy), but the sudden drop in the force after the first stress peak was missing, so part IIa and IIb can be defined together as part IIa+b. The failure of the MMSFs was smoother, without large and sudden force drops. In Fig. 3a a cross section from the linear part is magnified; the hollow spheres are in good shape, and there is no sign of any crack. After the first stress peak, the hollow spheres began to crash (above the dashed line in Fig. 3b). At the end of the test the specimen were fully compacted, and only a few hollow spheres remained unharmed in the compression cone near to the surface of the tool (Fig. 3c).

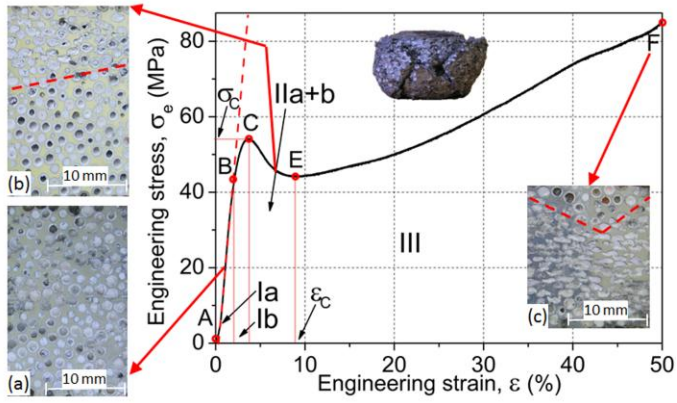


Figure 3

Typical compressive diagram of Al99.5-Globocer-H/D=1.5 MMSF specimen

The mechanical properties of the MMSFs can be characterized through the analysis of the above mentioned four material properties. The most important among them is the compressive strength, which is responsible for the load bearing capacity (Fig. 4).

The smaller hollow spheres ensured higher compressive strength than larger ones. This effect can be observed in the micrometer range also. The specimens containing smaller SL150 microballoons (average diameter 100 µm) have about 10% higher compressive strength than the ones containing larger SL300 (average diameter 150 µm) type microballoons. As shown in Table 2, the SL type microballoons are significantly smaller and they also have thinner wall. The smaller diameter and higher curvature give higher compressive strength and mechanical stability to the microballoons. Moreover, smaller wall thickness ensures lower probability for deflections; therefore the small SL type microballoons have higher compressive strength than the larger, Globocer type microballoons with thicker walls and more defects.

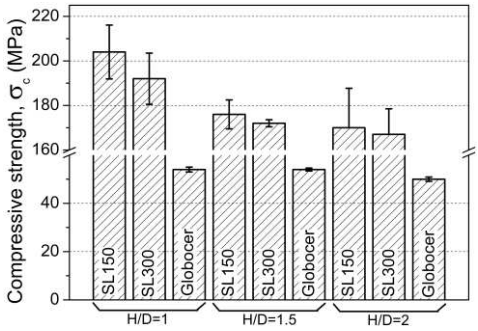


Figure 4
The effect of the microballoon size and aspect ratio on the compressive strength

Ceramics are especially sensitive for deflections; any small rupture or cavity can be the starting point of a crack. The effect of aspect ratio is also evident from Fig. 4. As the specimen height-specimen diameter ratio increased, the compressive strength decreased respectively. The MMSFs reinforced with smaller, SL type microballoons were more sensitive to this effect. In their case, the compressive strength drop is large (more than 30 MPa between H/D=1 and H/D=2), the specimens were buckled and shearing mode failure was observed even for specimens with H/D=1.5. In the case of bulky materials, this effect normally takes place if the aspect ratio is larger than 2.4. We can explain this phenomenon by the properties of the ceramic materials. They are quite sensitive to shear stresses; therefore, if there was any minimal shearing (due to not perfect specimen alignment for example), the negative effect of the shearing loading would be amplified by the sensitivity of the ceramic hollow spheres. The larger hollow spheres in the case of Globocer reinforced MMSFs were compressed rather than sheared, and therefore the aspect ratio has negligible effect on the compressive strength in this case; it remained almost constant with very narrow scatter. The next investigated property was the fracture strain (Fig. 5).

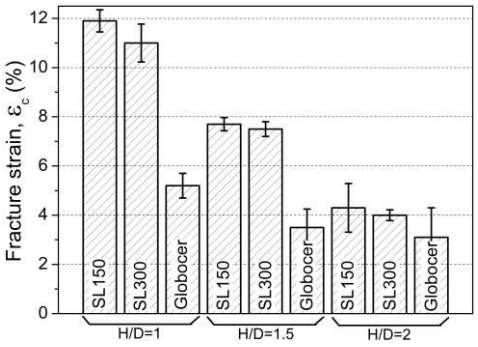


Figure 5
The effect of the microballoon size and aspect ratio on the fracture strain

The fracture strain showed similar behaviour as the compressive strength. The decrease in the strain was almost linear in the case of smaller hollow spheres, and larger spheres caused failure at smaller strains. These trends can be confirmed by the observation of the structural stiffness values (Fig. 6).

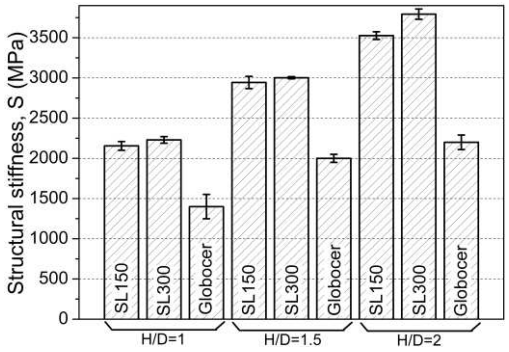


Figure 6

The effect of the microballoon size and aspect ratio on the structural stiffness

The initial slope of the plotted stress - strain curves increased linearly in all cases. The SL300 type reinforcement (larger microballoons) ensured higher structural stiffness than the smaller SL150 microballoon reinforcement, but the even larger Globocer type hollow spheres showed lower stiffness. This phenomenon can be explained by the different failure modes of the hollow spheres. The smaller ones were sheared and the larger ones were compressed. Finally, the consumed specific mechanical energies during the compressive tests are shown in Fig. 7.

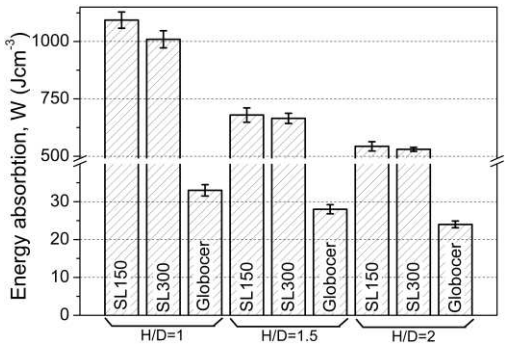


Figure 7

The effect of the microballoon size and aspect ratio on the consumed mechanical energy

As can be observed in Fig. 7, the consumed mechanical energy decreased significantly by the increment of the hollow spheres' size. Moreover the energy also decreased by the increasing aspect ratio. This phenomenon is indicated by the lower strength of the MMSFs. The lower compressive strength caused lower

plateau strength and, due to this, a smaller area under the compressive curve and lower consumed specific energy. This effect was more pronounced in the case of smaller hollow spheres; the H/D increment caused about 50% reduction in consumed energy. In the case of Globocer type hollow spheres, this reduction was only about 5-10%.

The influence of elevated test temperature is shown in Fig. 8 and presented by the example of SL type hollow spheres reinforced MMSFs.

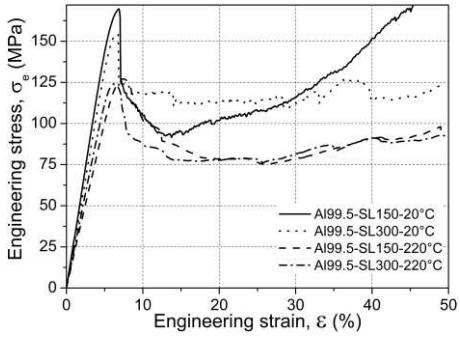


Figure 8
The effect of test temperature on the compressive behaviour of Al.995-SL300 MMSFs

In Fig. 8, the diagrams of the SL type hollow sphere reinforced MMSFs compressed at room and at elevated (220°C) temperature are shown and compared. Due to the elevated temperature, the compressive strength dropped by ~30-35%. In addition to this, the formability increased significantly, and due to this dual effect, the transmitted stress between the matrix and the hollow spheres increased slower than at room temperature and the first fracture appeared at higher strain. In short: the fracture strain increased by ~5% in all cases. The MMSFs became more ductile, but they remained strong enough and can be applied as structural elements: the compressive strengths were still above 120 MPa. This capability at elevated temperature is a serious advantage compared to the conventional metal and polymer foams and makes the MMSFs a good choice for structural parts in the neighbourhood of combustion engines or other heat producing systems. The absorbed specific energies were also decreased due to the lower compressive strength induced lower plateau strength. The effect of the microballoons' type was the same at elevated temperature too. The MMSFs with SL300 type microballoons showed ~5% lower compressive strength.

Conclusions

From the results of the above mentioned and discussed measurements the following conclusions can be drawn:

- The typical compression diagram of the MMSFs can be divided into three main parts containing five sections. The peak strength (compressive

strength), its strain (fracture strain), the structural stiffness and the area below the graph (the absorbed specific mechanical energy) can be applied as characterizing values of the compressive behaviour.

- The smaller, SL type microballoons ensured higher compressive strength, higher fracture strain and higher structural stiffness than the larger Globocer hollow spheres in all circumstances. In addition to the higher curvature and therefore higher compressive strength, the thinner wall of the smaller microballoons contains fewer defects than the thicker wall of the larger ones. The differences were significant between SL150 and SL300 microballoons too; again, the smaller SL150 type microballoons were stronger.
- The increased test temperature caused a ~30% drop in the compressive strength, while the fracture strain increased by ~5%. The structural stiffness decreased, and thus the MMSFs were more ductile than at room temperature. The decrease of compressive strength is significant but not too large; therefore, the MMSFs can be applied as structural elements at elevated temperatures.

In summary the compressive properties of MMSFs can be tailored by proper selection of the materials and by careful design for individual and unique applications.

Acknowledgement

The Metal Matrix Composites Laboratory is supported by Grant # GVOP 3.2.1-2004-04-0145/3.0. This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The investigations were supported by The Hungarian Research Fund, NKTH-OTKA PD 83687. This work is connected to the scientific program of the " Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project. This project is supported by the New Széchenyi Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002). Thanks to C. H. Erbslöh Hungaria Ltd. and R. Tóth for providing the E-spheres.

References

[1] Babcsán N, Leitlmeier D, Banhart J: Metal Foams—High Temperature Colloids Part I. Ex Situ Analysis of Metal Foams. Colloids and Surfaces A: Physicochemical and Engineering Aspects. 2005;261(1-3):123-30

[2] Babcsán N, Moreno FG, Banhart J: Metal Foams—High Temperature Colloids Part II: In Situ Analysis of Metal Foams. Colloids and Surfaces A: Physicochemical and Engineering Aspects. 2007;309(1-3):254-63

[3] Erikson R: A Survey of Current Technology. 5th Aerospace Materials. Von Braun Center, Huntsville, Alabama 2002

[4] Orbulov IN, Dobránszky J: Producing Metal Matrix Syntactic Foams by Pressure Infiltration. Periodica Polytechnica Mechanical Engineering. 2008;52(1):35-42

[5] Sphere Services Inc, http://wwwsphereservicescom/, last accessed: 6th December 2011

[6] Envirospheres Ltd, http://wwwenvirospherescom/productsasp, last accessed: 6th December 2011

[7] Tudorache T, Popescu M: FEM Optimal Design of Wind Energy-based Heater. Acta Polytechnica Hungarica. 2009;6(2):55-70

[8] Palmer R, Gao K, Doan T, Green L, Cavallaro G: Pressure Infiltrated Syntactic Foams—Process Development and Mechanical Properties. Materials Science and Engineering: A. 2007;464(1-2):85-92

[9] Balch D, Odwyer J, Davis G, Cady C, Grayiii G, Dunand D: Plasticity and Damage in Aluminum Syntactic Foams Deformed under Dynamic and Quasi-Static Conditions. Materials Science and Engineering A. 2005;391(1-2):408-17

[10] Balch D, Dunand D: Load Partitioning in Aluminum Syntactic Foams Containing Ceramic Microspheres. Acta Materialia. 2006;54(6):1501-11

[11] Rohatgi PK, Guo RQ, Iksan H, Borchelt EJ, Asthana R: Pressure Infiltration Technique for Synthesis of Aluminum–Fly Ash Particulate Composite. Materials Science and Engineering A. 1998;244:22-30

[12] Rohatgi P, Kim J, Gupta N, Alaraj S, Daoud A: Compressive Characteristics of A356/fly Ash Cenosphere Composites Synthesized by Pressure Infiltration Technique. Composites Part A: Applied Science and Manufacturing. 2006;37(3):430-7

[13] Kiser M, He MY, Zok FW: The Mechanical Response of Ceramic Microballoon Reinforced Aluminum Matrix Composites under Compressive Loading. Acta Materialia. 1999;47(9):2685-94

[14] Wu G, Dou Z, Sun D, Jiang L, Ding B, He B: Compression Behaviors of Cenosphere–Pure Aluminum Syntactic Foams. Scripta Materialia. 2007;56(3):221-4

[15] Tao XF, Zhang LP, Zhao YY: Al Matrix Syntactic Foam Fabricated with Bimodal Ceramic Microspheres. Materials & Design. 2009;30(7):2732-6

[16] Tao XF, Zhao YY: Compressive Behavior of Al Matrix Syntactic Foams Toughened with Al Particles. Scripta Materialia. 2009;61(5):461-4

[17] Zhang LP, Zhao YY: Mechanical Response of Al Matrix Syntactic Foams Produced by Pressure Infiltration Casting. Journal of Composite Materials. 2007;41(17):2105-17

[18] Mondal DP, Das S, Ramakrishnan N, Uday Bhasker K: Cenosphere-filled Aluminum Syntactic Foam Made through Stir-Casting Technique.

Composites Part A: Applied Science and Manufacturing. 2009;40(3):279-88

[19] Mondal DP, Das S, Jha N: Dry Sliding Wear Behaviour of Aluminum Syntactic Foam. Materials & Design. 2009;30(7):2563-8

[20] Rabiei A, Oneill A: A Study on Processing of a Composite Metal Foam via Casting. Materials Science and Engineering: A. 2005;404(1-2):159-64

[21] Ramachandra M, Radhakrishna K: Synthesis-Microstructure-Mechanical Properties-Wear and Corrosion Behavior of an Al-Si (12%)—Flyash Metal Matrix Composite. Journal of Materials Science. 2005;40(22):5989-97

[22] Ramachandra M, Radhakrishna K: Effect of Reinforcement of Flyash on Sliding Wear, Slurry Erosive Wear and Corrosive Behavior of Aluminium Matrix Composite. Wear. 2007;262(11-12):1450-62

[23] Couteau O, Dunand D: Creep of Aluminum Syntactic Foams. Materials Science and Engineering: A. 2008;488(1-2):573-9

[24] Kuffová M, Nečas P: Fracture Mechanics Prevention: Comprehensive Approach-based Modelling? Acta Polytechnica Hungarica. 2010;7(5):5-17

[25] Trumble KP: Spontaneous Infiltration of Non-Cylindrical Porosity Close Packed Spheres. Acta Materialia. 1998;46(7):2363-7

[26] Bárczy T, Kaptay G: Modelling the Infiltration of Liquid Metals unto Porous Ceramics. Materials Science Forum. 2005;473-474:297-302

[27] Asthana R, Rohatgi PK, Tewari N: Infiltration Processing of Metal - Matrix Composites: a Review. Processing of Advanced MAterials. 1992;2:1-17

[28] Hollomet GmbH, http://wwwhollometcom/cms/, last accessed 28th November 2011

[29] Jaeger HM, Nagel SR: Physics of the Granular State. Science. 1992;5051:1523-31

[30] Testing of Metallic Materials - Compression Test of Metallic Cellular Materials, DIN50134 standard, October 2008

[31] Gupta N, Kishore, Woldesenbet E, Sankaran S: Studies on Compressive Failure Features in Syntactic Foam Material. Journal of Materials Science. 2001;36:4485-91

[32] Bunn P, Mottram JT: Manufacture and Compression Properties of Syntactic Foams. Composites. 1993;24(7):565-71

Stable Design of a Class of Nonlinear Discrete-Time MIMO Fuzzy Control Systems

Radu-Emil Precup[1], Marius-Lucian Tomescu[2], Emil M. Petriu[3], Stefan Preitl[1], Claudia-Adina Dragoş[1]

[1] Department of Automation and Applied Informatics, "Politehnica" University of Timisoara, Bd. V. Parvan 2, RO-300223 Timisoara, Romania
E-mail: radu.precup@aut.upt.ro, stefan.preitl@aut.upt.ro, claudia.dragos@aut.upt.ro

[2] Faculty of Computer Science, "Aurel Vlaicu" University of Arad, Complex Universitar M, Str. Elena Dragoi 2, RO-310330 Arad, Romania
E-mail: tom_uav@yahoo.com

[3] 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 presents a new stability analysis approach dedicated to a class of nonlinear discrete-time multi input-multi output (MIMO) Takagi-Sugeno fuzzy control systems (FCSs). The theorem presented in this paper offers sufficient conditions for the global stability of the FCSs. The applicability of the theoretical results is illustrated by the stable design of Takagi-Sugeno fuzzy controllers for the level control of spherical three tank systems as nonlinear MIMO processes. Digital simulation results are included.

Keywords: eigenvalues; MIMO fuzzy control systems; stability analysis; Takagi-Sugeno fuzzy controllers; three tank systems

1 Introduction

The stable design of fuzzy control systems (FCSs) is important because it contributes to the fulfilment of very good performance. Many popular stability analysis solutions concerning Takagi-Sugeno (T-S) FCSs are offered in this context, and their usual formulation is done in the linear matrix inequality (LMI) framework. The main features of these solutions are:

- The linearization can result in uncertainties and inaccuracies of fuzzy models.
- The quadratic Lyapunov functions may lead usually to conservative stability conditions.

- Although the LMIs are computationally solvable, they require numerical algorithms implemented by software tools.

Some approaches to the stability analysis of multi input-multi output (MIMO) T-S FCSs have been reported recently in the literature. Based on a novel fuzzy Lyapunov-Krasovskii functional, a stability analysis and stabilization for a class of discrete-time Takagi-Sugeno fuzzy systems is developed in [1]. A useful property of the staircase membership functions and a set of linear-matrix-inequality (LMI), the stability conditions for fuzzy control systems are offered in [2]-[4]. Sufficient conditions for the exponential stability of type-1 and type-2 T-S FCSs are given in [5]-[7] in fuzzy positive systems formulations. Fuzzy control design based on adaptive control schemes are proposed in [8]-[11].

The new contribution of this paper with respect to the state of the art is a stability analysis theorem dedicated to nonlinear MIMO processes controlled by T-S fuzzy controllers (FCs). Our original proof of the stability analysis theorem is based on the eigenvalues of the matrices of quadratic forms. Since these matrices are actually vector functions of vector arguments, their eigenvalues are functions of state variables. Similar approaches but with different stability formulations and proofs are reported in [12]-[15].

The specific features of the stability analysis theorem proposed in this paper concern the avoidance of both process linearization and the LMIs in the derivation and proof of the stability conditions because there is no need to calculate common positive definite matrices. Those are the reasons why the suggested approach proves to be advantageous with respect to LMI-based stability analysis solutions. Furthermore, the stability analysis method is formulated here so as to be well suited for T-S FC designs dedicated to a wide class of nonlinear processes [16]-[26].

This paper is organized as follows. Section 2 defines the structure of T-S FCSs which control a class of nonlinear MIMO processes. Section 3 gives the stability theorem for discrete-time MIMO FCSs. A case study presented in Section 4 offers the stable design of T-S FCSs dedicated to the level control of spherical three tank systems and digital simulation results. The conclusions are discussed in Section 5.

2 Fuzzy Control System Structure

The MIMO FCS structure is presented in Figure 1. Let $X \subset R^n$ ($n \in N$, $n > 0$) be the universe of discourse. The nonlinear MIMO process is characterized by the discrete-time input affine state-space model

$$\begin{aligned} &\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}(t),\ t \in N,\ \mathbf{x}(0) = \mathbf{x}_0 \in X, \\ &\mathbf{y}(t) = \mathbf{g}(\mathbf{x}(t)). \end{aligned} \tag{1}$$

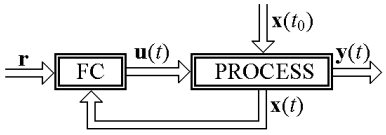


Figure 1
MIMO FCS structure

Figure 1 illustrates: **r** – the reference input vector which is constant for stabilizing control systems, **y** – the controlled output vector, $\mathbf{x} \in X$ – the state vector, $\mathbf{x}(t) = [x_1(t) \quad x_2(t) \quad ... \quad x_n(t)]^T \in X$; the superscript T stands for matrix transposition, t is the time variable (with the initial time moment $t_0 = 0$), $\mathbf{x}_0$ is the initial condition vector, $\mathbf{f} : R^n \to R$, $\mathbf{B} : R^n \to R^{n \times m}$ – the continuous vector-valued functions which describe the dynamics of the process,

$$\mathbf{f}(\mathbf{x}(t)) = \begin{bmatrix} f_1(\mathbf{x}(t)) \\ f_2(\mathbf{x}(t)) \\ ... \\ f_n(\mathbf{x}(t)) \end{bmatrix}, \; f_i : R^n \to R, \; i = 1...n,$$
$$\mathbf{B}(\mathbf{x}(t)) = \begin{bmatrix} \mathbf{b}_1^T(\mathbf{x}(t)) \\ \mathbf{b}_2^T(\mathbf{x}(t)) \\ ... \\ \mathbf{b}_n^T(\mathbf{x}(t)) \end{bmatrix}, \; \mathbf{b}_i^T(\mathbf{x}(t)) = [b_{i1}(\mathbf{x}(t)) \quad b_{i2}(\mathbf{x}(t)) \quad ... \quad b_{im}(\mathbf{x}(t))], \tag{2}$$

and $\mathbf{u}(t) = [u_1(t) \quad u_2(t) \quad ... \quad u_m(t)]^T$ – the control signal vector applied to the process. The actuators and measuring instrumentation are included in the nonlinear process.

The i^{th} fuzzy control rule in the rule base of the T-S FC, referred to as $\mathrm{R}^i, i = 1...n_{RB}, n_{RB} \geq 2$ (n_{RB} – the number of rules), is expressed as

$$\begin{aligned} &\mathrm{R}^i : \text{IF } x_1(t) \text{ IS } \tilde{X}_{1i} \text{ AND } x_2(t) \text{ IS } \tilde{X}_{2i} \text{ AND} ... \text{ AND } x_n(t) \text{ IS } \tilde{X}_{ni} \\ &\quad \text{THEN } \mathbf{u} = \mathbf{u}^i(\mathbf{x}(t)), \; i = 1...n_{RB}, \end{aligned} \tag{3}$$

where $\tilde{X}_{ki}$ are fuzzy sets with the universes X_{ki}, $k = 1...n$, corresponding to the linguistic terms (LTs) afferent to the state variables x_i, $u^i(\mathbf{x})$ is the control signal produced by the rule R^i with the firing strength $\alpha^i = \alpha^i(\mathbf{x})$

$$\begin{aligned} &\alpha^i(\mathbf{x}) = \text{AND}(\mu_{\tilde{X}_{1i}}(x_1), \mu_{\tilde{X}_{2i}}(x_2), ..., \mu_{\tilde{X}_{ni}}(x_n)), \forall \mathbf{x} \in X \;\; \exists i = 1...n_{RB}, \\ &0 < \alpha^i(\mathbf{x}) \leq 1, \end{aligned} \tag{4}$$

where the function AND is a t-norm, and $\mu_{\tilde{X}_{ki}}$ are the membership functions of the fuzzy sets of LTs $\tilde{X}_{ki}$. An active region of the rule R^i is defined as the set $X_i^A = \{\mathbf{x} \in X \mid \alpha^i(\mathbf{x}) \neq 0\}$.

The control signal vector $\mathbf{u}$ is a function of α^i and $\mathbf{u}^i$ which depends on the inference engine and on the defuzzification method. The weighted sum defuzzification method produces the control signal vector $\mathbf{u}(\mathbf{x}(t))$, which will also be referred to as $\mathbf{u}(t)$ in the sequel for the sake of simplicity:

$$\mathbf{u}(\mathbf{x}(t)) = \frac{\sum_{i=1}^{n_{RB}} \alpha^i(\mathbf{x}(t))\mathbf{u}^i(\mathbf{x}(t))}{\sum_{i=1}^{n_{RB}} \alpha^i(\mathbf{x}(t))}. \tag{5}$$

3 Stability Analysis Theorem

Let the process be characterized by the state-space model defined in (1), and let V be a radially unbounded function $V: X \to R,\ V(\mathbf{x}) > 0, \forall \mathbf{x} \in X, \mathbf{x} \neq \mathbf{0}$. The first difference of the function $V(\mathbf{x}(t))$ along the trajectory of (1), denoted by $\Delta V(\mathbf{x}(t))$, is

$$\Delta V(\mathbf{x}(t)) = V(\mathbf{x}(t+1)) - V(\mathbf{x}(t)). \tag{6}$$

Using the notation $V_i(\mathbf{x}(t))$ for the Lyapunov function candidate $V(\mathbf{x}(t))$, which is considered along the trajectory of the system (1) for $\mathbf{u}(t) = \mathbf{u}^i(\mathbf{x}(t))$, the first difference of $V_i(\mathbf{x}(t))$ is $\Delta V_i(\mathbf{x}(t))$:

$$\Delta V_i(\mathbf{x}(t)) = V_i(\mathbf{x}(t+1)) - V_i(\mathbf{x}(t)), \forall\, \mathbf{x} \in X_i^A. \tag{7}$$

The following original stability analysis theorem is derived on the basis of Lyapunov's theorem for discrete-time systems using the formulation given in [27]:

Theorem 1. Let the FCS be described by the discrete-time input affine MIMO system modelled in (1), the T-S FC characterized by equations (3)–(7), and $\mathbf{x} = \mathbf{0}$ be an equilibrium point of (1). If there exists

$$V: X \to R,\ V(\mathbf{x}(t)) = \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t), \text{ continuous in } \mathbf{x}, \tag{8}$$

where $\mathbf{P} \in R^{n \times n}$ is a positive definite matrix such that

$$\Delta V_i(\mathbf{x}(t)) \leq 0,\ \forall\, \mathbf{x} \in X_i^A,\ i = 1...n_{RB}, \tag{9}$$

then $\mathbf{x} = \mathbf{0}$ is stable.

Proof. The hypotheses of the theorem result in

$$\Delta V_i(\mathbf{x}(t)) = V_i(\mathbf{x}(t+1)) - V_i(\mathbf{x}(t)) < 0,\ \forall \mathbf{x} \in X_i^A,\ i = 1...n_{RB}. \tag{10}$$

The term $\mathbf{x}(t+1)$ is next substituted from (1) into (10):

$$\begin{aligned}
&\Delta V_i(\mathbf{x}(t)) = V_i(\mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)) - V_i(\mathbf{x}(t)) \\
&= [\mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)]^T\ \mathbf{P}\,[\mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)] \\
&- \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t) = [\mathbf{f}^T(\mathbf{x}(t)) + (\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))]\,\mathbf{P}\,[\mathbf{f}(\mathbf{x}(t)) \\
&+ \mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)] - \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t) = [\mathbf{f}^T(\mathbf{x}(t))\mathbf{P} \\
&+ (\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}][\mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)] - \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t) \\
&= \mathbf{f}^T(\mathbf{x}(t))\mathbf{P}\mathbf{f}(\mathbf{x}(t)) + \mathbf{f}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t) \\
&+ (\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{f}(\mathbf{x}(t)) + (\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t) \\
&- \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t) < 0,\ i = 1...n_{RB}.
\end{aligned} \tag{11}$$

The multiplication of (11) by $\alpha^i(\mathbf{x}(t))$ and the calculation of the sum result in

$$\begin{aligned}
&[\mathbf{f}^T(\mathbf{x}(t))\mathbf{P}\mathbf{f}(\mathbf{x}(t))]\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t)) + [\mathbf{f}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))]\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))\mathbf{u}^i(t) \\
&+ \sum_{i=1}^{n_{RB}}[(\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{f}(\mathbf{x}(t))\alpha^i(\mathbf{x}(t))] \\
&+ \sum_{i=1}^{n_{RB}}[(\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)\alpha^i(\mathbf{x}(t))] \\
&- [\mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t)]\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t)) < 0.
\end{aligned} \tag{12}$$

Equation (12) is divided by $\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t)) > 0$ and equation (5) is applied to transform the resulted sums as follows:

$$\begin{aligned}
&\mathbf{f}^T(\mathbf{x}(t))\mathbf{P}\mathbf{f}(\mathbf{x}(t)) + \mathbf{f}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}(t) + \mathbf{u}^T(t)\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{f}(\mathbf{x}(t)) \\
&+ \frac{\sum_{i=1}^{n_{RB}}[(\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)\alpha^i(\mathbf{x}(t))]}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))} - \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t) < 0.
\end{aligned} \tag{13}$$

The expression of $\Delta V(\mathbf{x}(t))$ results from (1) and (6):

$$\begin{aligned}
&\Delta V(\mathbf{x}(t)) = V(\mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}(t)) - V(\mathbf{x}(t)) \\
&= [\mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}(t)]^T\ \mathbf{P}\,[\mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}(t)] - \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t) \\
&= [\mathbf{f}^T(\mathbf{x}(t))\mathbf{P} + \mathbf{u}^T(t)\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}][\mathbf{f}(\mathbf{x}(t)) + \mathbf{B}(\mathbf{x}(t))\mathbf{u}(t)] - \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t) \\
&= \mathbf{f}^T(\mathbf{x}(t))\mathbf{P}\mathbf{f}(\mathbf{x}(t)) + \mathbf{f}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}(t) + \mathbf{u}^T(t)\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{f}(\mathbf{x}(t)) \\
&+ \mathbf{u}^T(t)\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}(t) - \mathbf{x}^T(t)\,\mathbf{P}\,\mathbf{x}(t) < 0.
\end{aligned} \tag{14}$$

In the following we prove that

$$\mathbf{u}^T(t)\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}(t) \le \frac{\sum_{i=1}^{n_{RB}}[(\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)\alpha^i(\mathbf{x}(t))]}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))}. \quad (15)$$

The terms $\mathbf{u}^T(t)\mathbf{B}^T(\mathbf{x}(t))\,\mathbf{P}\,\mathbf{B}(\mathbf{x}(t))\mathbf{u}(t)$ and $(\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\,\mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)$ are quadratic forms because the matrix

$$\mathbf{M}(\mathbf{x}(t)) = \mathbf{B}^T(\mathbf{x}(t))\,\mathbf{P}\,\mathbf{B}(\mathbf{x}(t)) \in R^{m\times m} \quad (16)$$

is symmetric. The matrix $\mathbf{M}(\mathbf{x}(t))$ has the following spectral decomposition (Jordan decomposition):

$$\mathbf{M}(\mathbf{x}(t)) = \mathbf{\Gamma}(\mathbf{x}(t))\,\mathbf{\Lambda}(\mathbf{x}(t))\,\mathbf{\Gamma}^T(\mathbf{x}(t)) = \sum_{i=1}^{m}[\lambda_i(\mathbf{x}(t))\boldsymbol{\gamma}_i(\mathbf{x}(t))\boldsymbol{\gamma}_i^T(\mathbf{x}(t))], \quad (17)$$

where

$$\mathbf{\Lambda}(\mathbf{x}(t)) = \mathrm{diag}(\lambda_1(\mathbf{x}(t)),\lambda_2(\mathbf{x}(t)),...,\lambda_m(\mathbf{x}(t))), \quad (18)$$

$\lambda_j(\mathbf{x}(t)),\ j=1...m,$ are the eigenvalues of $\mathbf{M}(\mathbf{x}(t))$. The orthogonal matrix

$$\mathbf{\Gamma}(\mathbf{x}(t)) = [\boldsymbol{\gamma}_1(\mathbf{x}(t)) \quad \boldsymbol{\gamma}_2(\mathbf{x}(t)) \quad ... \quad \boldsymbol{\gamma}_m(\mathbf{x}(t))] \quad (19)$$

consists of the eigenvectors $\boldsymbol{\gamma}_i(\mathbf{x}(t))$ of $\mathbf{M}(\mathbf{x}(t))$.

Considering the linear transformation

$$\mathbf{u} \mapsto \Gamma^T\mathbf{u} = \mathbf{w} = [w_1 \quad w_2 \quad ... \quad w_m]^T,\ w_j = \boldsymbol{\gamma}_j^T\mathbf{u},\ j=1...m, \quad (20)$$

equations (16), (17) and (20) lead to

$$\begin{aligned}&\mathbf{u}^T(t)\,\mathbf{B}^T(\mathbf{x}(t))\,\mathbf{P}\,\mathbf{B}(\mathbf{x}(t))\,\mathbf{u}(t) = \mathbf{u}^T(t)\,\mathbf{M}(\mathbf{x}(t))\,\mathbf{u}(t)\\ &= \mathbf{u}^T(t)\,\mathbf{\Gamma}(\mathbf{x}(t))\,\mathbf{\Lambda}(\mathbf{x}(t))\,\mathbf{\Gamma}^T(\mathbf{x}(t))\,\mathbf{u}(t) = (\mathbf{w}(t))^T\,\mathbf{\Lambda}(\mathbf{x}(t))\,\mathbf{w}(t)\\ &= \sum_{j=1}^{m}[\lambda_j(\mathbf{x}(t))w_j^2(\mathbf{x}(t))].\end{aligned} \quad (21)$$

The expression of w_j from (20) is substituted into (21) leading to the following result:

$$\mathbf{u}^T(t)\,\mathbf{B}^T(\mathbf{x}(t))\,\mathbf{P}\,\mathbf{B}(\mathbf{x}(t))\,\mathbf{u}(t) = \sum_{j=1}^{m}\{\lambda_j(\mathbf{x}(t))[\boldsymbol{\gamma}_j^T(\mathbf{x}(t))\mathbf{u}(t)]^2\}. \quad (22)$$

The following relationship results from (22) by replacing $\mathbf{u}(t)$ with $\mathbf{u}^i(t)$:

$$(\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\,\mathbf{P}\,\mathbf{B}(\mathbf{x}(t))\,\mathbf{u}^i(t) = \sum_{j=1}^{m}\{\lambda_j(\mathbf{x}(t))[\boldsymbol{\gamma}_j^T(\mathbf{x}(t))\mathbf{u}^i(t)]^2\}. \quad (23)$$

The expression of $\mathbf{u}(t)$ is substituted from (5) into the right-hand side of (22), and the sums are manipulated as follows:

$$\mathbf{u}^T(t)\,\mathbf{B}^T(\mathbf{x}(t))\,\mathbf{P}\,\mathbf{B}(\mathbf{x}(t))\,\mathbf{u}(t) = \sum_{j=1}^{m}\{\lambda_j(\mathbf{x}(t))[\boldsymbol{\gamma}_j^T(\mathbf{x}(t))\left(\frac{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))\mathbf{u}^i(t)}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))}\right)]^2\} \tag{24}$$

$$= \sum_{j=1}^{m}\{\frac{\lambda_j(\mathbf{x}(t))}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))}[\frac{\left(\sum_{i=1}^{n_{RB}}[\alpha^i(\mathbf{x}(t))\boldsymbol{\gamma}_j^T(\mathbf{x}(t))\mathbf{u}^i(t)]\right)^2}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))}]\}.$$

The application of Cauchy-Buniakovski-Schwarz's inequality to the second fraction in the right-hand side of (24) leads to

$$\mathbf{u}^T(t)\,\mathbf{B}^T(\mathbf{x}(t))\,\mathbf{P}\,\mathbf{B}(\mathbf{x}(t))\,\mathbf{u}(t) \le \sum_{j=1}^{m}\left\{\frac{\lambda_j(\mathbf{x}(t))}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))}\sum_{i=1}^{n_{RB}}\{\alpha^i(\mathbf{x}(t))[\boldsymbol{\gamma}_j^T(\mathbf{x}(t))\mathbf{u}^i(t)]^2\}\right\}, \tag{25}$$

and next to

$$\mathbf{u}^T(t)\,\mathbf{B}^T(\mathbf{x}(t))\,\mathbf{P}\,\mathbf{B}(\mathbf{x}(t))\,\mathbf{u}(t) \le \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_{RB}}\{\lambda_j(\mathbf{x}(t))[\boldsymbol{\gamma}_j^T(\mathbf{x}(t))\mathbf{u}^i(t)]^2\alpha^i(\mathbf{x}(t))\}}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))}. \tag{26}$$

The multiplication of (23) by $\alpha^i(\mathbf{x}(t))$, the calculation of the sum and the division by $\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t)) > 0$ result in the expression of the right-hand side of (15):

$$\frac{\sum_{i=1}^{n_{RB}}[(\mathbf{u}^i(t))^T\mathbf{B}^T(\mathbf{x}(t))\mathbf{P}\mathbf{B}(\mathbf{x}(t))\mathbf{u}^i(t)\alpha^i(\mathbf{x}(t))]}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))} = \frac{\sum_{i=1}^{n_{RB}}\sum_{j=1}^{m}\{\lambda_j(\mathbf{x}(t))[\boldsymbol{\gamma}_j^T(\mathbf{x}(t))\mathbf{u}^i(t)]^2\alpha^i(\mathbf{x}(t))\}}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))}$$

$$= \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_{RB}}\{\lambda_j(\mathbf{x}(t))[\boldsymbol{\gamma}_j^T(\mathbf{x}(t))\mathbf{u}^i(t)]^2\alpha^i(\mathbf{x}(t))\}}{\sum_{i=1}^{n_{RB}}\alpha^i(\mathbf{x}(t))}. \tag{27}$$

Therefore equations (26) and (27) demonstrate the inequality (15). The inequality (15) is applied to (13) and (14), which result finally in

$$\Delta V(\mathbf{x}(t)) \le 0. \tag{28}$$

Therefore, the equilibrium point at the origin will be stable. The proof is now complete. Concluding, Theorem 1 offers sufficient stability conditions concerning the class of fuzzy control systems defined in Section 2.

4 Case Study

The case study applies Theorem 1 to the design of T-S FCSs dedicated to the level control of spherical three tank systems. The process structure presented in Figure 2 illustrates the three spherical tanks, T1, T2 and T3, with the same radius R, in series connection by two connecting pipes of inner area S. All three tanks are equipped with piezo-resistive pressure sensors (viz. the level sensors LS1, LS2 and LS3) to measure the liquid levels. The FC actuates (by means of the pumps P1 and P2) the flow rates q_{p1} and q_{p2} in order to control independently the levels in the tanks T1 (h_1) and T2 (h_2), and the following constraints imposed to the levels result from the process structure:

$$0 < h_i < 2R,\ i = 1...3. \tag{29}$$

A typical control objective pointed out in [13] is to keep the liquid levels h_1 and h_2 at the imposed levels while the liquid level in the tank T3 (h_3) is uncontrollable. The level sensors give the measured levels h_{m1}, h_{m2} and h_{m3} used by the FC. The connecting pipes and tanks are equipped with manually adjustable valves and outlets to simulate clogs and leaks.

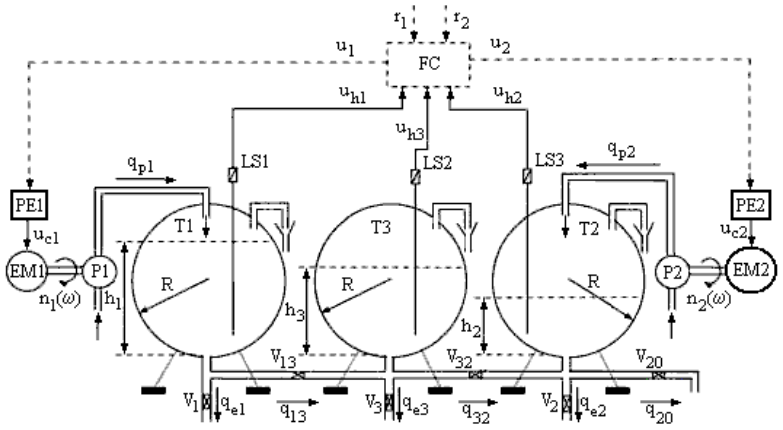


Figure 2
FCS and process structure

The simplified FCS structure is presented in Figure 3, where $\mathbf{d}=[d_{e1}=1-\mu_{e1} \quad d_{e2}=1-\mu_{e2} \quad d_{e3}=1-\mu_{e3}]^T$ is the disturbance input vector, $\mathbf{r}=[r_1 \quad r_2]^T$ is the reference input vector, e_i

$$e_i = r_i - u_{hi} = r_i - k_{mi}h_i,\ i=1...2, \tag{30}$$

are the control errors grouped in the control error vector $\mathbf{e}=[e_1 \quad e_2]^T$, μ_{e1}, μ_{e2} and μ_{e3} are the deterministic disturbance inputs are the positions of the valves V_1, V_2 and V_3, $0\le\mu_{e1},\mu_{e2},\mu_{e3}\le 1$, with the notations 0 for the completely close valves and 1 for the completely open valves, and k_{mi}, $i=1...3$, are the sensor gains.

Using the notation $A(h_i)=\pi h_i(2R-h_i),\ i=1...3$, for the transversal section area of sphere (i.e., tank) i at height (liquid level) h_i, the first principle mathematical model of the process proposed in [13] is discretized, and the following disturbed discrete-time process model is used in T-S FC design:

$$\begin{aligned}
e_1(t+1) &= \frac{k_{m1}(t)}{A(u_{h1}(t)/k_{m1}(t))}[S\mathrm{sgn}(\frac{r_1-e_1(t)}{k_{m1}(t)}-\frac{u_{h3}(t)}{k_{m3}(t)})\\
&\sqrt{2g\,|\frac{r_1-e_1(t)}{k_{m1}(t)}-\frac{u_{h3}(t)}{k_{m3}(t)}|}+d_{e1}S_V\sqrt{2g\frac{r_1-e_1(t)}{k_{m1}(t)}}-k_{p1}(t)u_1(t)],\\
e_2(t+1) &= \frac{k_{m2}(t)}{A(u_{h2}(t)/k_{m2}(t))}[-S\mathrm{sgn}(\frac{u_{h3}(t)}{k_{m3}(t)}-\frac{r_2-e_2(t)}{k_{m2}(t)})\\
&\sqrt{2g\,|\frac{u_{h3}(t)}{k_{m3}(t)}-\frac{r_2-e_2(t)}{k_{m2}(t)}|}+d_{e2}S_V\sqrt{2g\frac{r_2-e_2(t)}{k_{m2}(t)}}\\
&+S_V\sqrt{2g\frac{r_2-e_2(t)}{k_{m2}(t)}}-k_{p2}(t)u_2(t)],\\
u_{h3}(t+1) &= \frac{k_{m3}(t)}{A(u_{h3}(t)/k_{m3}(t))}[S\mathrm{sgn}(\frac{u_{h1}(t)}{k_{m1}(t)}-\frac{u_{h3}(t)}{k_{m3}(t)})\sqrt{2g\,|\frac{u_{h1}(t)}{k_{m1}(t)}-\frac{u_{h3}(t)}{k_{m3}(t)}|}\\
&-S\mathrm{sgn}(\frac{u_{h3}(t)}{k_{m3}(t)}-\frac{u_{h2}(t)}{k_{m2}(t)})\sqrt{2g\,|\frac{u_{h3}(t)}{k_{m3}(t)}-\frac{u_{h2}(t)}{k_{m2}(t)}|}-d_{e3}S_V\sqrt{2g\frac{u_{h3}(t)}{k_{m3}(t)}}],\\
h_1(t) &= r_1-e_1(t),\\
h_2(t) &= r_2-e_2(t),
\end{aligned} \tag{31}$$

where S_V is the inner area of outflow pipes, and k_{p1} and k_{p2} are the actuator gains. The relation between the variables in the model (31) and the variables in the discrete-time input affine MIMO state-space model (1) are

$$\mathbf{x}=[x_1 \quad x_2]^T=[e_1 \quad e_2]^T=\mathbf{e},\ \mathbf{y}=[y_1=h_1 \quad y_2=h_2]^T. \tag{32}$$

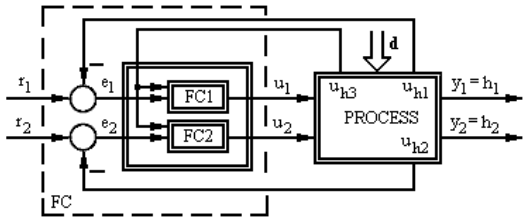


Figure 3
Simplified FCS structure

Figure 3 shows that the MIMO FC consists of two separately designed T-S FCs, FC1 and FC2. The fuzzification in FC is done using the input membership functions presented in Figure 4.

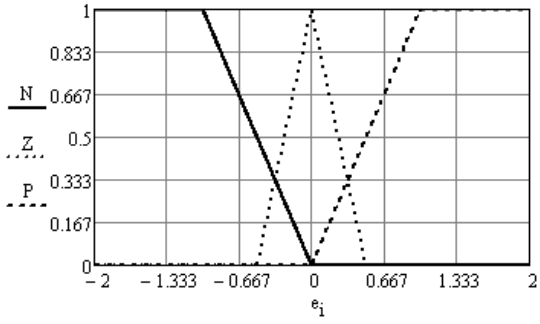


Figure 4
Input membership functions

The inference engine employs the MIN t-norm for the AND operator as specified in Section 2. The inference engine is assisted by the following complete rule base as $n_{RB} = 9$:

$$
\begin{aligned}
&R^1: \text{IF } e_1 \text{ IS P AND } e_2 \text{ IS P THEN } \mathbf{u} = \mathbf{u}_1,\\
&R^2: \text{IF } e_1 \text{ IS N AND } e_2 \text{ IS N THEN } \mathbf{u} = \mathbf{u}_2,\\
&R^3: \text{IF } e_1 \text{ IS N AND } e_2 \text{ IS P THEN } \mathbf{u} = \mathbf{u}_3,\\
&R^4: \text{IF } e_1 \text{ IS P AND } e_2 \text{ IS N THEN } \mathbf{u} = \mathbf{u}_4,\\
&R^5: \text{IF } e_1 \text{ IS P AND } e_2 \text{ IS Z THEN } \mathbf{u} = \mathbf{u}_5,\\
&R^6: \text{IF } e_1 \text{ IS Z AND } e_2 \text{ IS P THEN } \mathbf{u} = \mathbf{u}_6,\\
&R^7: \text{IF } e_1 \text{ IS N AND } e_2 \text{ IS Z THEN } \mathbf{u} = \mathbf{u}_7,\\
&R^8: \text{IF } e_1 \text{ IS Z AND } e_2 \text{ IS N THEN } \mathbf{u} = \mathbf{u}_8,\\
&R^9: \text{IF } e_1 \text{ IS Z AND } e_2 \text{ IS Z THEN } \mathbf{u} = \mathbf{u}_9,
\end{aligned}
\tag{33}
$$

where $\mathbf{u} = [u_1 \quad u_2]^T$, and the rule consequents $\mathbf{u}_k = [u_1^i \quad u_2^i]^T$, $i = 1...9$, are determined as follows on the basis of Theorem 1. More inputs can be considered, but this leads to the complication of the FCS structure and of the design, and rule base reduction techniques should be used [28]-[31]. The Lyapunov function candidate

$$V : R^2 \to R,\ V(\mathbf{e}) = 0.5e_1^2 + 0.5e_2^2 \tag{34}$$

is chosen in order to design stable FCSs for this MIMO process. For $\mathbf{d} = \mathbf{0}$ the time derivative of $V(\mathbf{e})$ along the trajectory of (31), referred to as $\dot{V}(\mathbf{e})$, is

$$\begin{aligned} &V(t+1) = e_1(t)e_1(t+1) + e_2e_2(t+1) = \frac{e_1(t)k_{m1}(t)}{A(u_{h1}(t)/k_{m1}(t))} \\ &[S\,\mathrm{sgn}(\frac{r_1 - e_1(t)}{k_{m1}(t)} - \frac{u_{h3}(t)}{k_{m3}(t)})\sqrt{2g\,|\frac{r_1 - e_1(t)}{k_{m1}(t)} - \frac{u_{h3}(t)}{k_{m3}(t)}|} - k_{p1}(t)u_1(t)] \\ &+ \frac{e_2(t)k_{m2}(t)}{A(u_{h2}(t)/k_{m2}(t))}[-S\,\mathrm{sgn}(\frac{u_{h3}(t)}{k_{m3}(t)} - \frac{r_2 - e_2(t)}{k_{m2}(t)})\sqrt{2g\,|\frac{u_{h3}(t)}{k_{m3}(t)} - \frac{r_2 - e_2(t)}{k_{m2}(t)}|} \\ &+ S_V\sqrt{2g\,\frac{r_2 - e_2(t)}{k_{m2}(t)}} - k_{p2}(t)u_2(t)]. \end{aligned} \tag{35}$$

The control laws in the rule consequents of MIMO FC are designed to fulfil the condition (9) in Theorem 1, which leads to

$$V_i(t+1) = F(\mathbf{e}(t)) + \mathbf{g}^T(\mathbf{e}(t))\mathbf{u}_i(\mathbf{e}(t)) \le 0,\ i = 1...9. \tag{36}$$

The condition (36) is important because it supports the formulation of the rule base of MIMO FC summarized in Table 1 and proved in Appendix 1. But the controller design depends on the process, and different expressions of Lyapunov function candidates can be used in other applications [32]-[41].

Concluding, Theorem 1 is verified. Therefore the T-S FCS designed in this section is stable.

The values of process parameters considered in this case study are

$$\begin{aligned} &S = 0.005\,\mathrm{m}^2,\ S_V = 0.005\,\mathrm{m}^2,\ R = 1\,\mathrm{m},\ g = 9.8\,\mathrm{m/s}^2, \\ &k_{p1} = k_{p2} = 0.094\,\mathrm{m}^3/(\mathrm{V\,s}),\ k_{m1} = k_{m2} = 1\,\mathrm{V/m}, \end{aligned} \tag{37}$$

and the sampling period was set to $T_s = 0.01\,\mathrm{s}$.

Three digital simulation scenarios were considered in order to illustrate the stable behaviour of our T-S FCS scenario 1 (reference inputs $r_1 = 1.5\,\mathrm{m}$ and $r_2 = 1.5\,\mathrm{m}$, and initial conditions $h_1(0) = 0.1\,\mathrm{m}$, $h_2(0) = 1.9\,\mathrm{m}$ and $h_3(0) = 1.5\,\mathrm{m}$ applied to T-S FCS), scenario 2 ($r_1 = 0.5\,\mathrm{m}$, $r_2 = 1.5\,\mathrm{m}$, $h_1(0) = 0.1\,\mathrm{m}$, $h_2(0) = 1.1\,\mathrm{m}$ and

$h_3(0)=1\,\mathrm{m}$ applied to T-S FCS) and scenario 3 ($r_1=0.5\,\mathrm{m}$, $r_2=1.5\,\mathrm{m}$, $h_1(0)=1\,\mathrm{m}$, $h_2(0)=0.1\,\mathrm{m}$ and $h_3(0)=1\,\mathrm{m}$ applied to T-S FCS). The trapezoidal function defined in [13] models the variations of deterministic disturbance inputs.

Table 1
Rule base of MIMO FC

R^i	Premise		Consequent	
	e_1	e_2	u_1	u_2
R^1	P	P	$2S\sqrt{gR}/k_{p1}$	$(2S\sqrt{gR}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$
R^2	N	N	$-2S\sqrt{gR}/k_{p1}$	$(-2S\sqrt{gR}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$
R^3	N	P	$-2S\sqrt{gR}/k_{p1}$	$(2S\sqrt{gR}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$
R^4	P	N	$2S\sqrt{gR}/k_{p1}$	$(-2S\sqrt{gR}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$
R^5	P	Z	$2S\sqrt{gR}/k_{p1}$	$(e_2-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}})\sqrt{2g\lvert\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}}\rvert}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$
R^6	Z	P	$(e_1+S\,\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}})\sqrt{2g\lvert\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}}\rvert})/k_{p1}$	$(2S\sqrt{gR}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$
R^7	N	Z	$-2S\sqrt{gR}/k_{p1}$	$(e_2-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}})\sqrt{2g\lvert\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}}\rvert}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$
R^8	Z	N	$(e_1+S\,\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}})\sqrt{2g\lvert\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}}\rvert})/k_{p1}$	$(-2S\sqrt{gR}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$
R^9	Z	Z	$(e_1+S\,\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}})\sqrt{2g\lvert\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}}\rvert})/k_{p1}$	$(e_2-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}})\sqrt{2g\lvert\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}}\rvert}+S_V\sqrt{2g\frac{r_2-e_2}{k_{m2}}})/k_{p2}$

The digital simulation results obtained for the simulation scenarios 1, 2 and 3 are presented in Figures 5, 6 and 7, respectively. These results highlight the stable behaviour of the T-S FCS for different inputs and initial conditions.

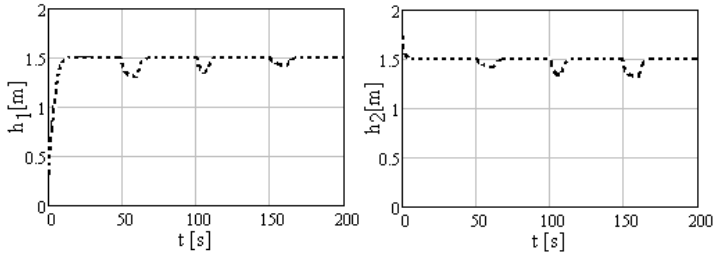


Figure 5
Digital simulation results (scenario 1)

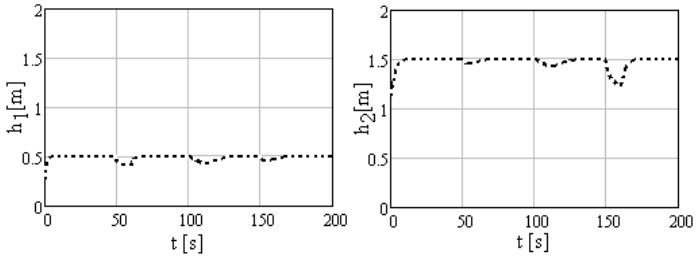


Figure 6
Digital simulation results (scenario 2)

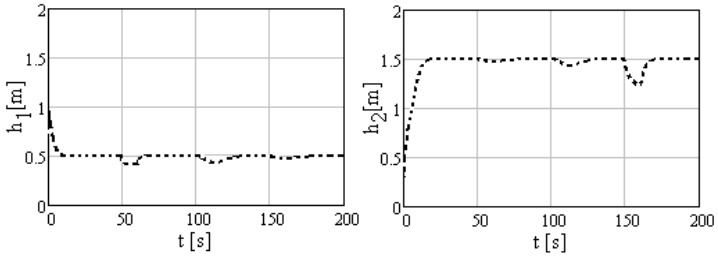


Figure 7
Digital simulation results (scenario 3)

Conclusions

A new stability approach to nonlinear MIMO process characterized by discrete-time input affine state-space models has been proposed. The approach has been applied to the stable design of a T-S FC for the level control of spherical three tank systems. Future research will be focused on the refinement of the stability analysis theorem in order to become less dependent on the process.

Acknowledgement

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS – UEFISCDI, project number PN-II-ID-PCE-2011-3-0109. The cooperation between the Óbuda University, Budapest, Hungary, the University of Ljubljana, Slovenia, and the "Politehnica" University of Timisoara, Romania, in the framework of the Hungarian-Romanian and Slovenian-Romanian Intergovernmental S & T Cooperation Programs is acknowledged.

References

[1] L. Wu, Z. Su, P. Shi P, J. Qiu: A New Approach to Stability Analysis and Stabilization of Discrete-Time T-S Fuzzy Time-Varying Delay Systems, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 41, No. 1, 2011, pp. 273-286

[2] H. K. Lam: LMI-based Stability Analysis for Fuzzy-Model-based Control Systems Using Artificial T-S Fuzzy Model, IEEE Transactions on Fuzzy Systems, Vol. 19, No. 3, 2011, pp. 505-513

[3] M. Narimani, H. K. Lam, R. Dilmaghani, C. Wolfe: LMI-based Stability Analysis of Fuzzy-Model-based Control Systems Using Approximated Polynomial Membership Functions, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 41, No. 3, 2011, pp. 713-724

[4] H. K. Lam, M. Narimani: Quadratic-Stability Analysis of Fuzzy-Model-based Control Systems Using Staircase Membership Functions, IEEE Transactions on Fuzzy Systems, Vol. 18, No. 1, 2010, pp. 125-137

[5] S. Jafarzadeh, M. S. Fadali, A. H. Sonbol: Stability Analysis and Control of Discrete Type-1 and Type-2 TSK Fuzzy Systems: Part I. Stability Analysis, IEEE Transactions on Fuzzy Systems, Vol. 19, No. 6, 2011, pp. 989-1000

[6] S. Jafarzadeh, M. S. Fadali, A. H. Sonbol: Stability Analysis and Control of Discrete Type-1 and Type-2 TSK Fuzzy Systems: Part II. Control Design, IEEE Transactions on Fuzzy Systems, Vol. 19, No. 6, 2011, pp. 1001-1013

[7] M. S. Fadali, S. Jafarzadeh: Fuzzy TSK Positive Systems: Stability and Control, Proceedings of American Control Conference (ACC 2011), San Francisco, CA, 2011, pp. 4964-4969

[8] I. Škrjanc, S. Blažič, D. Matko: Direct Fuzzy Model-Reference Adaptive Control, International Journal of Intelligent Systems, Vol. 17, No. 10, 2002, pp. 943-963

[9] M. Kratmüller: Combining Fuzzy/Wavelet Adaptive Error Tracking Control Design, Acta Polytechnica Hungarica, Vol. 7, No. 4, 2010, pp. 115-137

[10] Y.-S. Huang, D.-S. Xiao, X.-X. Chen, Q.-X. Zhu, Z.-W. Wang: H_∞ Tracking-based Decentralized Hybrid Adaptive Output Feedback Fuzzy

Control for a Class of Large-Scale Nonlinear Systems, Fuzzy Sets and Systems, Vol. 171, No. 1, 2011, pp. 72-92

[11] M. A. Khanesar, M. Teshnehlab: Model Reference Fuzzy Control of Nonlinear Dynamical Systems Using an Optimal Observer, Acta Polytechnica Hungarica, Vol. 8, No. 4, 2011, pp. 35-54

[12] R.-E. Precup, S. Preitl, I. J. Rudas, M. L. Tomescu, J. K. Tar: Design and Experiments for a Class of Fuzzy Controlled Servo Systems. IEEE/ASME Transactions on Mechatronics, Vol. 13, No. 1, 2008, pp. 22-35

[13] R.-E. Precup, M.-L. Tomescu, E. M. Petriu, S. Preitl, J. Fodor, D. Bărbulescu: Stability Analysis of a Class of MIMO Fuzzy Control Systems. Proceedings of 2010 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2010), Barcelona, Spain, 2010, pp. 2885-2890

[14] R.-E. Precup, M.-L. Tomescu, S. Preitl, E. M. Petriu, C.-A. Dragoş: Stability Analysis of Fuzzy Logic Control Systems for a Class of Nonlinear SISO Discrete-Time Systems, Preprints of 18th IFAC World Congress, Milano, Italy, 2011, pp. 13612-13617

[15] R.-E. Precup, E. M. Petriu, C.-A. Dragoş, R.-C. David: Stability Aanalysis Results Concerning the Fuzzy Control of a Class of Nonlinear Time-Varying Systems, Theory and Applications of Mathematics & Computer Science, Vol. 1, No. 1, 2011, pp. 2-10

[16] S. Blažič, I. Škrjanc, D. Matko: Globally Stable Direct Fuzzy Model Reference Adaptive Control, Fuzzy Sets and Systems, Vol. 139, No. 1, 2003, pp. 3-33

[17] A. Palcu, S. Nădăban, A. Şandru: Some Remarks on the Boson Mass Spectrum in a 3-3-1 Gauge Model, Romanian Journal of Physics, Vol. 56, Nos. 5-6, 2011, pp. 673-681

[18] D. Hládek, J. Vaščák, P. Sinčák: Multi-Robot Control System for Pursuit-Evasion Problem, Journal of Electrical Engineering, Vol. 60, No. 3, 2009, pp. 143-148

[19] Gy. Hermann, J. K. Tar, K. R. Kozlowsky: Design of a Planar High Precision Motion Stage, in: Robot Motion and Control 2009, K. R. Kozlowsky (Ed.), Springer-Verlag, Berlin, Heidelberg, 2009, pp. 371-379

[20] R. E. Haber, R. M. del Toro, A. Gajate: Optimal Fuzzy Control System Using the Cross-Entropy Method. A Case Study of a Drilling Process, Information Sciences, Vol. 180, No. 14, 2010, pp. 2777-2792

[21] J. A. Iglesias, P. Angelov, A. Ledezma, A. Sanchis: Evolving Classification of Agents' Behaviors: A General Approach, Evolving Systems, Vol. 1, No. 3, 2010, pp. 161-171

[22] Gy. Mester: Intelligent Mobile Robot Motion Control in Unstructured Environments, Acta Polytechnica Hungarica, Vol. 7, No. 4, 2010, pp. 153-165

[23] K. Y. Chan, C. K. Kwong, T. S. Dillon, Y. C. Tsim: Reducing Overfitting in Manufacturing Process Modeling Using a Backward Elimination-based Genetic Programming, Applied Soft Computing, Vol. 11, No. 2, 2011, 1648-1656

[24] N. J. Cotton, B. M. Wilamowski: Compensation of Nonlinearities Using Neural Networks Implemented on Inexpensive Microcontrollers, IEEE Transactions on Industrial Electronics, Vol. 58, No. 3, 2011, pp. 733-740

[25] N. Kasabov, H. N. A. Hamed: Quantum-Inspired Particle Swarm Optimisation for Integrated Feature and Parameter Optimisation of Evolving Spiking Neural Networks, International Journal of Artificial Intelligence, Vol. 7, No. A11, 2011, pp. 114-124

[26] O. Linda, M. Manic: Uncertainty-Robust Design of Interval Type-2 Fuzzy Logic Controller for Delta Parallel Robot, IEEE Transactions on Industrial Informatics, Vol. 7, No. 11, 2011, pp. 661-670

[27] J. J. E. Slotine, W. Li: Applied Nonlinear Control, Prentice-Hall, Englewood Cliffs, NJ, 1991

[28] P. Baranyi, K. F. Lei, Y. Yam: Complexity Reduction of Singleton-based Neuro-Fuzzy Algorithm, Proceedings of IEEE International Conference System, Man, and Cybernetics (SMC'00), Nashville, TN, USA, 2000, pp. 2503-2508

[29] P. Baranyi, D. Tikk, Y. Yam, R. J. Patton: From Differential Equations to PDC Controller Design via Numerical Transformation, Computers in Industry, Vol. 51, No. 3, 2003, pp. 281-297

[30] Zs. Cs. Johanyák, Sz. Kovács: Fuzzy Rule Interpolation Based on Polar Cuts, in: Computational Intelligence, Theory and Applications, B. Reusch (Ed.), Springer-Verlag, Berlin, Heidelberg, New York, 2006, pp. 499-511

[31] Zs. Cs. Johanyák: Student Evaluation Based on Fuzzy Rule Interpolation, International Journal of Artificial Intelligence, Vol. 5, No. A10, 2010, pp. 37-55

[32] L. Horváth, I. J. Rudas: Modeling and Problem Solving Methods for Engineers, Academic Press, Elsevier, Burlington, MA: 2004

[33] A. Palcu: Charged and Neutral Currents in a 3-3-1 Model with Right-Handed Neutrinos, Modern Physics Letters A, Vol. 23, No. 6, 2008, pp. 387-399

[34] B. Danković, S. Nikolić, M. Milojković, Z. Jovanović: A Class of Almost Orthogonal Filters, Journal of Circuits, Systems, and Computers, Vol. 18, No. 5, 2009, pp. 923-931

[35] J. Vaščák, L. Madarász: Adaptation of Fuzzy Cognitive Maps – A Comparison Study, Acta Polytechnica Hungarica, Vol. 7, No. 3, 2010, pp. 109-122

[36] Z.-Y. Zhao, W.-F. Xie, H. Hong: Hybrid Optimization Method of Evolutionary Parallel Gradient Search, International Journal of Artificial Intelligence, Vol. 5, No. A10, 2010, pp. 1-16

[37] J. K. Tar, I. J. Rudas, J. F. Bitó, J. A. Tenreiro Machado, K. R. Kozlowski: Adaptive Tackling of the Swinging Problem for a 2 DOF Crane – Payload System, in: Computational Intelligence in Engineering, I. J. Rudas, J. Fodor, J. Kacprzyk (Eds.), Springer-Verlag, Berlin, Heidelberg, 2010, pp. 103-114

[38] O. Linda, M. Manic: Fuzzy Force-Feedback Augmentation for Manual Control of Multi-Robot System, IEEE Transactions on Industrial Electronics, Vol. 58, No. 8, 2011, pp. 3213-3220

[39] K. Y. Chan, T. S. Dillon, C. K. Kwong: Modeling of a Liquid Epoxy Molding Process Using a Particle Swarm Optimization-Based Fuzzy Regression Approach, IEEE Transactions on Industrial Informatics, Vol. 7, No. 1, 2011, pp. 148-158

[40] A. Sadighi, W.-J. Kim: Adaptive-Neuro-Fuzzy-based Sensorless Control of a Smart-Material Actuator, IEEE/ASME Transactions on Mechatronics, Vol. 16, No. 2, 2011, pp. 371-379

[41] A. E. Ruano, C. L. Cabrita, P. M. Ferreira, L. T. Kóczy: Exploiting the Functional Training Approach in B-Splines, Preprints of 1st IFAC Conference on Embedded Systems, Computational Intelligence and Telematics in Control (CESCIT 2012), Würzburg, Germany, 2012, pp. 127-132

Appendix 1. Proof of stability condition

Theorem 1 is applied as follows in order to formulate the rule base of the MIMO FC for the spherical three tank system using the FCS structure given in Fig. 3. Since the design of the rule consequents is based on controller regions in the input space of the MIMO FC and on inequalities. Let the universe of discourse be $X=[-2,2]\times[-2,2]$, and $\mathbf{e}=\mathbf{0}\in X$. The Lyapunov function candidate defined in (37) is considered, and it is a continuously differentiable positive function on X.

The control laws in the rule consequents of MIMO FC are designed in order to fulfil (36). Therefore the following analysis is done for all rules.

Rule R^1. e_1 IS P and e_2 IS P. So $X_1^A=[0,2]\times[0,2]$. Accepting these conditions, equation (36) is transformed into

$$\dot{V}_1(\mathbf{e})\le 0,\ \forall \mathbf{e}\in X_1^A. \tag{38}$$

We choose u_1^1 and u_2^1 and we introduce the control laws in Table 1. So

$$\dot{V}_1(\mathbf{e}) = \frac{e_1 k_{m1}}{A(u_{h1}/k_{m1})}[S\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}} - \frac{u_{h3}}{k_{m3}})\sqrt{2g\,|\frac{r_1-e_1}{k_{m1}} - \frac{u_{h3}}{k_{m3}}|} - 2S\sqrt{gR}]$$
$$+ \frac{e_2 k_{m2}}{A(u_{h2}/k_{m2})}[-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}} - \frac{r_2-e_2}{k_{m2}})\sqrt{2g\,|\frac{u_{h3}}{k_{m3}} - \frac{r_2-e_2}{k_{m2}}|} - 2S\sqrt{gR}]. \quad (39)$$

The following inequalities hold:

$$S\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}} - \frac{u_{h3}}{k_{m3}})\sqrt{2g\,|\frac{r_1-e_1}{k_{m1}} - \frac{u_{h3}}{k_{m3}}|} < 2S\sqrt{gR}\,, \quad (40)$$

$$-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}} - \frac{r_2-e_2}{k_{m2}})\sqrt{2g\,|\frac{u_{h3}}{k_{m3}} - \frac{r_2-e_2}{k_{m2}}|} < 2S\sqrt{gR}\,, \quad (41)$$

because the terms in the modulus in the left-hand sides of (40) and (41) are in fact levels and they fulfil the constraints (29). Equations (39)–(41) lead to (38).

Rule R^2. e_1 IS N and e_2 IS N. So $X_2^A = [-2,0]\times[-2,0]$. Accepting these conditions, equation (36) is transformed into

$$\dot{V}_2(\mathbf{e}) \le 0,\ \forall \mathbf{e} \in X_2^A. \quad (42)$$

We choose u_1^2 and u_2^2, and we introduce the resulted control laws (that belong to the rule consequents) in Table 1. Therefore

$$\dot{V}_2(\mathbf{e}) = \frac{e_1 k_{m1}}{A(u_{h1}/k_{m1})}[S\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}} - \frac{u_{h3}}{k_{m3}})\sqrt{2g\,|\frac{r_1-e_1}{k_{m1}} - \frac{u_{h3}}{k_{m3}}|} + 2S\sqrt{gR}]$$
$$+ \frac{e_2 k_{m2}}{A(u_{h2}/k_{m2})}[-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}} - \frac{r_2-e_2}{k_{m2}})\sqrt{2g\,|\frac{u_{h3}}{k_{m3}} - \frac{r_2-e_2}{k_{m2}}|} + 2S\sqrt{gR}]. \quad (43)$$

Equations (40) and (41) hold for this rule, too. Equations (40), (41) and (43) lead to the fulfilment of the condition (42).

Rule R^3. e_1 IS N and e_2 IS P. So $X_3^A = [-2,0]\times[2,0]$. Therefore the condition (36) is transformed into

$$\dot{V}_3(\mathbf{e}) \le 0, \forall \mathbf{e} \in X_3^A. \quad (44)$$

We choose the expressions of the control laws u_1^3 and u_2^3, and these rule consequents are introduced in Table 1. Therefore

$$\dot{V}_3(\mathbf{e}) = \frac{e_1 k_{m1}}{A(u_{h1}/k_{m1})}[S\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}} - \frac{u_{h3}}{k_{m3}})\sqrt{2g\,|\frac{r_1-e_1}{k_{m1}} - \frac{u_{h3}}{k_{m3}}|} + 2S\sqrt{gR}]$$
$$+ \frac{e_2 k_{m2}}{A(u_{h2}/k_{m2})}[-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}} - \frac{r_2-e_2}{k_{m2}})\sqrt{2g\,|\frac{u_{h3}}{k_{m3}} - \frac{r_2-e_2}{k_{m2}}|} - 2S\sqrt{gR}]. \quad (45)$$

Equations (40), (41) and (45) lead to the fulfilment of (44).

Rule R^4. e_1 IS P and e_2 IS N. So $X_4^A = [0,2]\times[-2,0]$. Therefore the condition (36) is transformed into

$$\dot{V}_4(\mathbf{e}) \le 0, \forall \mathbf{e} \in X_4^A. \tag{46}$$

We choose u_1^4 and u_2^4, and these rule consequents are introduced in Table 1. So

$$\begin{aligned}\dot{V}_4(\mathbf{e}) = &\frac{e_1 k_{m1}}{A(u_{h1}/k_{m1})}[S\,\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}})\sqrt{2g\,|\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}}|}-2S\sqrt{gR}]\\ &+\frac{e_2 k_{m2}}{A(u_{h2}/k_{m2})}[-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}})\sqrt{2g\,|\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}}|}+2S\sqrt{gR}],\end{aligned} \tag{47}$$

and equations (40), (41) and (47) lead to the fulfilment of (46).

Rule R^5. e_1 IS P and e_2 IS Z. So $X_5^A = [0,2]\times[-0.5,0.5]$. Therefore the condition (28) is transformed into

$$\dot{V}_5(\mathbf{e}) \le 0, \forall \mathbf{e} \in X_5^A. \tag{48}$$

We choose the forms of u_1^5 and u_2^5, and these control laws are introduced as rule consequents in Table 1. Therefore

$$\begin{aligned}\dot{V}_5(\mathbf{e}) = &\frac{e_1 k_{m1}}{A(u_{h1}/k_{m1})}[S\,\mathrm{sgn}(\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}})\sqrt{2g\,|\frac{r_1-e_1}{k_{m1}}-\frac{u_{h3}}{k_{m3}}|}-2S\sqrt{gR}]\\ &-\frac{e_2^2 k_{m2}}{A(u_{h2}/k_{m2})}.\end{aligned} \tag{49}$$

Therefore equations (40) and (49) lead to the fulfilment of (48).

Rule R^6. e_1 IS Z and e_2 IS P. So $X_6^A = [-0.5,0.5]\times[0,2]$ and the condition (36) is transformed into

$$\dot{V}_6(\mathbf{e}) \le 0, \forall \mathbf{e} \in X_6^A. \tag{50}$$

We choose u_1^6 and u_2^6, and these rule consequents are introduced in Table 1. So

$$\begin{aligned}\dot{V}_6(\mathbf{e}) = &-\frac{e_1^2 k_{m1}}{A(u_{h1}/k_{m1})}\\ &+\frac{e_2 k_{m2}}{A(u_{h2}/k_{m2})}[-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}})\sqrt{2g\,|\frac{u_{h3}}{k_{m3}}-\frac{r_2-e_2}{k_{m2}}|}-2S\sqrt{gR}].\end{aligned} \tag{51}$$

Equations (41) and (51) lead to the fulfilment of (50).

Rule R^7. e_1 IS N and e_2 IS Z. So $X_7^A = [-2,0]\times[-0.5,0.5]$ and the condition (36) is transformed into

$$\dot{V}_7(\mathbf{e}) \le 0, \forall \mathbf{e} \in X_7^A. \tag{52}$$

We choose u_1^8 and u_2^8, and these control laws (that belong to the rule consequents) are introduced in Table 1. Therefore

$$\begin{aligned}\dot{V}_8(\mathbf{e}) &= -\frac{e_1^2 k_{m1}}{A(u_{h1}/k_{m1})} \\ &+ \frac{e_2 k_{m2}}{A(u_{h2}/k_{m2})}[-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}} - \frac{r_2 - e_2}{k_{m2}})\sqrt{2g\,|\frac{u_{h3}}{k_{m3}} - \frac{r_2 - e_2}{k_{m2}}|} + 2S\sqrt{gR}],\end{aligned} \tag{53}$$

and equations (41) and (53) lead to the fulfilment of (52).

Rule R^8. e_1 IS Z and e_2 IS N. So $X_8^A = [-0.5,0.5]\times[0,2]$ and the condition (36) is transformed into

$$\dot{V}_8(\mathbf{e}) \le 0, \forall \mathbf{e} \in X_8^A. \tag{54}$$

We choose the forms of u_1^8 and u_2^8, and these control rules are introduced as rule consequents in Table 1. Therefore

$$\begin{aligned}\dot{V}_8(\mathbf{e}) &= -\frac{e_1^2 k_{m1}}{A(u_{h1}/k_{m1})} \\ &+ \frac{e_2 k_{m2}}{A(u_{h2}/k_{m2})}[-S\,\mathrm{sgn}(\frac{u_{h3}}{k_{m3}} - \frac{r_2 - e_2}{k_{m2}})\sqrt{2g\,|\frac{u_{h3}}{k_{m3}} - \frac{r_2 - e_2}{k_{m2}}|} + 2S\sqrt{gR}],\end{aligned} \tag{55}$$

and equations (41) and (55) lead to the fulfilment of (54).

Rule R^9. e_1 IS Z and e_2 IS Z. So $X_9^A = [-0.5,0.5]\times[-0.5,0.5]$ and the condition (36) is transformed into

$$\dot{V}_9(\mathbf{e}) \le 0, \forall \mathbf{e} \in X_9^A. \tag{56}$$

We choose u_1^9 and u_2^9, and these rule consequents are introduced in Table 1. So

$$\dot{V}_9(\mathbf{e}) = -\frac{e_1^2 k_{m1}}{A(u_{h1}/k_{m1})} - \frac{e_2^2 k_{m2}}{A(u_{h2}/k_{m2})} < 0. \tag{57}$$

Therefore equation (57) guarantees the fulfilment of (56).

Concluding, the formulation of the rule base of the MIMO FC for the spherical three tank system (Table 1) was done such that to fulfil the condition (36). This condition is fulfilled because we proved equivalent conditions for all rules.

Ranking Decision Making Units Based on DEA-like Nonreciprocal Pairwise Comparisons

János Fülöp[1], Rita Markovits-Somogyi[2]

[1] Research Group of Operations Research and Decision Systems
Computer and Automation Research Institute, Hungarian Academy of Sciences
Kende u. 13-17, H-1111 Budapest, Hungary
E-mail: fulop@sztaki.hu

[2] Department of Transport Economics
Budapest University of Technology and Economics
Bertalan L. u. 2, H-1111 Budapest, Hungary
E-mail: rsomogyi@kgazd.bme.hu

Abstract: Ever since the birth of data envelopment analysis (DEA) the question of ranking the decision making units (DMUs) has been one of the focal points of research in the area. Among several other approaches, promising attempts have been made to marry DEA with the analytic hierarchy process (AHP) method. Keeping the idea of using DEA-based pairwise comparisons between the DMUs, as proposed in some DEA-AHP variants published in the literature, a new method is presented for combining DEA with techniques for eliciting weights from pairwise comparison matrices. The basic idea is to apply a variant of the CCR problem instead of the classic one. The ensuing scores are then utilized to build a nonreciprocal pairwise comparison matrix which serves as the basis for eliciting the ranking values of the DMUs. The main advantage of this new method is the wider range of the resulting ranking values which subsequently leads to better distinction between the DMUs. Besides the eigenvector method, optimization based methods are also considered for eliciting the ranking values from the nonreciprocal pairwise comparison matrix. Numerical examples are supplied for comparing the proposed techniques.

Keywords: data envelopment analysis; ranking decision making units; pairwise comparisons; techniques for eliciting weights

1 Introduction

Data envelopment analysis (DEA) is a very powerful tool for the efficiency evaluation of decision making units (DMUs) with multiple inputs and outputs. One of the main advantages of this non-parametric linear programming method is its capability of discerning the efficient DMUs by creating an efficiency frontier as based on the observed data and thus not requiring any *a priori* information about the relationship between the inputs and outputs.

One of its shortcomings, however, is its inability to fully rank the decision making units. Ever since it was created by Charnes, Cooper and Rhodes [9] on the basis of the idea of Farrel [20], the question of full ranking has been in the frontline of research.

A multitude of different DEA variants with diverse backgrounds have been developed with the aim of solving this problem. Perhaps the most widely known and applied ranking method is the super-efficiency DEA model. Developed by Andersen and Petersen [3], this technique creates the best practice frontier first without evaluating DMU_0 and then with its inclusion. Next the extent to which the envelopment frontier becomes extended is investigated. Several extensions and variants of this method are available, e.g. [25, 29, 30]. The problem with super-efficiency DEA is that under certain conditions infeasibility occurs, which limits the applicability of the technique. (For details see [37]).

Another approach is that of cross-efficiency introduced by Sexton et al. [39] and extended by Doyle and Green [15] where the individual decision making units are not only assessed by their own weights, but the weights of all the other DMUs are also incorporated into the value judgement. This score is more representative of efficiency than the traditional DEA-score, but at the same time the connection to the multiplier weights is lost [2].

Torgensen et al. [42] investigate the extent to which the different DMUs are peers to each other, and through this benchmarking procedure is full ranking achieved.

If developed purposefully, the utilization of common weights can contribute to reaching a full ranking as well. The main goal of Wang et al. [43] is to introduce a minimum weight restriction and, as a side effect, common weights and then full ranking is also achieved. Common weights are also achieved by different multivariate statistical analyses and these can also lead to full ranking. For instance, canonical correlation, linear discriminant analysis or the discriminant analysis of ratios can also be employed [2].

Another way of approaching the question of full ranking is to take into account the slacks present in the slack-adjusted DEA model. Bardhan et al. [4] rank inefficient units this way. Tone [41] composes a method that can rank all the DMUs. Du et al. [16] create an extension to this model and Chen and Sherman [10] also use slacks for the development of a non-radial super-efficiency DEA model.

Wen and Li [44] aim to utilize fuzzy information in data envelopment analysis and as a side effect full ranking is achieved.

Multi-criteria decision making methods can also be combined with DEA to provide full ranking. Sinuany-Stern et al. [40] integrate analytic hierarchy process with DEA: the pairwise comparison matrix of AHP is created through the objective evaluation of pairs of DMUs by DEA.

The method developed by Sinuany-Stern et al. [40] has been applied for instance by Guo et al. [23] for supply chain evaluation. Royendegh and Erol [34, 35] also build upon the idea of [40] but extend the method to the analytic network process (ANP), the more generalized form of AHP. Zhang et al. [46] combine DEA with AHP for 4PL vendor selection, but their approach is different. After the construction of an input-output structure, AHP is utilized for a preliminary data analysis with the help of which the importance of the different criteria is determined. The results of the AHP are then used as preferential information in a modified DEA model. A pairwise comparison matrix is created with the evolving efficiency values and then AHP is applied again for the evaluation of the matrix.

The authors have found the work of Sinuany-Stern et al. [40] particularly inspiring and, appreciating the results therein, wish to further improve the original method by enabling the pairwise comparison matrix to be nonreciprocal, which contributes to the possibility of a more accurate evaluation. Even more so, since in the course of applying DEA to the case of logistic centres [31, 32], it has been revealed that the thumb rule in connection with the number of DMUs to be found in the literature is very difficult to be adhered to. According to this rule, the number of observations should be three times greater than the number of the inputs plus outputs; and the number of DMUs should be equal or larger than the product of the number of inputs and outputs [5]. Some authors are less strict in their conduct; Wu and Goh [45] for example argue that the number of DMUs should only be minimally two times as much as the sum of the number of inputs and outputs. However, under certain conditions, even this requirement might be difficult to satisfy.

The new technique proposed in this article intends to provide a solution for both the problem of full ranking and the difficulty inherent in the thumb rule cited above. Nonetheless, it is not in the scope of the present paper to explore and compare all the DEA-connected techniques aimed at resolving the question of full ranking, or even to measure up the proposed method to all the rest of existing solutions. It is its goal, however, to provide numerical examples of its utilization, with special attention paid to the cases presented in articles closely related to the technique at hand.

In Section 2, a variant of the CCR problem is proposed instead of the classic one used in [40]. Easy ways for computing the related reciprocal and nonreciprocal pairwise comparison matrices are also presented. Section 3 addresses the issue of how to elicit ranking weights from the nonreciprocal pairwise comparison matrix obtained by the new approach. Besides the eigenvector method used in [40], the logarithmic least squares method and the weighted least squares method are also considered. In Section 4, numerical examples are supplied for comparing the proposed techniques.

2 DEA-like Pairwise Comparisons of the Decision Making Units

Let us assume that there are n decision making units each producing s different outputs from m different inputs. X_{ij} is the input i of unit j, while Y_{rj} is the output r of unit j. We assume that all X_{ij}s and Y_{rj}s are positive. In the original approach of Sinuany-Stern et al. [40], traditional DEA runs are executed for any pair of DMUs, as if only these two decision making units existed. The runs are based on the DEA CCR model adapted to the case of two DMUs. For this, let A and B be a pair of units and let us consider the CCR model as if only these two units existed:

$$\begin{aligned} F_{\mathrm{AB}} = \max \quad & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{A}} \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{m} v_i X_{i\mathrm{A}} = 1, \\ & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{A}} \le 1, \\ & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{B}} - \sum_{i=1}^{m} v_i X_{i\mathrm{B}} \le 0, \\ & u_r \ge 0, r = 1, \dots, s,\ v_i \ge 0, i = 1, \dots, m. \end{aligned} \tag{1}$$

In [40] the notation E_{AA} is used for F_{AB}. We think, however, that F_{AB} is a more appropriate notation since both units appear in it, and the precedence of A over B means that it is the efficiency of A which is being evaluated by using the two units. The change from E to F is necessary in order to avoid its confusion with another efficiency value later on. Consider another problem being in a close relation to (1):

$$\begin{aligned} \hat{F}_{\mathrm{AB}} = \max \quad & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{A}} \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{m} v_i X_{i\mathrm{A}} = 1, \\ & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{B}} - \sum_{i=1}^{m} v_i X_{i\mathrm{B}} = 0, \\ & u_r \ge 0, r = 1, \dots, s,\ v_i \ge 0, i = 1, \dots, m. \end{aligned} \tag{2}$$

Comparing (2) with (1), the main difference between the two models becomes clearly visible: the second constraint of (1) representing an upper bound for the objective function of (1) is omitted. The reason behind this is the basic idea of the new model (2): the aim with the exclusion is to provide an opportunity for a full comparison between the two decision making units, without limiting the evolving score. If that constraint is left untouched, the resulting efficiency value will very frequently be the unity; and thus real distinction is not achieved between the two DMUs. A similar idea of omitting the upper bound on the objective function of (1) has already appeared in the case of the super-efficiency ranking techniques [2, 3] too.

A further minor remark concerns the inequality in the third constraint of (1), which changes to equality in (2). This can be explained by the following: should we leave the inequality in (2), it would clearly hold as an equality for any optimal

solution of (2). Since we shall only be interested in the optimal solutions of (2), we can consider that constraint as an equality.

Proposition 1:

$$\hat{F}_{\mathrm{AB}} = \max_{(r,i)} \frac{Y_{r\mathrm{A}}/X_{i\mathrm{A}}}{Y_{r\mathrm{B}}/X_{i\mathrm{B}}} = \max_{r=1,\dots,s} \frac{Y_{r\mathrm{A}}}{Y_{r\mathrm{B}}} \cdot \max_{i=1,\dots,m} \frac{X_{i\mathrm{B}}}{X_{i\mathrm{A}}} \tag{3}$$

and

$$F_{\mathrm{AB}} = \min\{1, \hat{F}_{\mathrm{AB}}\}. \tag{4}$$

Proof: Problem (2) has a finite optimal solution taken at a basic feasible solution. It is clear from the special structure of (2) that any basic feasible solution has exactly two positive variables: one from the v_i and one from the u_r variables. Given a basic feasible solution of (2), let i_0 and r_0 denote the indices of those positive variables. It is easy to see that

$$v_{i_0} = \frac{1}{X_{i_0\mathrm{A}}} \quad \text{and} \quad u_{r_0} = \frac{v_{i_0} X_{i_0\mathrm{B}}}{Y_{r_0\mathrm{B}}} = \frac{X_{i_0\mathrm{B}}/X_{i_0\mathrm{A}}}{Y_{r_0\mathrm{B}}},$$

and the value of the objective function is

$$\frac{Y_{r_0\mathrm{A}}/X_{i_0\mathrm{A}}}{Y_{r_0\mathrm{B}}/X_{i_0\mathrm{B}}}. \tag{5}$$

Finding the optimal basic solution means finding the pair (r_0, i_0) with the maximal value of (5). This implies (3) directly.

Keeping in mind the upper bounding role of the second constraint of (1), it is evident that F_{AB} is equal to $\hat{F}_{\mathrm{AB}}$ if $\hat{F}_{\mathrm{AB}} \leq 1$, and $F_{\mathrm{AB}} = 1$ otherwise. □

In essence, $\hat{F}_{\mathrm{AB}}$ is the resulting value of the pairwise comparison; it represents the efficiency of A in comparison with B. According to (3), it can also be interpreted as the product of an output and an input efficiency value. The first is the maximal output ratio of A compared with B, and the second one is the maximal input ratio of B compared with A. This can also be interpreted in a way that, in this comparison, that single input and that single output will be chosen which is most satisfying from unit A's point of view. It is clear from (3) that this pairwise comparison is not reciprocal, i.e. $\hat{F}_{\mathrm{AB}} = 1/\hat{F}_{\mathrm{BA}}$ does not necessarily hold.

The original approach proposed in [40] determines F_{AB} in the first step; then, in the second step a cross evaluation of unit B is performed based on the idea of [33]:

$$\begin{aligned} E_{\mathrm{BA}} = \max \quad & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{B}} \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{m} v_i X_{i\mathrm{B}} = 1, \\ & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{B}} \leq 1, \\ & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{A}} - F_{\mathrm{AB}} \sum_{i=1}^{m} v_i X_{i\mathrm{A}} = 0, \\ & u_r \geq 0, r = 1, \dots, s, \; v_i \geq 0, i = 1, \dots, m. \end{aligned} \tag{6}$$

Actually, E_{AB} is the optimal cross evaluation of unit B, while the output/input ratio of unit A is fixed at F_{AB}. Using a similar reasoning as with (1) and (2), and omitting the second constraints of (6), i.e. an upper bound restriction for the objective function, the following auxiliary problem can be established for (6):

$$\begin{aligned} \hat{E}_{\mathrm{BA}} = \max \quad & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{B}} \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{m} v_i X_{i\mathrm{B}} = 1, \\ & \textstyle\sum_{r=1}^{s} u_r Y_{r\mathrm{A}} - F_{\mathrm{AB}} \sum_{i=1}^{m} v_i X_{i\mathrm{A}} = 0, \\ & u_r \geq 0, r = 1, \dots, s, \; v_i \geq 0, i = 1, \dots, m. \end{aligned} \tag{7}$$

The following Proposition 2 can be proved in the same way as Proposition 1; only the constant F_{AB} appearing in the second constraint of (7) requires some extra attention. We leave the proof to the reader.

Proposition 2:

$$\hat{E}_{\mathrm{BA}} = F_{\mathrm{AB}} \cdot \max_{(r,i)} \frac{Y_{r\mathrm{B}}/X_{i\mathrm{B}}}{Y_{r\mathrm{A}}/X_{i\mathrm{A}}} = F_{\mathrm{AB}} \cdot \max_{r=1,\dots,s} \frac{Y_{r\mathrm{B}}}{Y_{r\mathrm{A}}} \cdot \max_{i=1,\dots,m} \frac{X_{i\mathrm{A}}}{X_{i\mathrm{B}}} \tag{8}$$

and

$$E_{\mathrm{BA}} = \min \{1, \hat{E}_{\mathrm{BA}}\}. \tag{9}$$

□

Collating (8) and (3), changing the role of A and B in the latter, we obtain

Corollary 1:

$$\hat{E}_{\mathrm{BA}} = F_{\mathrm{AB}} \hat{F}_{\mathrm{BA}}. \tag{10}$$

□

Proposition 3: If $F_{\mathrm{AB}} < 1$, then $E_{\mathrm{BA}} = 1$.

Proof: From (4) and $F_{\mathrm{AB}} < 1$, we get $F_{\mathrm{AB}} = \hat{F}_{\mathrm{AB}}$. Then, from (3) and (8),

$$\hat{E}_{\mathrm{BA}} = \max_{(r,i)} \frac{Y_{r\mathrm{A}}/X_{i\mathrm{A}}}{Y_{r\mathrm{B}}/X_{i\mathrm{B}}} \cdot \max_{(r,i)} \frac{Y_{r\mathrm{B}}/X_{i\mathrm{B}}}{Y_{r\mathrm{A}}/X_{i\mathrm{A}}} = \frac{\max\limits_{(r,i)} \frac{Y_{r\mathrm{A}}/X_{i\mathrm{A}}}{Y_{r\mathrm{B}}/X_{i\mathrm{B}}}}{\min\limits_{(r,i)} \frac{Y_{r\mathrm{A}}/X_{i\mathrm{A}}}{Y_{r\mathrm{B}}/X_{i\mathrm{B}}}} \geq 1.$$

From (9), we get $E_{\mathrm{BA}} = 1$ immediately. □

As a consequence of Proposition 3, if $F_{\mathrm{AB}} < 1$, then it is unnecessary to solve problems (6) or (7), we get $E_{\mathrm{BA}} = 1$ directly.

Proposition 4: If there exist (r_1, i_1) and (r_2, i_2) such that

$Y_{r_1\mathrm{A}}/X_{i_1\mathrm{A}} \geq Y_{r_1\mathrm{B}}/X_{i_1\mathrm{B}}$ and $Y_{r_2\mathrm{A}}/X_{i_2\mathrm{A}} \leq Y_{r_2\mathrm{B}}/X_{i_2\mathrm{B}}$,

then $F_{\mathrm{AB}} = E_{\mathrm{BA}} = 1$.

Proof: From (3) and (4), we obtain $F_{\mathrm{AB}} = 1$. Then, (8) implies $\hat{E}_{\mathrm{BA}} \geq 1$, thus $E_{\mathrm{BA}} = 1$. □

Remark 1: It is easy to see that the condition of Proposition 4 does not hold if and only if

$$Y_{r\mathrm{A}}/X_{i\mathrm{A}} < Y_{r\mathrm{B}}/X_{i\mathrm{B}} \text{ for all } (r,i) \tag{11}$$

or

$$Y_{r\mathrm{A}}/X_{i\mathrm{A}} > Y_{r\mathrm{B}}/X_{i\mathrm{B}} \text{ for all } (r,i). \tag{12}$$

Remark 2: We have also pointed out in Propositions 1 and 2 that we do not have to use any optimization software to obtain the optimal solution of (1), (2), (6) and (7).

Since n DMUs are given, and the values F_{AB}, $\hat{F}_{\mathrm{AB}}$, E_{BA} and $\hat{E}_{\mathrm{BA}}$ are to be determined for all pairs A and B of the DMUs, we introduce the following notations. Let $\mathrm{DMU}_1,\ldots,\mathrm{DMU}_n$ denote the decison making units. Considering DMU_j as A and DMU_k as B, let $F_{jk} = F_{\mathrm{AB}}$, $\hat{F}_{jk} = \hat{F}_{\mathrm{AB}}$, $E_{kj} = E_{\mathrm{BA}}$ and $\hat{E}_{kj} = \hat{E}_{\mathrm{BA}}$.

Sinuany-Stern et al. [40] construct an $n \times n$ matrix $A = [a_{jk}]$ of the entries

$$a_{jk} = \frac{F_{jk}+E_{jk}}{F_{kj}+E_{kj}}, \ j,k = 1,\ldots,n. \tag{13}$$

The nominator of a_{jk} is the sum of the efficiency evaluation and the cross evaluation of DMU_j in comparison with DMU_k. The denominator can be interpreted similarly by changing the role of j and k.

Clearly,

$$a_{jk} > 0, \ j,k = 1,\ldots,n, \tag{14}$$

and

$$a_{jk} = 1/a_{kj}, \ j,k = 1,\ldots,n. \tag{15}$$

An $n \times n$ matrix A with the properties (14) and (15) is called a *pairwise comparison matrix* [36].

Constructing the pairwise comparison matrix A by (13) is, however, in some cases problematic. Namely, if considering A as DMU_j and B as DMU_k and neither (11) nor (12) hold, then $F_{jk} = E_{kj} = F_{kj} = E_{jk} = 1$, consequently, $a_{jk} = a_{kj} = 1$. In some practical or randomly generated cases, it is not very probable that the dominance of (11) or (12) holds. This may lead to the phenomenon observed in some numerical examples applying the approach of [41] that the pairwise comparison matrices constructed by (13) comprise strikingly many 1 elements [23, 34, 40, 46]. The result 1 of a pairwise comparison means that the two DMUs cannot be considered as different. Therefore, a large number of unities in the pairwise comparison matrix may hinder the strict ranking of DMUs as the ranking

weights elicited from the pairwise comparison matrix may be identical or very close to each other.

In this paper, instead of using (13), we propose to construct an $n \times n$ matrix $A = [a_{jk}]$ of the entries

$$a_{jk} = \hat{F}_{jk}\ ,\ j,k = 1,\dots,n. \tag{16}$$

As mentioned earlier, $\hat{F}_{jk}$ provides the opportunity to compare DMU_j against DMU_k without limiting the score from above. It is undeniable that DMU_j is in a privileged position in the course of this comparison. It can also be considered as a football match where DMU_j has the home field advantage. Of course, the role of DMU_j and DMU_k is reversed when the value $a_{kj} = \hat{F}_{kj}$ is determined, and thus a level playing field can be assured.

It was already pointed out that $\hat{F}_{\mathrm{AB}}$ and $1/\hat{F}_{\mathrm{BA}}$ may be different, i.e. the reciprocity property (15) does not necessarily hold for a matrix A constructed by (16). Actually, matrix A of (16) is simply a positive matrix, but we call it a *nonreciprocal* pairwise comparison matrix to indicate the context.

3 Eliciting Weights from the Nonreciprocal Pairwise Comparison Matrix

After having constructed the pairwise comparison matrix A by (13), Sinuany-Stern et al. [40] follow the standard AHP methodology [36]. By applying the Eigenvector Method (EM), the maximal eigenvalue $\lambda_{\max}$ and its corresponding eigenvector w^{EM} of

$$Aw = \lambda w \tag{17}$$

are determined. The real number $\lambda_{\max}$ and the vector w^{EM} are positive and unique. In addition, $\lambda_{\max} \geq n$, and $\lambda_{\max} = n$ if and only if A is consistent, i.e.

$a_{ij}a_{jk} = a_{ik}$ for all $i,j,k = 1,\dots,n.$

Having determined the vector w^{EM}, the DMUs are ranked in the following way. Rank 1 is assigned to the DMU with the maximal value of w_j^{EM}, and in decreasing order of w_j^{EM} are the further ranks allocated to the remaining DMUs.

In the standard AHP methodology, matrix A is assumed to be reciprocal. The matrix A of (16) is, however, nonreciprocal. As far as the authors know, nonreciprocal matrices in a pairwise comparison context and the question of how to elicit the weight vector w from them appeared first in [28]. Although double wine testing is mentioned as a main example for the necessity of the relaxation of the reciprocity condition, further examples of application from other fields of life have also been published in the literature.

The lack of reciprocity may even occur if the comparisons are performed by the same person at different times. In [14] an experiment is reported. Postgraduate students were asked to fill in the upper part of a pairwise comparison matrix, where the items to be compared were in the scope of their studies. Some weeks later, they were asked to fill in the lower part of the PC matrix. It turned out that for none of the matrices obtained in this way did the reciprocity property hold.

As mentioned in [24], nonreciprocal pairwise comparison matrices may also appear when one compares financial assets denominated in different currencies. Because of transaction costs, the resulting matrices are not reciprocal, even if no subjectivity is involved.

The classical methods used in case of reciprocal pairwise comparison matrices can also be extended to the nonreciprocal case in more or less direct ways. EM can also be interpreted without the reciprocity condition since the Perron-Frobenius and the Frobenius theorems, the mathematical bases of the EM, do not require A to be reciprocal; see [38] for details. The property $\lambda_{\max} \geq n$ does not necessarily hold for a nonreciprocal matrix A; thus, the consistency index $\mathrm{CI} = (\lambda_{\max} - n)/(n-1)$ playing an important role in AHP [36] may be negative. However, since the pairwise comparison matrices are constructed from objective data in our case, we are not concerned about the consistency of A.

The Eigenvector Method is not the only way to elicit ranking weights from a pairwise comparison matrix. A group of approaches applies optimization methods and proposes different ways for minimizing the difference between A and consistent pairwise comparison matrices. The optimization methods are based on the basic property that A is consistent if and only if

$$a_{ij} = \frac{w_i}{w_j}, \quad i,j = 1,\dots,n,$$

where w is a positive n-vector and is unique after a normalization. Most of the optimization approaches can be directly extended to the nonreciprocal case as well. If the difference to be minimized is measured in the least-squares sense, i.e., with the Frobenius norm, then we get the Least Squares Method (LSM) [11]:

$$\min \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{ij} - \frac{w_i}{w_j}\right)^2 \tag{18}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} w_i = 1, \;\; w_i > 0, i = 1,\dots,n.$$

Under special conditions, (18) can be transcribed into the form of a convex optimization problem and can be solved by simple local search techniques [21, 22]. However, without the special conditions, problem (18) may be a difficult nonconvex optimization problem with multiple local optima and even with multiple isolated global optimal solutions [26, 27].

In order to elude the difficulties caused by the possible nonconvexity of (18), several other, more easily solvable problem forms are proposed to derive priority

weights from a pairwise comparison matrix. The Weighted Least Squares Method (WLSM) [6, 11] in the form of

$$\min \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{ij}w_j - w_i\right)^2 \tag{19}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} w_i = 1, \;\; w_i > 0, i = 1, \dots, n$$

involves a convex quadratic optimization problem whose unique optimal solution is easy to obtain.

The Logarithmic Least Squares Method (LLSM) [12, 13] in the form

$$\min \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\log a_{ij} - \log\frac{w_i}{w_j}\right)^2 \tag{20}$$

$$\text{s.t.} \quad \prod_{i=1}^{n} w_i = 1, \;\; w_i > 0, i = 1, \dots, n$$

is based on an optimization problem whose unique optimal solution, in the reciprocal case, is the geometric mean of the rows of matrix A. This result was extended to the nonreciprocal case in [28], and the following optimal solution to (20) was obtained:

$$w_i = \sqrt[2n]{\frac{\prod_{j=1}^{n} a_{ij}}{\prod_{j=1}^{n} a_{ji}}}, \;\; i = 1, \dots, n. \tag{21}$$

In this paper, for the numerical experiment, we used the EM, LLSM and WLSM approaches to elicit ranking weights from pairwise comparison matrices. For further approaches, see [7, 8, 17, 18, 19, 21] and the references therein.

Of course, the different approaches may result in different weight vectors w, and consequently, in different ranking orders of some DMUs. This well-known phenomenon occurred in the following numerical examples, too.

4 Numerical Examples

We compared the original AHP/DEA ranking methodology proposed in [40] with the new method presented in this paper. The original and three variants of the new method were tested in parallel. The latter differ in the way the weight vector w, used for ranking the DMUs, is elicited from the appropriate pairwise comparison matrix. EM1 – the original method – applies the Eigenvector Method on the reciprocal pairwise comparison matrix of elements yielded by (13), while EM2, LLSM and WLSM apply the Eigenvector Method, the Logarithmic Least Squares Method and the Weighted Least Squares Method, respectively, on the nonreciprocal pairwise comparison matrix $\hat{F}$.

Example 1

Example 1 is the nursing home example developed in [39]. This example was also used in the review paper [2] to compare the majority of the techniques surveyed there. The example comprises six DMUs, two inputs and two outputs: staff hours per day (StHr) and supplies per day (Supp) as inputs, and total Medicare plus Medicaid reimbursed patient days (MCPD) and total private patient days (PPD) as outputs. The raw data are presented in Table 1.

Table 1
Input and output data of Example 1

DMU	Inputs		Outputs	
	StHr	Supp	MCPD	PPD
A	150	0.2	14000	3500
B	400	0.7	14000	21000
C	320	1.2	42000	10500
D	520	2.0	28000	42000
E	350	1.2	19000	25000
F	320	0.7	14000	15000

By applying (2), we obtain the matrix $\hat{F}$:

$$\hat{F} = \begin{pmatrix} 1.00000 & 3.50000 & 2.00000 & 5.00000 & 4.42105 & 3.50000 \\ 2.25000 & 1.00000 & 3.42857 & 1.42857 & 1.44000 & 1.40000 \\ 1.40625 & 3.75000 & 1.00000 & 2.50000 & 2.41776 & 3.00000 \\ 3.46154 & 1.53846 & 2.46154 & 1.00000 & 1.13077 & 1.72308 \\ 3.06122 & 1.55102 & 2.38095 & 1.13095 & 1.00000 & 1.52381 \\ 2.00893 & 1.25000 & 2.44898 & 1.42857 & 1.26316 & 1.00000 \end{pmatrix}.$$

Since every element of $\hat{F}$ is greater than or equal to 1, from (4), we get $F_{jk} = 1$, $j,k=1,\ldots,n$. Then, from (10), we obtain $\hat{E}_{jk} \geq 1$, and from (9), $E_{kj} = 1$ for all $j,k=1,\ldots,n$. Consequently, the reciprocal pairwise comparison matrix A constructed by (13) consists only of unity elements. This matrix is consistent, the maximal eigenvector consists of equal weights, and all DMUs get the ranking 1-6 as shown in Table 2 in the columns under EM1.

Methods EM2, LLSM and WLSM were run on the nonreciprocal pairwise comparison matrix $\hat{F}$. The results are also displayed in Table 2. Although a deeper numerical comparison of these methods against other ranking methods is beyond the scope of this paper, it is worth noting that the ranking by EM2 is the same as that of the super-efficiency approach in [2]; moreover, the correlation of the vector of weights by EM2 to that by the super-efficiency technique is 0.99329. In the case of LLSM and WLSM, the ranking orders are the same only in the first two positions but the correlations are still 0.97891 and 0.97337, respectively.

Table 2
Weights and ranking of the DMUs in Example 1

DMU	Weights				Ranking			
	EM1	EM2	LLSM	WLSM	EM1	EM2	LLSM	WLSM
A	0.16667	0.23760	0.19755	0.19196	1-6	1	1	1
B	0.16667	0.14988	0.15839	0.15899	1-6	4	5	5
C	0.16667	0.17439	0.16565	0.16947	1-6	2	2	2
D	0.16667	0.15985	0.16437	0.14730	1-6	3	4	6
E	0.16667	0.14951	0.16438	0.16461	1-6	5	3	4
F	0.16667	0.12877	0.14966	0.16767	1-6	6	6	3

Example 2

The raw data of Example 2 is from Table 3 of [40], and is shown in Table 3 below.

Table 3
Input and output data of Example 2

DMU	Inputs		Outputs	
	X_1	$X2$	Y_1	Y_2
A	50	55	10	56
B	130	60	12	78
C	68	96	45	9
D	45	30	35	18
E	5	3	99	3

By applying (2), we obtain the matrix $\hat{F}$:

$$\hat{F} = \begin{pmatrix} 1.00000 & 2.16667 & 10.86061 & 2.80000 & 1.86667 \\ 1.27679 & 1.00000 & 13.86667 & 2.16667 & 1.30000 \\ 3.30882 & 7.16912 & 1.00000 & 0.85084 & 0.22059 \\ 6.41667 & 8.42593 & 6.40000 & 1.00000 & 0.66667 \\ 181.50000 & 214.50000 & 70.40000 & 28.28571 & 1.00000 \end{pmatrix}. \tag{22}$$

The reciprocal pairwise comparison matrix of the elements (13) is

$$A = \begin{pmatrix} 1.00000 & 1.00000 & 1.00000 & 1.00000 & 1.00000 \\ 1.00000 & 1.00000 & 1.00000 & 1.00000 & 1.00000 \\ 1.00000 & 1.00000 & 1.00000 & 0.85084 & 0.22059 \\ 1.00000 & 1.00000 & 1.17531 & 1.00000 & 0.66667 \\ 1.00000 & 1.00000 & 4.53333 & 1.50000 & 1.00000 \end{pmatrix}. \tag{23}$$

Two versions of Example 2 were solved in [40]. In the first one only four DMUs, Unit A to Unit D, were considered. The matrices $\hat{F}$ and A corresponding to this case can be easily obtained by taking the 4×4 upper-left submatrix of (22) and (23), respectively. The weights and the ranking of the DMUs in the case with four DMUs are shown in Table 4. The second version of Example 2 is with five DMUs. The corresponding weights and ranking can be found in Table 5.

Table 4
Weights and ranking of the DMUs in Example 2 with four DMUs

DMU	Weights				Ranking			
	EM1	EM2	LLSM	WLSM	EM1	EM2	LLSM	WLSM
A	0.24980	0.23732	0.26086	0.22924	2-3	3	2	2
B	0.24980	0.24654	0.20025	0.11203	2-3	2	3	3
C	0.24010	0.19087	0.14399	0.04622	4	4	4	4
D	0.26031	0.32527	0.39490	0.61251	1	1	1	1

Table 5
Weights and ranking of the DMUs in Example 2 with five DMUs

DMU	Weights				Ranking			
	EM1	EM2	LLSM	WLSM	EM1	EM2	LLSM	WLSM
A	0.19057	0.06436	0.07302	0.00538	2-3	2	3	4
B	0.19057	0.05037	0.05606	0.00451	2-3	3	4	5
C	0.14070	0.02515	0.04031	0.01349	5	5	5	3
D	0.17608	0.04872	0.11055	0.03955	4	4	2	2
E	0.30208	0.81140	0.72007	0.93708	1	1	1	1

In both versions, methods EM2, LLSM and WLSM yield full ranking orders although they are different in several positions. It is also a striking property of the methods based on the nonreciprocal pairwise comparison matrix $\hat{F}$ that the weights are more distinguished. The largest weights are significantly larger and the smallest weights are significantly smaller than those of EM1. Also, it is worth observing how the favorable input and output values of DMU E are reflected in the corresponding weights in Table 5.

Example 3

Example 3 comes from [23] and the raw data are given in Table 6.

Table 6
Input and output data of Example 3

DMU	Inputs			Outputs		
	X_1	X_2	X_3	Y_1	Y_2	Y_3
DMU_1	15	15	0.05	0.80	0.800	0.42
DMU_2	70	25	0.10	0.90	0.900	0.53
DMU_3	45	16	0.07	0.96	0.885	0.47
DMU_4	40	30	0.12	0.85	0.750	0.32
DMU_5	35	25	0.11	0.75	0.845	0.44
DMU_6	60	18	0.15	0.85	0.755	0.25
DMU_7	55	20	0.08	0.70	0.850	0.51
DMU_8	30	12	0.09	0.95	0.700	0.46

The nonreciprocal pairwise comparison matrix $\hat{F}$ is obtained by (2), and the reciprocal A by (13):

$$\hat{F} = \begin{pmatrix} 1.00000 & 4.14815 & 2.71186 & 3.50000 & 2.48889 & 6.72000 & 4.19048 & 2.28571 \\ 0.75714 & 1.00000 & 0.78936 & 1.98750 & 1.32500 & 3.18000 & 1.02857 & 1.15714 \\ 1.12500 & 1.66667 & 1.00000 & 2.75391 & 2.01143 & 4.02857 & 1.71429 & 1.62551 \\ 0.53125 & 1.65278 & 0.99609 & 1.00000 & 1.03889 & 1.92000 & 1.66964 & 0.80357 \\ 0.63375 & 1.87778 & 1.22760 & 1.65000 & 1.00000 & 3.01714 & 1.68367 & 1.03469 \\ 0.88542 & 1.31173 & 0.78704 & 1.67778 & 1.57407 & 1.00000 & 1.34921 & 0.71905 \\ 0.91071 & 1.22470 & 0.94947 & 2.39062 & 1.59375 & 3.82500 & 1.00000 & 1.36607 \\ 1.48437 & 2.46296 & 1.48437 & 3.59375 & 2.63889 & 3.68000 & 2.48810 & 1.00000 \end{pmatrix},$$

$$A = \begin{pmatrix} 1.00000 & 1.32075 & 1.00000 & 1.88235 & 1.57791 & 1.12941 & 1.09804 & 1.00000 \\ 0.75714 & 1.00000 & 0.78936 & 1.00000 & 1.00000 & 1.00000 & 1.00000 & 1.00000 \\ 1.00000 & 1.26685 & 1.00000 & 1.00392 & 1.00000 & 1.27059 & 1.05322 & 1.00000 \\ 0.53125 & 1.00000 & 0.99609 & 1.00000 & 1.00000 & 1.00000 & 1.00000 & 0.80357 \\ 0.63375 & 1.00000 & 1.00000 & 1.00000 & 1.00000 & 1.00000 & 1.00000 & 1.00000 \\ 0.88542 & 1.00000 & 0.78704 & 1.00000 & 1.00000 & 1.00000 & 1.00000 & 0.71905 \\ 0.91071 & 1.00000 & 0.94947 & 1.00000 & 1.00000 & 1.00000 & 1.00000 & 1.00000 \\ 1.00000 & 1.00000 & 1.00000 & 1.24444 & 1.00000 & 1.39073 & 1.00000 & 1.00000 \end{pmatrix}.$$

The weights and the ranking orders obtained by the tested methods are as follows:

Table 7

Weights and ranking of the DMUs in Example 3

DMU	Weights				Ranking			
	EM1	EM2	LLSM	WLSM	EM1	EM2	LLSM	WLSM
DMU_1	0.15290	0.22575	0.21761	0.26630	1	1	1	1
DMU_2	0.11616	0.09546	0.09978	0.08828	6	6	6	6
DMU_3	0.13279	0.13641	0.14839	0.15489	3	3	3	3
DMU_4	0.11203	0.08305	0.08474	0.07254	8	8	7	7
DMU_5	0.11729	0.10275	0.10874	0.10563	5	5	5	4
DMU_6	0.11391	0.08334	0.07143	0.04668	7	7	8	8
DMU_7	0.12172	0.11261	0.10932	0.09334	4	4	4	5
DMU_8	0.13318	0.16063	0.16001	0.17233	2	2	2	2

The ranking orders by EM1, EM2 and LLSM coincide in the first six positions. WLSM renders a further rank reversal in positions 4 and 5. The greater separation in the weights by EM2, LLSM and WLSM, in comparison to those by EM1, can be observed at this example, too. We mention that the weight vector w published in [23] is w =(0.152, 0.117, 0.134, 0.112, 0.118, 0.114, 0.122, 0.133)^T yielding the ranking (1,6,2,8,5,7,4,3). The slight difference to the weights obtained by EM1 may come from a larger stopping tolerance when computing the maximal eigenvector in [23]. But even a slight difference implies a rank reversal for DMU_3 and DMU_8.

Conclusions

The method proposed in this paper seems to be a promising new tool for ranking DMUs. It keeps the idea of using DEA-based pairwise comparisons between the decision making units (DMUs), as was proposed originally in [40]. The basic new idea is to apply a variant of the CCR problem instead of the classic one. The ensuing scores are then utilized to build a nonreciprocal pairwise comparison

matrix which serves as the basis of eliciting the ranking values of the DMUs. The main advantage of this new method is the wider range of the resulting ranking values, which subsequently leads to better distinction between the DMUs. This useful property was also confirmed by numerical examples. In addition to the eigenvector method, optimization based methods, such as the logarithmic least squares method and the weighted least squares method, were also tested for eliciting the ranking values from the nonreciprocal pairwise comparison matrix.

The numerical examples show that applying the new variant of the CCR problem and eliciting ranking weights from nonreciprocal pairwise comparison matrices remedy some shortcomings of the original method proposed in [40]. On the other hand, based upon the numerical examples, one cannot give a definite answer to the question of which of the techniques tested for eliciting ranking weights from nonreciprocal pairwise comparison matrices is the best. This question is argued but is undecided in the more general multiattribute decision making, too.

Acknowledgement

This work is connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" supported by the New Hungary Development Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002). The authors also gratefully acknowledge the support provided by the National Development Agency and the Hungarian National Scientific Research Fund (OTKA CNK 78168 and in part OTKA grant K 77420).

References

[1] Adler, N., Berechman, J., Measuring Airport Quality from the Airlines' Viewpoint: an Application of Data Envelopment Analysis, Transport Policy 8:171-181, 2001

[2] Adler, N., Friedman, L., Sinuany-Stern, Z., Review of Ranking Methods in the Data Envelopment Analysis Context, European Journal of Operation Research 140:249-265, 2002

[3] Andersen, A., Petersen, C. N., A Procedure for Ranking efficient Units in DEA, Management Science 39:1261-1264, 1993

[4] Bardhan, I., Bowlin, W. F., Cooper, W. W., Sueyoshi, T., Models for Efficiency Dominance in Data Envelopment Analysis. Part I: Additive Models and MED Measures, Journal of the Operations Research Society of Japan 39:322-332, 1996

[5] Bazargan, M., Vasigh, B., Size Versus Efficiency: a Case Study of US Commercial Airports, Journal of Air Transport Management 9:187-193, 2003

[6] Blankmeyer, E., Approaches to Consistency Adjustments, Journal of Optimization Theory and Applications 54:479-488, 1987

[7] Bozóki, S., Solution of the Least Squares Method Problem of Pairwise Comparisons Matrices, Central European Journal of Operations Research 16:345-358, 2008

[8] Carrizosa, E., Messine, F., An Exact Global Optimization Method for Deriving Weights from Pairwise Comparison Matrices, Journal of Global Optimization 38:237-247, 2007

[9] Charnes, A., Cooper, W. W., Rhodes, E., Measuring the Efficiency of Decision Making Units, European Journal of Operational Research 2:429-444, 1978

[10] Chen, Y., Sherman, H. D., The Benefits of Non-Radial vs. Radial Super-Efficiency DEA: an Application to Burden-Sharing Amongst NATO Member Nations, Socio-Economic Planning Sciences 38:307-320, 2004

[11] Chu, A. T. W., Kalaba, R. E., Spingarn, K., A Comparison of Two Methods for Determining the Weight Belonging to Fuzzy Sets, Journal of Optimization Theory and Applications 4:531-538, 1979

[12] Crawford, G., Williams, C., A Note on the Analysis of Subjective Judgment Matrices, Journal of Mathematical Psychology 29:387-405, 1985

[13] De Jong, P., A Statistical Approach to Saaty's Scaling Method for Priorities, Journal of Mathematical Psychology 28:467-478, 1984

[14] Diaz-Balteiro, L., González-Pachón, J., Romero, C., Forest Management with Multiple Criteria and Multiple Stakeholders: An Application to Two Public Forests in Spain, Scandinavian Journal of Forest Research 24(1): 87-93, 2009

[15] Doyle, J. R., Green, R., Efficiency and Cross-Efficiency in Data Envelopment Analysis: Derivatives, Meanings and Uses, Journal of the Operational Research Society 45(5):567-578, 1994

[16] Du, J., Liang, L., Zhu, J., A Slacks-based Measure of Super-Efficiency in Data Envelopment Analysis: A comment, European Journal of Operational Reearch 204:694-697, 2010

[17] Farkas, A., The Analysis of the Principal Eigenvector of Pairwise Comparison Matrices, Acta Polytechnica Hungarica, 4(2):99-116, 2007

[18] Farkas, A., Lancaster, P., Rózsa, P., Consistency Adjustment for Pairwise Comparison Matrices, Numerical Linear Algebra with Applications 10:689-700, 2003

[19] Farkas, A., Rózsa, P., On the Non-Uniqueness of the Solution to the Least Squares Optimization of Pairwise Comparison Matrices, Acta Polytechnica Hungarica 1:1-22, 2004

[20] Farrel, M. J., The Measurement of Productive Efficiency, Journal of Royal Statistical Society A 120:253-281, 1957

[21] Fülöp, J., A Method for Approximating Pairwise Comparison Matrices by Consistent Matrices, Journal of Global Optimization 42:423-442, 2008

[22] Fülöp, J., Koczkodaj, W. W., Szarek, S. J., On Some Convexity Properties of the Least Squares Method for Pairwise Comparisons Matrices without the Reciprocity Condition, Journal of Global Optimization, 2012 (in print)

[23] Guo, J., Jia, L., Qiu, L., Research on Supply Chain Performance Evaluation based on DEA/AHP Model, Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06), 0-7695-2751-5/06, 2006

[24] Hovanov, N. V., Kolari, J. W., Sokolov, M. V., Deriving Weights from General Pairwise Comparison Matrices, Mathematical Social Sciences 55:205-220, 2008

[25] Jahanshahloo, G. R., Junior, H. V., Lotfi, F. H., Akbarian, D., A New DEA Ranking System Based on Changing the Reference Set, European Journal of Operational Research 181:331-337, 2007

[26] Jensen, R. E., Comparison of Eigenvector, Least Squares, Chi Squares and Logarithmic Least Squares Methods of Scaling a Reciprocal Matrix, Working paper 153, Trinity University, 1983

[27] Jensen, R. E., Alternative Scaling Method for Priorities in Hierarchical Structures, Journal of Mathematical Psychology 28:317-332, 1984

[28] Koczkodaj, W. W., Orlowski, M., Computing a Consistent Approximation to a Generalized Pairwise Comparisons Matrix, Computers and Mathematics with Applications 37(2):79-85, 1999

[29] Lee, H.-S., Chu, C.-W., Zhu, J., Super-Efficiency DEA in the Presence of Infeasibility, European Journal of Operational Research 213(1):359-360, 2011

[30] Lotfi, F. H., Jahanshahloo, G. R., Esmaeili, M., Sensitivity Analysis of Efficient Units in the Presence of Non-Discretionary Inputs, Applied Mathematics and Computation 190:1185-1197, 2007

[31] Markovits-Somogyi, R., Data Envelopment Analysis and Its Key Variants Utilized in the Transport Sector, Periodica Polytechnica Ser. Transport, 39 (2):1-6, 2011

[32] Markovits-Somogyi, R., Gecse, G., Bokor, Z., Basic Efficiency Measurement of Hungarian Logistics Centres Using Data Envelopment Analysis, Periodica Polytechnica, Social and Management Sciences, 19 (2): 97-101, 2011

[33] Oral, M., Kettani, O., Lang, P., A Methodology for Collective Evaluation and Selection of Industrial R&D Projects, Management Science 37(7):871-885, 1991

[34] Royendegh, B. D., Erol, S., A DEA-ANP Hybrid Algorithm Approach to Avaluate a University's Performance, International Journal of Basic & Applied Sciences 9(10):115-129, 2009

[35] Rouyendegh, B. D., Erol, S., The DEA—FUZZY ANP Department Ranking Model Applied in Iran Amirkabir University, Acta Polytechnica Hungarica 7(4):103-114, 2010

[36] Saaty, T. L., The Analytic Hierarchy Process, McGraw-Hill, New York, 1980

[37] Seiford, L. M., Zhu, J., Infeasibility of Super-Efficiency Data Envelopment Analysis Models, INFOR 37:174-187, 1999

[38] Sekitani, K., Yamaki, N., A Logical Interpretation for the Eigenvalue Method in AHP, Journal of the Operations Research Society of Japan 42:219-232, 1999

[39] Sexton, T. R., Silkman, R. H., Hogan, A. J., Data Envelopment Analysis: Critique and Extensions. In: Silkman, R. H. (Ed.) Measuring Efficiency: An Assessment of Data Envelopment Analysis, Jossey-Bass, San Francisco, CA, 73-105, 1986

[40] Sinuany-Stern, Z., Mehrez, A., Hadad, Y., An AHP/DEA Methodology for Ranking Decision Making Units. International Transactions in Operational Research 7(2):109-124, 2000

[41] Tone, K., A Slacks-based Measure of Super-Efficiency in Data Envelopment Analysis, European Journal of Operational Research 143:32-41, 2002

[42] Torgersen, A. M., Forsund, F. R., Kittelsen, S. A. C., Slack-adjusted Efficiency Measures and Ranking of Efficient Units, The Journal of Productivity Analysis 7:379-398, 1996

[43] Wang, Y.-M., Luo, Y., Lan, Y.-X., Common Weights for Fully Ranking Decision Making Units by Regression Analysis, Expert Systems with Applications, 38: 9122-9128, 2011

[44] Wen, M., Li, H., Fuzzy Data Envelopment Analysis (DEA): Model and Ranking Method, Journal of Computational and Applied Mathematics 223:872-878, 2009

[45] Wu, Y. C. J., Goh, M., Container Port Efficiency in Emerging and More Advanced Countries, Transportation Research Part E 46(6):1030-1042, 2010

[46] Zhang, H., Li, X., Liu, W., An AHP/DEA Methodology for 3PL Vendor Selection in 4PLW, Shen et al. (Eds.): CSCWD 2005, LNCS 3865, Springer-Verlag Berlin Heidelberg, 646-655, 2006

Mobile Detection Algorithm in Mobile Device Detection and Content Adaptation

Zlatko Čović[1], Miodrag Ivković[2], Biljana Radulović[2]

[1] Subotica Tech – College of Applied Sciences, Department of Informatics
Marka Oreškovića 16, 24000 Subotica, Serbia, chole@vts.su.ac.rs

[2] Technical Faculty "Mihajlo Pupin", Department of Informatics
Đure Đakovića BB, 23000 Zrenjanin, Serbia
E-mails: misa.ivkovic@gmail.com; bradulov@ptt.rs

Abstract: This paper describes several approaches in the development of mobile web sites, including the concept of detection and adaptation for mobile content. A mobile site lacking optimization can lead to prolonged downloading of data as well as difficulties with browsing. For that reason, we need to optimize web content on such sites. Our research was conducted to determine the most adequate and most effective detection technique. This technique was then used in the development of the Mobile Detection Algorithm Based on Tera-Wurfl (MDABTW). The MDABTW algorithm is based on the Tera-Wurfl library and allows for the generation of web content in several markup languages. The basic principles of how this algorithm works are described in this paper. The MDABTW algorithm was further implemented in the regional mobile CMS, called Mobko, which is the main part of an integrated health IS providing counseling and education services for youth in the Subotica region.

Keywords: mobile web sites; mobile device detection; content adaptation; mobile algorithm

1 Introduction

The Web has revolutionized the way we interact, as well as how we collect and publish information. Until recently, web content was available only to desktop users. With the advent of mobile technology and with the proliferation of mobile devices featuring Internet access, the global availability of information has exploded. Standard (2G) mobile phones and smart phones (3G) have advanced so rapidly that nowadays even an entry level mobile phone allows for Internet access of some kind. This situation opens up a whole new arena for web content that has been adapted specifically for mobile devices [7].

Today, increasing numbers of people access the Internet via their mobile device instead of a PC. The number of mobile devices in use has already surpassed the

number of personal computers. It is estimated that the difference between these two numbers will only increase in the years to come [9].

The first problem that can occur when creating a mobile web site is how to distinguish between mobile users and desktop users.

2 Mobile Web Sites and the Problem with Content

Most often we have seen that there is no optimization and adaptation of existing sites or new sites for mobile users. The web content might be too wide to fit the screen of mobile device – user equipment (UE). Font size can also be problematic, being too small or too large to read on UE easily. If the web content also includes images, it is uncertain whether those can be displayed correctly, if at all. The same applies to multimedia files, which frequently cannot be viewed on mobile devices. Often, web pages that have been initially designed for desktop computers are too encumbered with content, so they are practically unsuitable for users accessing them via mobile devices.

Creating content for mobile devices has some specific restrictions imposed by the limited resources on the UE, including: a small size screen, limited wireless bandwidth, small data storage capacity and processing power, and a limited number of keyboard buttons to be used for web navigation. Sites designed for mobile devices should deliver content that is tailored to the characteristics of the accessing device. In order for a mobile content to load faster, they should be adapted and formatted, taking into consideration both the users' needs as well as the capabilities of mobile devices.

Mobile web sites can be created by using several technologies (e.g. WML, HTML, and XHTML MP). Existing mobile devices have different levels of support for these technologies, and for this reason, the creator of a mobile web site should create multiple versions of the web site. This means that the developer should be very familiar with all the required technologies in order to implement them properly. This could be an issue as it requires additional investment in development and learning.

3 Content Adaptation

Content adaptation, which is sometimes also called multi-serving, is the delivery of content based on the capabilities of the access device.

Detection, adaptation and support for multiple devices have historically been a painful point in design and development of mobile content. Although there are

several strategies and techniques available to tackle this problem, each of them use in some form the DAD (*Detect Deliver Text*) mode. This mode includes:

1 Detection
2 Adaptation
3 Deliver

Different techniques are used for the adaptation, including the detection, redirection, set up of correct MIME types, the changing of links, and the removal or scaling graphics. The LCD method (or the "Lowest Common Denominator") establishes a minimum set of characteristics which are expected from the mobile device. Web content is then developed by following such guidelines. The minimal set of features is also called DDC (Default Delivery Context) [4].

Adaptation in accordance with the capabilities of accessing mobile device is an ideal solution for the delivery of mobile content. At the same time, most programmers prefer to first start with the LCD approach before going into adaptation. The reasons for this are many. Adaptation involves additional costs and complexity. It also requires changes on the server side to detect and deliver content in a tailored manner, which cannot be accomplished for all device types. However, the LCD method can be sufficient in the case of a limited use of mobile content [4].

Based on the paper by Adzic, Kalva and Furht [1], there are three possible levels at which the adaptation can be performed: 1) at the server which hosts web content; 2) at the intermediary proxy server; 3) and at the client side (i.e. UE). The main difference among the three lies in what supplies the content, and what the content is [12].

Server-side adaptation involves committing additional resources and software on the server that stores content for delivery. The main advantage of this approach is that the content at every web page can be converted and tailored for the specific needs of a receiving mobile device.

The proxy server or intermediary adaptation is performed between the server (content provider) and the mobile client (content consumer). It can be implemented by the content provider or by the third party device [1]. However, this form of adaptation is usually outside of your control.

In the client-side adaptation, the user can define preferences and determine the type and scale of the adaptation process [1]. Client side adaptation usually relies on using the media selectors associated with CSS (Cascading Style Sheets). While this can be a useful form of adaptation, it is limited. The mobile browser downloads the same XHTML markup as the desktop browser, which might unnecessarily consume the user's air time and consequently cost for the content that ends up not being displayed [5].

The following section of this paper will present several approaches in the development of mobile sites.

3.1 Approaches in Mobile Development

In the development of mobile web sites, there can be defined several approaches:

a) Do nothing
b) Remove formatting
c) Filter media
d) Find the target device - redirection
e) Implement the full detection and adaptation

Some of them use some kind of adaptation which is executed at different levels (i.e. server side, client side or at the intermediary proxy server).

3.1.1 Do Nothing

The first approach is still used for many sites because the site owner or developer does not fully understand the significance of mobile-oriented content or prefers to wait until the UE technology is mature enough to handle their web site as is. In other words, the web content gets published only once, whereas the mobile device is expected to be smart enough to handle it and display web pages as appropriate. The most important adjustment techniques are adopted in the Opera's Small Screen Rendering (SSR) system, used for the new Apple iPhone and Nokia Browser, which provides very good results. The new generation of mobile phones have web browsers which attempt to optimize the web content on the UE in real time. Often it only involves reducing the page views without reducing actual download time, thus contributing additional (and unnecessary) cost to the user.

The issue with this approach is that it does not work for standard (2G) non-smart phones, which do not have web browsers capable of optimizing content in real time. Considering that large number of users who still have regular mobile phones, this approach will deprive them of getting proper information. Even smart phone users, with browsers featuring optimization in real time, can sometime give up on accessing classic web sites because the site is either just not functional enough or has pure navigation system with does not allow for the easy finding of the desired information. In this approach the transcoders are often used to reformat the content of classical sites and transform them into a more mobile "friendly" version.

When to use this approach

- If you do not expect too many mobile visitors to your web site
- If the majority of your users are using smart phones or other devices featuring large screens and an optimized web browser
- If people want to use your website in exactly the same way as used via PC, (i.e. no need for mobile-specific features)
- If you do not have time or resources to implement another method(s)
- If you do not have sufficient expertise to include multiple technologies in your development

When not to use this approach

- If you want to reach the maximum number of potential visitors to your site
- If mobile users visiting your site have ~~a~~ specific task they want to accomplish quickly and efficiently
- If you want to provide the best browsing experience for mobile users

3.1.2 Remove Formatting

One of the biggest problems for mobile browsers is to parse the HTML code and page layout. Complex formatting rules mean more computer operations that may be a challenge due to UE limitations in regards to processor and memory. Most mobile web users pay for service on the basis of downloaded data in kilobytes; therefore large HTML pages and images will not be a good solution. If you remove the formatting from the page layout, including all images, the mobile user will get text and links only.

When to use this approach

- If you want a quick way to create a mobile version of your web site
- If you want to cover the majority of mobile browsers
- If your site features mostly textual content and has good navigation structure

When not to use this approach

- If your site features good user interface design
- If the site content is not very useful for mobile users

3.1.3 Filter Media

If you want to use the same XHTML site for both desktop and mobile devices, the solution is to change the site design and format related pages by using a CSS file. The CSS standard enables you to create multiple styles for each document, taking into account that site content might be displayed on various types of media. This approach involves creating two or more versions of CSS files and using those as media attribute to generate corresponding pages.

When to use this approach

- When you are confident that the access devices fully support the filtering of media
- When you make a mobile website that is dedicated to a specific group of users who have devices with the possibility of filtering media
- When you are familiar only with the client side of internet technologies

When not to use this approach

- When you are not sure whether all access devices support the filtering of media
- When you create a mobile web site that is expected to draw a large number of visitors with different types of phones

3.1.4 Find the Target Device – Redirection

This approach uses a detection technique to find the type of target device. Specifically, it tries to decide if the user is accessing the web site via a mobile device or via computer. If the detection is successful, the user is redirected to the appropriate version of the site (mobile or desktop). An enhanced version of this strategy involves the detection of supported Markup Language.

When to use this approach

- If you only need to determine whether the user accesses you site via mobile devices or desktop computers
- If it is not necessary to create content that is fully optimized for the characteristics of the accessing device
- If the user provides multiple choices to select the mobile version of the site because it does not detect the preferred hypertext markup language

When not to use this approach

- If you create a site whose content partly needs to be optimized for the features of the device
- If it is necessary to determine the preferred hypertext markup language

3.1.5 Implement the Full Detection and Adaptation

This approach requires adaptation in the framework by taking into consideration several related variables. Firstly, it requires the detection of the user agent and its characteristics. This is usually determined by:

- Preferred markup language
- Screen size in pixels
- The existence of the touch screen
- Support for certain graphics formats
- Support for certain multimedia files and Java
- Support for styles

Based on the detection of the preferred language and UE features, the user gets redirected to the corresponding site version. In this strategy, you need to create multiple versions of your site. In the event that the detection of the preferred markup language has failed or the user wants to switch to a different version of your site, the option should be provided for selecting the language for content generation.

When to use this approach

- If you want to create a site which fully supports the features of access device
- If you do expect a large number of visitors to your mobile site
- If you do not have to worry about the resources on your server

When not to use this approach

- If you have limited resources on your server
- If you have no need for partial or full optimization

3.2 Header of Request

A request header sent from the mobile browser to the web server has many attributes defined by the reader in the case of direct client-server communication (i.e. without using any proxy, gateway or a transcoder). Some of the headers that may be useful are listed in Table 1.

Table 1
Listing of useful headers

User-Agent	Name of browser, user agent or platform from which the request originates
Accept	Comma separated values (CSV) which are MIME types that are supported by user agent.
Accept-Charset	Specifies the character set that supports the user agent.
Accept-Language	Contains information about the supported languages in the user agent
X-wap-profile and Profile	Provides information about the UAProf (User Agent Profile) XML file that is unique to each device.

3.2.1 User Agent

The user agent is a client application that implements the network protocol used in the client-server communication. It typically contains a series of different information, from which can be identified: which device is used, on what operating system it is based, what version of software is running, who the manufacturer is, and what type of content the device is able to read. The format of information differs between manufacturers of mobile devices.

The User-Agent header is useful in the following situations:

- If we need to identify specific models of mobile devices.
- If we need to distinguish between mobile devices or user agents from different companies.
- If we need to determine whether the user agent browser to your desktop computer or a mobile device microreaders.

3.2.2 UAProof

The UAProf (*User Agent Profile*) is a standard defined by the *Open Mobile Alliance* (OMA, earlier WAP Forum). It provides information on UE capabilities in the form of an XML file which the server can access via the Internet. After this file has been uploaded onto server, it must be parsed in order to retrieve the desired information. The drawback of the *UAProf* specification is that many older devices do not support it. In addition, it can slow down the content generation on the server side as each application must retrieve and parse the XML file separately in order to obtain the pertinent information. A possible solution to speed up this process could be to store partial contents of the file in a database, after its initial upload, or to save the file local copy on the server [9].

3.2.3 WURFL

The WURFL (*Wireless Universal Resource File*) is a wireless universal resource file or an XML file that is regularly updated with information on mobile devices. It contains information on virtually all mobile devices that currently exist, and all their options. Practically, this file is a collection of a large number of UAProf files.

4 A Survey for Detection Techniques

As mentioned in Chapter 3, there are several ways to carry out the detection of an access device. Most of these detections can be classified into two general categories: a) detection using only information from the HTTP headers; and b) detection using the WURFL file. For the purpose of our research, we utilized both methods, namely:

1 The so-called “Simple detection”

2 Detection by using the Tera-Wurfl library.

For "Simple Detection" we used PHP script, which makes a distinction between computers and mobile devices by comparing the user agent data and values from HTTP headers with specific text values. This technique also determines the preferred hypertext markup language.

The advantages of this technique include:

- No installation process
- There is no need for a database
- Takes up less space on the server
- Fast execution time

The disadvantages of this technique are:

- Relatively poor accuracy
- No community that contributes to the development
- Limited set of functions to detect more capabilities
- Inability to complete the adaptation

Tera-Wurfl is a PHP- and MySQL-based library which uses WURFL to detect the individual features of mobile phones. The advantage of Tera-Wurfl over other systems for detection is that it relies on a database to find the best matching mobile phone. Tera-WURFL loads the data from the WURFL file into a MySQL database for faster access and uses it to determine which device is the most similar to the one requesting your content. The library then returns the capabilities

associated with such device to your scripts via PHP associative array. In our research we used the WURFL with 29 groups featuring a total of 531 capabilities.

Tera-WURFL takes the requestor's user agent and puts it through a filter to determine which UserAgentMatcher to use. Each UserAgentMatcher is specifically designed to best match the device from a group of similar devices based on the Reduction in String and/or the *Levenshtein Distance algorithm*.

The advantages of this technique are as follows:

- High performance
- Accurate detection of mobile devices
- Fast detection of desktop vs. mobile devices
- Possibility of full adaptation

The disadvantages of this technique are:

- Need for a database
- Requirement for more space on the server

Our survey involved 80 students. This survey was conducted at Subotica Tech in the period from November to December 2010. Each student was given access to web pages on a server via his mobile phone. One page used the "Simple Detection" technique and the second used detection based on the Tera-Wurfl library. The aim of this research was to check how exact these two detection techniques compare. The results are shown below.

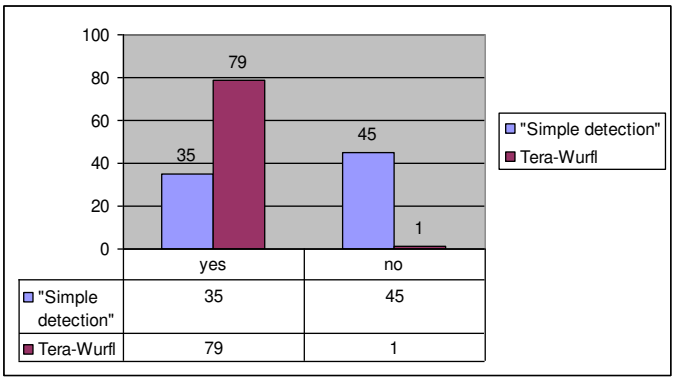


Figure 1
Result of accuracy of detection methods

Based on these results, it can be seen that the detection technique using the Tera-Wurfl library is more accurate and more suitable for implementation in mobile development.

5 Mobile Detection Algorithm Based on Tera-Wurfl (MDABTW)

The current research results described in Section 4 demonstrate that the Tera-Wurfl technique is the better one for detecting the characteristics of accessing devices. This library includes a large amount of data for each of the mobile devices, which is contained in the WURFL file.

The Tera-Wurl library also contains an optimizer for the WURFL file. The Optimizer is a very useful feature since, based on the selection of certain characteristics, it can reduce the file and database size, reduce the occupancy of space on the server and speed up the process of detection. In addition to optimization, device detection also requires the selection of significant device characteristics to be utilized for detection. For that purpose, we have developed the Mobile Detection Algorithm Based on Tera-Wurfl (MDABTW) algorithm. The MDABTW is an algorithm for the detection of mobile devices based on Tera-Wurfl library. It allows for the detection and generation of mobile content in WML, XHTML MP and XHTML languages, using WAP CSS or CSS as needed.

In this algorithm, web sites for mobile devices are divided into four groups: simple web sites – *SIMPLE,* multimedia web sites – *MULTIMEDIA*, dynamic web sites – *DYNAMIC* and integrated web sites – *INTEGRATED*.

We have implemented this algorithm in the mobile CMS called Mobko. Mobko plays a key role in an integrated model for mobile learning in the health information system and counseling services for youth in Subotica region.

The algorithm works as outlined in Fig. 2. After a visit to the website, the processing of the request is started by detection of accessing device. For this purpose we used user agent data. In the event that the access device is not a mobile device, it generates content for the standard web site, using XHTML language and CSS technique. In the case that the access device is mobile device, the MDABTW algorithm starts with one of subprograms. Which subprogram will be used depends on the chosen model of web site and the mobile web site group.

Every group has associated subgroups. Subgroups vary based on the application of certain elements. These groups differ according to the type of content that is delivered and the manner of interaction with the user.

For detection, the same procedure is used for each of the above groups. However, in the subsequent process of the adaptation and generation of web content, some special actions are taken based on the customized WURFL file containing the essential features. As mentioned earlier, the WURFL file used in the current case contained 29 groups with a total of 531 capabilities. This constitutes a large amount of information. For faster detection, it is better to use and optimize only the characteristics that are important for the kind of web site involved.

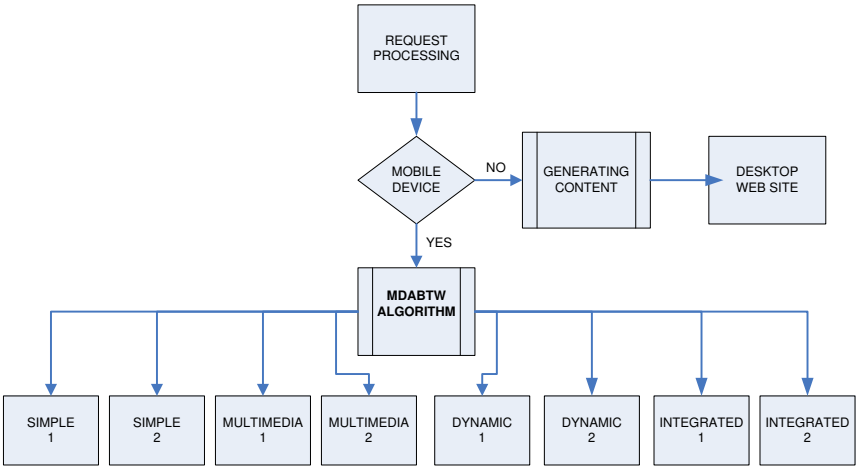


Figure 2
Detection process schema

Table 1
Short description of subprograms

Subprogram	**Description**
Simple1	Handles web sites with static content only.
Simple2	Handles web sites with static content and supports the use of web forms.
Multimedia1	Handles static web sites with multimedia content.
Multimedia2	Handles web sites with multimedia content and web forms in it.
Dynamic1	Handles dynamic web sites where the content is generated based on the data from database.
Dynamic2	Handles dynamic web sites where the content is generated based on the data from database and supports use of web forms.
Integrated1	Handles web sites which have multimedia and dynamic content.
Integrated2	Handles web sites which have multimedia and dynamic content and supports the use of web forms.

For each type of website, the detection process identifies only the selected capabilities which are site-specific rather than being group capabilities. However, there are also situations when the detection process needs to identify group capabilities as well.

These are the groups of capabilities used in the algorithm:

- Product_info (PI)
- Playback (P)
- Wap_push (WP)
- MMS (M2)
- Display (D)
- Image_format (IF)

- Wml_ui (WU)
- Xhtml_ui (XU)
- Html_ui (HU)
- Streaming (S)
- Css (C)
- Markup (M)
- Bugs (B)
- Object_download (OD)
- Sound_format (SF)
- Ajax (A)

Table 2 provides an overview of the subprograms and capability groups where F indicates when "Full", P indicates when "Partial" detection is done and the sign "-" indicates that this group is not used.

Table 2

List of capability groups and subprogram

	Capability group															
Subprogram	**PI**	**P**	**WP**	**M2**	**D**	**IF**	**WU**	**XU**	**HU**	**S**	**C**	**M**	**B**	**OD**	**SF**	**A**
Simple1	P	-	-	-	F	P	P	P	P	-	P	P	-	-	-	-
Simple2	P	-	-	-	F	P	P	P	P	-	P	P	P	-	-	-
Multimedia1	P	F	F	F	F	F	P	P	P	F	P	P	-	F	F	-
Multimedia2	P	F	F	F	F	F	P	P	P	F	P	P	P	F	F	-
Dynamic1	P	-	-	-	F	P	P	P	P	-	P	P	-	-	-	F
Dynamic2	P	-	-	-	F	P	P	P	P	-	P	P	P	-	-	F
Integrated1	P	F	F	F	F	F	P	P	P	F	P	P	-	-	F	F
Integrated2	P	F	F	F	F	F	P	P	P	F	P	P	P	F	F	F

5.1 Description of Subprogram

The Simple1 subprogram works based on the following principles:

1 Firstly, the device pointing method is detected, which means that the access device can be either a phone with a touch screen or a standard mobile phone without touch screen. This information is very important because it drives the design of the complete navigation structure for a web site.

2 The next step is to use an appropriately optimized WURFL file for the detection of the device's capabilities. This includes the preferred markup language, DTD (Document Type Definition) and MIME type.

3 If the access device is a standard mobile phone without a touch screen, after the capabilities detection in point 2, the algorithm checks if the preferred markup language is WML. This step is missing in mobile devices that have touch screens because these phones in most cases do not support the WML language and the performance of these phones impose a need to create richer content.

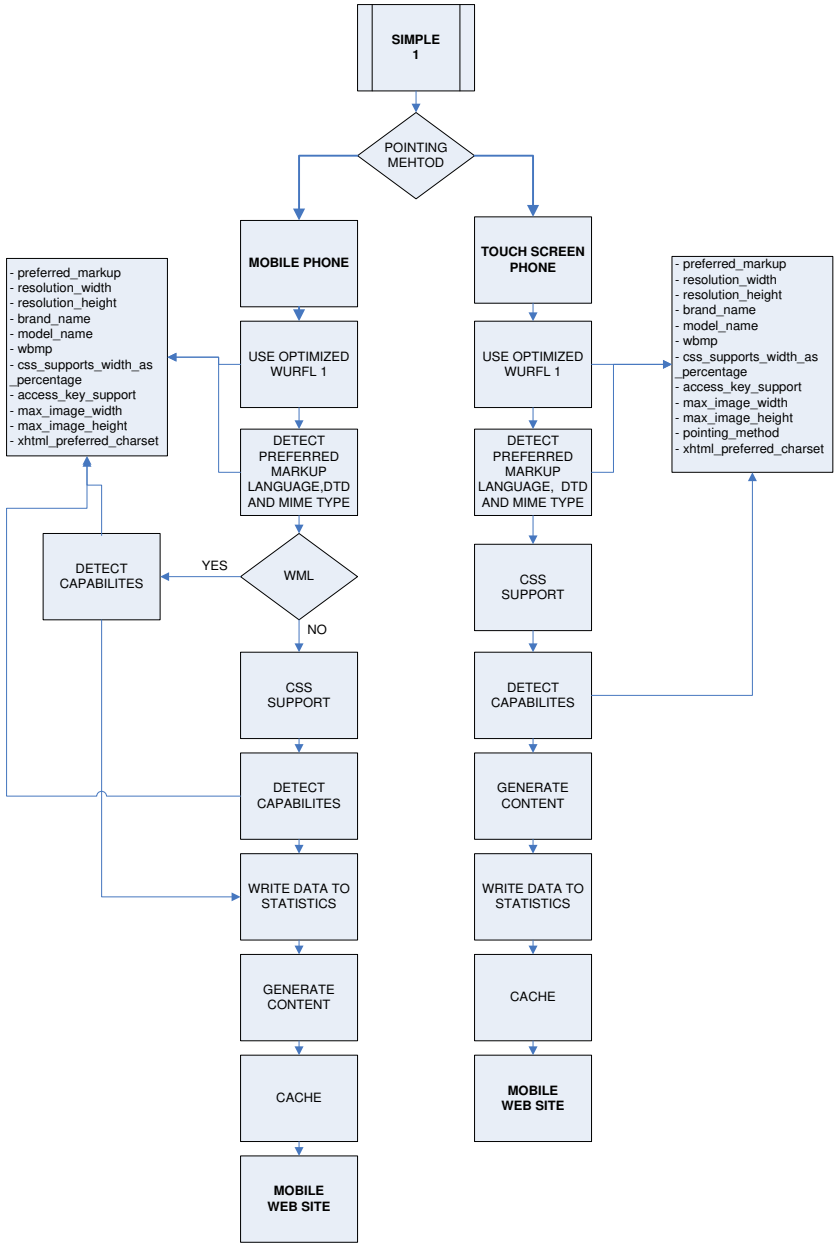


Figure 3
Working schema of Simple1 subprogram

4 If these data are related to the WML language, there is no need to determine support for the CSS as WML does not support it. After that, the capabilities that are necessary for WML content generation are considered detected.

5 In the case when detected data do not belong to WML, the CSS support is important since, based on that, the device capabilities are detected.

6 After detection, data are written in the database for statistics and the optimized content of the web site is generated. The optimized content could be generated in any of three versions: WML, XHTML or XHTML MP.

7 The generated site is saved in the cache memory for faster access next time

8 Finally, the user is provided with access to the optimized mobile web site.

Simple2 subprogram differs from the Simple1 in regard to the detection of input masks as well as some other capabilities that are significant for the use on web forms. The other subprograms work like the Simple1 subprogram but the main difference lies in the characteristics that are detected.

5.2 Configuration and Implementation

The configuration file *TeraWurflConfig.php* needs to be set properly in order to use one of the chosen subprograms. The following parameters are the most important ones:

```
public static $WURFL_FILE = 'wurfl-integrated2.xml';

public static $LOG_FILE = 'wurfl-integrated2.log';
```

The first parameter contains the name of the optimized WURFL file while the second parameter is a log file for this WURFL file. These settings are enough when installing the system according to the tutorial and updating WURFL database from the local optimized WURFL file. After the initial installation, there is the possibility to update the WURFL database with the latest version of this xml file from the internet. As the optimized files are used in the algorithm, one does not want to update and insert all data of the WURFL file from internet, just the essential characteristics for every subprogram. For that purpose, a $CAPABILITY_FILTER variable within the *TeraWurflConfig.php* configuration file should be set. This variable is of the array type. It is possible to set complete capability groups or individual capabilities in this variable. To this end, every subprogram configures the specific parameters in a different way.

In the following code, the settings for the *Simple1* subprogram are shown.

```
public static $CAPABILITY_FILTER = array(
"pointing method",
"model name",
"is_wireless_device",
"brand_name",
```

```
"wml_make_phone_call_string",
"access_key_support",
"xhtml_preferred_charset",
"xhtml_avoid_accesskeys",
"xhtml_make_phone_call_string",
"xhtml_display_accesskey",
"xhtmlmp_preferred_mime_type",
"css_supports_width_as_percentage",
"preferred_markup",
"max_image_width",
"resolution_height",
"resolution_width",
"max_image_height",
"jpg",
"gif",
"wbmp",
"png",
);
```

With these parameters, TeraWURFL allows one to store only the specified capabilities from the WURFL file.

The following code presents the usage. First the session is started, then the important files are included. In the next step, the instance of the Tera-WURFL object is created, and data from the HTTP_USER_AGENT header is received. This data helps to determine if the user comes with mobile device (is_wireless_device) or not. If not, the program redirects the user to a non-mobile version of the page. If the user agent is a mobile device, the associative array $_SESSION[‘d_c’] is set and the relevant characteristics for the used subprogram are detected and put in it. This array contains only the characteristics that are important for the generation of the web page layout and they are available in the whole web page due to the use of session variable.

```
// start the session
session_start();
// include the necessary files
include("include/config.php");
include("include/functions.php");
// include the Tera-WURFL file
require_once('../../Tera-Wurfl/TeraWurfl.php');
// Instantiate the Tera-WURFL object
$wurflObj = new TeraWurfl();
// Get the data from HTTP_USER_AGENT header
$ua = $_SERVER["HTTP_USER_AGENT"];
// Get the capabilities from the object
$matched = $wurflObj->getDeviceCapabilitiesFromAgent($ua);
if(!check_session_variables())
{
if(!$wurflObj->getDeviceCapability("is_wireless_device")){
        header("Location:../index.php");
        exit();
}
//Get device capabilities
$_SESSION['d_c']=array();
$_SESSION['d_c']['pointing_method'] = $wurflObj-
>getDeviceCapability("pointing_method");
...
```

5.3 Testing the Algorithm in Practice

To test the MDABTW algorithm, it was used in an integrated learning system within youth counseling in Subotica. Figure 5 show the results of this algorithm implementation for two types of commercial handsets. On the right side there is the screenshot of the Nokia E51 phone (featuring a standard - non-touch screen) whereas the left side contains the screen shot of the Huawei U8110 phone (with touch screen option). These 2 phones differ in many aspects, including: preferred markup language, DTD, markup MIME type, screen size, pointing method, etc. Despite all these differences, Fig. 5 clearly demonstrates the complete success of the *MDABTW* algorithm at work. Namely, the process involving device detection and generating optimized content with the graphic files and text.

The most significant form of successful implementation of the algorithm is in the reduction of the size of the WURFL file and the MySQL database and the reduction in the size of the optimized web content. The size of the non-optimized WURFL file is 16 MB. After data entry from this file to the database, the MySQL database takes up 29 MB of space on the server. Table 3 presents the results of the reduction with the use of the MDABTW algorithm in every subprogram.

Table 3

Results of reduction with the use of MDABTW algorithm

Subprogram	Size of optimized WURFL file	Size of used MySQL database	Reduction of WURFL file in percentages	Reduction of MySQL database in percentages
Simple1	5.58 MB	17.2 MB	67%	41%
Simple2	5.62 MB	17.2 MB	66%	41%
Multimedia1	12.80 MB	21.3 MB	24%	27%
Multimedia2	12.91 MB	21.4 MB	23%	26%
Dynamic1	5.66 MB	17.2 MB	66%	41%
Dynamic2	5.7 MB	17.3 MB	66%	41%
Integrated1	12.99 MB	24.2 MB	23%	17%
Integrated2	13.03 MB	24.3 MB	22%	16%

Applying the MDABTW algorithm reduces the size of the generated files that are delivered to the client. The following table presents the size of the files generated for different access devices with different characteristics

Table 4

Size of optimized web content for different user agents with the use of MDABTW algorithm

	index.php	servis.php	lekcije.php	testovi.php	login.php	arhiva.php
Web browser	1296 KB	1295 KB	1264 KB	1295 KB	-	1299 KB
NokiaE51	8.3 KB	8.7 KB	9.8 KB	11.3 KB	8.5 KB	50 KB
Huawei U8110	10.4 KB	9.8 KB	12 KB	13.8 KB	9.7 KB	52.1 KB

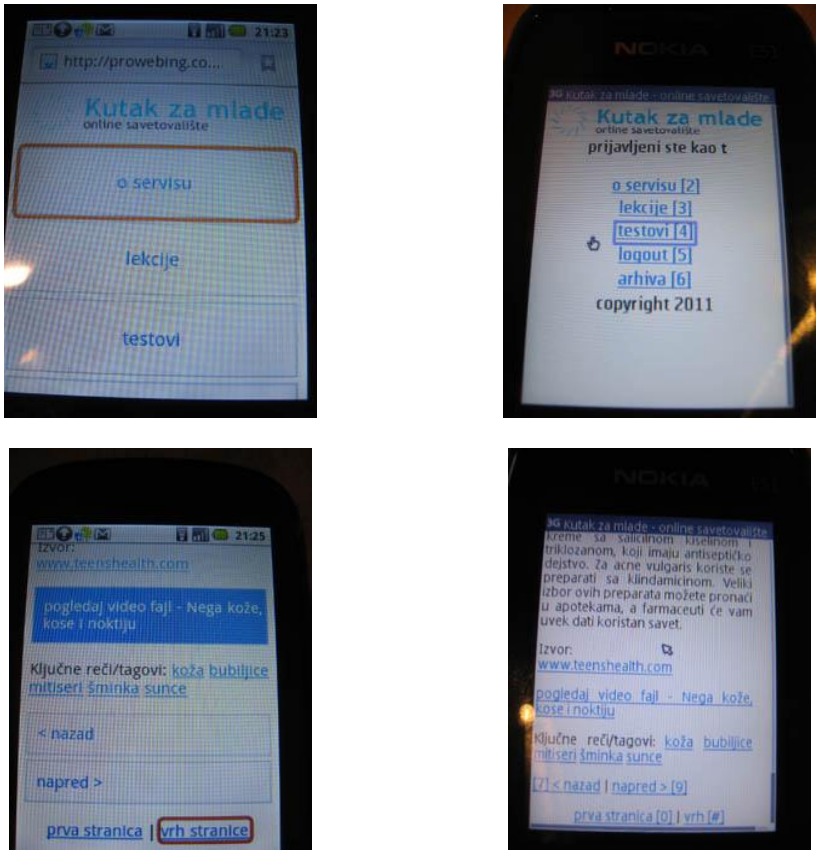


Figure 5
Algorithm in practice – optimized content for HuaweiU8110 and Nokia E51

Conclusions

This paper describes several approaches for the development of mobile web sites, including the concept of device type detection and content adaptation. Many web pages that have been originally designed for desktop computers are currently too encumbered with content, which makes them practically unsuitable for users accessing via mobile devices. For that reason, we need to optimize the content of such web sites. On the basis of our research, the appropriate detection technique was selected and used for creation of the *MDABTW* mobile detection algorithm. This algorithm utilizes the Tera-Wurfl library and allows for the generation of web content in WML, XHTML MP and XHTML languages, using WAP CSS or CSS as needed. Support for HTML5 and CSS3 can be easily incorporated into the *MDABTW* algorithm when these techniques become a widely adopted standard.

After initial testing and tuning, the *MDABTW* algorithm was implemented for the regional mobile CMS called Mobko – part of an integrated system for health counseling of young people in the Subotica region. This system includes the following: a) CMS Mobko; b) Web portal access to counseling services for the PC that enables viewing and downloading of educational materials and testing; c) Mobile version of the Internet portal for access to counseling services by mobile devices, where all content will be adjusted and optimized for the possibility of a user's mobile device; d) Stand-alone application for mobile phones allowing for training and testing in an on-line or off-line mode as well as for the SMS service used to distribute important information.

Our next goal is to investigate further the use of the mobile learning services within integrated IS and how it impacts the health education of the younger population, who are known to be technology savvy. The result of this planned investigation will be compared to other available theoretical and practical research results. In addition, we also plan to conduct several surveys intended to define guidelines for the future development and improvement of the existing *MDABTW* algorithm.

References

[1] Adzic, V., Kalva, H., Furht, B.: A Survey of Multimedia Content Adaptation for Mobile Devices, Multimedia Tools and Applications Journal, Vol. 51, No. 1, DOI: 10.1007/s11042-010-0669-x, 2011, pp. 379-396

[2] Xinyou, Z., Toshio, O.: A Device-Independent System Architecture for Adaptive Mobile Learning, Eighth IEEE International Conference on Advanced Learning Technologies (ICALT 2008), 2008, pp. 22-25

[3] Bianchi, D., Mordonini, M.: Content Adaptation for M-learning, Proceedings of the IADIS International Conference on Mobile Learning, Qwara, Malta, 28-30 June, 2005, pp. 246-250

[4] Firtman, M.: Programming the Mobile Web, O'Reilly, 2010

[5] Fling, B.: Mobile Design and Development, O'Reilly, 2010

[6] Liu, Y., Yang, Z., Deng, X., Bu, J., Chen, C.: Media Browsing for Mobile Devices based on Resolution Adaptive Recommendation, International Conference on Communications and Mobile Computing, 2009, pp. 285-290

[7] Čović, Z., Horvat Cinger, N., Ivković, M.: Mobile Learning in Practice, Proceedings of the 11th International Symposium on Computational Intelligence and Informatics, CINTI 2010, Budapest, Hungary, November 18-20, 2010, IEEE Catalog Number: CFP1024M-PRT, ISBN: 978-1-4244-9278-7, pp. 315-318

[8] Cserkúti, P., Szabó, Z., Eppel, T., Pál, J.: SmartWeb – Web Content Adaptation for Mobile Devices, Proceedings of 4th Slovakian-Hungarian

Joint Symposium on Applied Machine Intelligence, SAMI 2006, Herlany, Slovakia, January 20-21, 2006, pp. 1-11

[9] Čović, Z., Ivković, M.: Adaptation of Content for Mobile Web Sites, Proceedings of the 17th International Conference on Computer Science and Information Technology YU INFO 2011, March 6-9, Kopaonik, Serbia, 2011, pp. 1-4

[10] Metso, M., Löytynoja, M., Korva, J., Määttä, P., Sauvola, J.: Mobile Multimedia Services-Content Adaptation, 3rd International Conference on Information, Communications and Signal Processing, Singapore (CD-ROM), 2001, pp. 1-6

[11] Content Management for Wireless Networks, 2008-2013 - Insight Research Report

[12] Laakko, T., Hiltunen, T.: Adapting Web Content to Mobile User Agents, IEEE Internet Computing, Vol. 9, Issue 2, 2005, pp. 46-53

[13] Yang, S. J. H., Chen, I. Y. L., Chen, R.: Applying Content Adaptation to Mobile Learning, Proc. of Innovative Computing, Information and Control, 2007, ICICIC '07, Kumamoto, Japan, September 5-7, 2007, pp. 1-4

[14] Content Management for Mobile Delivery, Posted by Apoorv, PCM.Blog, May 26, 2007

[15] Tomášek, M.: Language for a Distributed System of Mobile Agents, Acta Polytechnica Hungarica Journal, Vol. 8, No. 2, 2011, pp. 117-127

[16] Čović, Z. Blažin, J. Ivković, M.: Implementation of Integrated Learning System Within Youth Counseling, Proceedings of the 9th International Symposium on Intelligent Systems and Informatics (SISY 2011) Subotica, Serbia, September 8-10, 2011, DOI: 10.1109/SISY.2011.6034377, ISBN: 978-1-4577-1975-2, pp. 489-494

Different Chromosome-based Evolutionary Approaches for the Permutation Flow Shop Problem

Krisztián Balázs[1], Zoltán Horváth[2], László T. Kóczy[1,3]

[1] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary, balazs@tmit.bme.hu

[2] Department of Mathematics and Computational Sciences, Széchenyi István University, Győr, Hungary, horvathz@sze.hu

[3] Department of Automation, Széchenyi István University, Győr, Hungary, koczy@sze.hu

Abstract: This paper proposes approaches for adapting chromosome-based evolutionary methods to the Permutation Flow Shop Problem. Two types of individual representation (i.e. encoding methods) are proposed, which are applied on three different chromosome based evolutionary techniques, namely the Genetic Algorithm, the Bacterial Evolutionary Algorithm and the Particle Swarm Optimization method. Both representations are applied on the two former methods, whereas one of them is used for the latter optimization technique. Each mentioned algorithm is involved without and with local search steps as one of its evolutionary operators. Since the evolutionary operators of each technique are established according to the applied representation, this paper deals with a total number of ten different chromosome-based evolutionary methods. The obtained techniques are evaluated via simulation runs carried out on the well-known Taillard's benchmark problem set. Based on the experimental results the approaches for adapting chromosome based evolutionary methods are compared to each other.

Keywords: Chromosome-based evolutionary methods; Memetic algorithms; Combinatorial optimization; Permutation Flow Shop Problem

1 Introduction

One of the most intensively studied combinatorial optimization problems is the Permutation Flow Shop Problem (PFSP) [1]. In this problem, there are given n jobs and m machines. All the jobs should be processed by all the machines one after another. The machines are deployed in a line and a machine can handle one

single job at once, that is, the process of the jobs is pipeline-like. There is also given an *n*-by-*m* processing time matrix defining the necessary amount of time a job has to stay on a machine, for each job-machine pair. A job can be processed on any machine only if the machine is free (the preceding job has finished on the machine) and the job has already been processed on the preceding machine.

The task is to find a permutation (a sequence) of the jobs, in case of which the total processing time of all the jobs on all the machines (i.e. the so-called makespan) is minimal.

This problem is known to be NP-hard [2], thus there are no efficient algorithms to exactly solve this task (and there is not much hope of finding one). This means that every method guaranteeing optimal solutions has impractically long computational time for even moderate problem sizes. Hence, only heuristics resulting in so called quasi-optimal solutions are viable. Over the past few decades, a number of such heuristics have been invented and published (e.g. [3]-[5]).

Since due to the nature of the PFSP problem these heuristics cannot be evaluated analytically, their evaluation and their comparison to other techniques can be made experimentally, i.e. based on results of simulation runs carried out on standard reference tasks, called benchmark problems. Several such comparisons have been made involving a large part of the so far proposed methods (e.g. [3]-[5]). These comparative studies are mostly based on the well-known Taillard's benchmark problem set [6].

This paper proposes approaches for adapting chromosome based evolutionary methods to the Permutation Flow Shop Problem. The proposal includes two types of individual representation (i.e. encoding method): a permutation and a real value based one. They are applied on three different chromosome based evolutionary techniques, namely the Genetic Algorithm [7], the Bacterial Evolutionary Algorithm [8] and the Particle Swarm Optimization [9] method. Both representations are applied on the two former methods, whereas the real value based one is used for the latter optimization technique. Each mentioned algorithm is involved without and with local search steps as one of its evolutionary operators. Since the evolutionary operators of each technique are established according to the applied representation, this paper deals with a total number of ten different chromosome-based evolutionary methods.

The obtained techniques are evaluated via simulation runs carried out on the above mentioned Taillard's benchmark problem set. Based on these experimental results, the methods, and thus the chromosome-based evolutionary technique adapting approaches themselves, are compared to each other.

The next section gives a formal definition to the PFSP problem. Within this, the search space and the makespan function as the objective function are defined.

Then, the third section gives a brief overview of the chromosome-based evolutionary techniques being adapted to the PFSP task. The basic concept and the main steps of the algorithms are also presented. The new encoding approaches for the PFSP problem are proposed in section four. After that, the evolutionary operators constructed based on the newly proposed individual representations are described. The sixth section enumerates the algorithms, which can be established by using the discussed approaches and which are compared via simulation runs. The experimental results and the observed characteristics are discussed in section seven. Finally, in the last section our work is summarized and some conclusions are drawn.

2 The Permutation Flow Shop Problem

As was described in the Introduction, in this problem there are given the number of jobs n, the number of machines m and an n-by-m processing time matrix **P** defining the necessary amount of time a job has to stay on a machine, for each job-machine pair. That is, the elements of the matrix are positive and an element $p_{i,j}$ denotes the time the i^{th} job stays on the j^{th} machine.

All the jobs should be worked by all the machines one after another. The machines are deployed in a line and a machine can handle one single job at once. That is, the process of the jobs is pipeline-like. A job can be processed on a machine only if the machine is free (the preceding job has finished on the machine) and the job has already been processed on the preceding machine.

The task is to find a permutation (an order) of the jobs, in case of which the total processing time of all the jobs on all the machines (i.e. the so called makespan) is minimal.

For example, if there are three jobs the permutation *(2,3,1)* denotes the case when the second job goes first, the third goes next, and finally the first goes last. This should not be confused with another interpretation of a permutation, where the same permutation would mean a case when the first job goes second, the second one goes third and the last one goes first. In our interpretation this latter will be referred as the 'inverse' of the permutation defined above. (It can be easily seen that this 'inverse' is the algebraic inverse of the permutation; consequently, the inverse of the inverse of the permutation is the permutation itself.)

Clearly, the search space is the set of the n-order permutations S_n, and the objective function is defined over this search space and its range is the set of positive numbers R^+.

Formally, the objective or makespan function f can be defined as follows (see e.g. [1]):

$$
\begin{aligned}
&f : S_n \to R^+ \\
&f(\sigma) = t(n, m, \sigma) \\
&t(0, j, \sigma) \equiv 0 \\
&t(i, 0, \sigma) \equiv 0 \\
&t(i, j, \sigma) = \max(t(i, j-1, \sigma), t(i-1, j, \sigma)) + p_{\sigma(i), j}
\end{aligned}
\tag{1}
$$

where $p_{\sigma(i),j}$ and $t(i,j,\sigma)$ denote the processing and the completion times of the i^{th} job of the σ permutation on the j^{th} machine, respectively.

The task is to find a σ permutation for which the makespan is optimal (i.e. minimal).

3 Overview of the Evolutionary Techniques Applied

A famous, frequently studied and applied family of iterative stochastic optimization techniques is called chromosome based evolutionary algorithms. These methods, like the Genetic Algorithm (GA) [7] or the Bacterial Evolutionary Algorithm (BEA) [8], imitate the abstract model of the evolution of populations observed in nature. Their aim is to change the individuals in the population (set of individuals) by the evolutionary operators to obtain better and better ones. The goodness of an individual can be measured by its 'fitness'. If an individual represents a candidate solution for a given problem, the algorithms try to find the optimal solution for the problem. Thus, in the case of optimization problems, the individuals represent elements of the search space and the fitness function is a transformation of the objective function. If an evolutionary algorithm uses an elitist strategy, it means that the best ever individual will always survive and appear in the next generation. As a result, at the end of the algorithm the best individual will represent the (quasi-) optimal element of the search space.

The individuals are usually represented by chromosomes (this is why these methods are called chromosome-based evolutionary algorithms), which are most often vectors holding numbers in their components (i.e. in their genes). The manner in which the individuals are represented as chromosomes is the encoding method.

The steps of the algorithms changing the chromosomes, and thus the candidate solutions, are called evolutionary operators. The evolutionary operators are in strong connections with the encoding technique, since the encoding determines the form of the chromosomes the operators work with.

It is quite obvious that besides the formation of the skeleton of the evolutionary algorithm, the design of the encoding method and the evolutionary operators also play key roles in the efficient government of the evolution process.

There are a huge number of chromosome based evolutionary algorithms. Some of them will be presented below; these are the ones that were investigated in our work.

3.1 Genetic Algorithm

One of the most (if not the most) widely applied chromosome based evolutionary techniques is the Genetic Algorithm (GA) [7]. It comprises the following steps:

1 Initialization:
An initial population is created by selecting random elements of the search space according to some distribution, or by using an initial heuristic.

2 Selection:
Individuals are selected according to their fitness values. The higher fitness value an individual has, the bigger its probability to be selected. There are a number of selection methods, e.g. roulette wheel technique, or stochastic universal sampling.
The selected individuals are called parents.

3 Crossover:
Pairs are formed from the set of parents and a random point of the chromosome is selected for each pair. Then the parents change the sequence of their genes between each other after the selected point. The resulting individuals are called offspring.

4 Mutation:
The genes of each offspring are mutated with a certain probability and take new random values.

5 Substitution:
The offspring are substituted in the population, i.e. they overwrite individuals in it. The individuals to overwrite are selected according to their fitness values. However, unlike in the selection step, in this case the higher fitness value an individual has, the smaller its probability to be selected. If the above mentioned elitist strategy is applied, the best individual will not be overwritten. With minor modifications, the same algorithms can be used for the selection of individuals to overwrite as in the selection step.

The main iteration loop of the algorithm contains steps 2 – 5. A single iteration is called generation. The algorithm stops if at the end of a generation one of the termination criteria fulfills (generation limit reached, time limit exceeded, etc.). After termination the best individual represents the quasi-optimal solution.

3.2 Bacterial Evolutionary Algorithm

Compared to GA, a somewhat different evolutionary technique is called the Bacterial Evolutionary Algorithm (BEA). This algorithm was introduced by Nawa and Furuhashi in [8]. The first version of this algorithm was called the Pseudo-Bacterial Genetic Algorithm (PBGA) [10], which proposed a modified mutation operator called bacterial mutation, based on the natural phenomenon of microbial evolution. The Bacterial Evolutionary Algorithm introduced a new operator called the gene transfer operator. While PBGA incorporates bacterial mutation and crossover operator, the BEA substitutes the classical crossover with the gene transfer operation. Both of these new operators were inspired by bacterial evolution. Bacteria can transfer genes to other bacteria, and thus gene transfer allows the bacteria to directly transfer information to the other individuals in the population.

BEA comprises the following steps:

1 Initialization:
   An initial population is created by selecting random elements of the search space according to some distribution, or by using an initial heuristic.
2 Bacterial mutation:
   All bacteria are mutated in all their genes multiple times in random orders. In case of each mutation step, if the original value was better, then it is restored; if the new one makes the individual have higher fitness value, then it is kept.
3 Gene transfer:
   The population is divided into two parts according to the fitness values. The individuals possessing higher fitness values form the superior and the ones having lower values form the inferior part of the population.
   Then pairs are formed, where the first members of the pairs are from the superior part (superior individuals) and the second members come from the inferior part (inferior individuals). For each pair a random point of the chromosome is selected. Then the value of the gene at the selected point in the inferior bacterium is overwritten with the value of the gene at the selected point in the superior bacterium.

The main iteration loop of the algorithm contains steps 2 and 3. The algorithm stops if at the end of a generation one of the termination criteria fulfils (generation limit reached, time limit exceeded, etc.). After termination the best individual represents the quasi-optimal solution.

3.3 Particle Swarm Optimization

Another type of iterative methods is called swarm intelligence techniques. These algorithms, like the nowadays famous Particle Swarm Optimization (PSO) [9], are inspired by social behavior observed in nature, e.g. bird flocking or fish schooling. In these methods a number of individuals try to find better and better places by exploring their environment, led by their own experiences and the experiences of the whole community. Since these methods are also based on processes of the nature, like GA or BEA, and since there is also a type of evolution in them ('social evolution'), they can be categorized amongst the evolutionary algorithms.

Similarly, as was mentioned above, these techniques can also be applied as optimization methods if the individuals represent candidate solutions.

PSO comprises the following steps:

1 Initialization:
An initial population is created by selecting random elements of the search space according to some distribution, or by using an initial heuristic.
In the case of each individual, their current place is considered as its local best place (see its reason below). Moreover, the place of the best individual is stored as the global best place in the search space.

2 Moving the individuals:
The individuals are moved (their position within the search space x_i is changed) based on their local best place l_i and on the global best place g_i (the subscript i denotes the number of the current iteration). The new position x_i is determined by the following equations:

$$v_0 = 0$$

$$v_{i+1} = \varphi_v v_i + \varphi_l r_l (l_i - x_i) + \varphi_g r_g (g_i - x_i) \quad (2)$$

$$x_{i+1} = x_i + v_{i+1}$$

where φ_v, φ_l and φ_g are parameters of the algorithm, r_l and r_g are random values.

3 Updating local and global best places:
The individuals are evaluated and if an individual is at a better position than its local best place, the local best place is set to the current position. If the currently best individual in the population has higher fitness value than the fitness value of the global best place, then the global best place is set to the position of the currently best individual.

The main iteration loop of the algorithm contains steps 2 and 3. The algorithm stops if at the end of a generation one of the termination criteria fulfils (generation limit reached, time limit exceeded, etc.). After termination the global best place represents the quasi-optimal solution.

3.4 Memetic Algorithms

The techniques causing minor modifications to the candidate solutions iteration by iteration and thus exploring only the 'neighborhood' of particular elements of the search space are called local search methods.

After a proper amount of iterations, as a result of these minor modification steps, the local search algorithms find the 'nearest' local minimum quite accurately. However, these techniques are very sensible to the location of the starting point. In order to find the global optimum, the starting point must be located close enough to it, in the sense that no local optima separate the two points.

Evolutionary computation techniques explore the whole objective function, because of their characteristic, so they find the global optimum, but they approach it slowly. Local search based algorithms, meanwhile, find only the nearest local optimum; however, they converge to it faster.

Avoiding the disadvantages of the two different technique types, evolutionary algorithms (including swarm intelligence techniques) and local search methods may be combined [11], for example, if in each iteration for each individual one or more local search steps are applied. Expectedly, this way the advantages of both local search and evolutionary techniques can be exploited: the local optima can be found quite accurately on the whole objective function, i.e. the global optimum can be obtained quite accurately.

There are several results in the literature confirming this expectation in the following aspect. Usually, the more difficult the applied local search step is, the higher convergence speed the algorithm has in terms of number of generations. It must be emphasized that most often these results discuss the convergence speed in terms of number of generations. However, the more difficult an algorithm is, the greater computational demand it has, i.e. each iteration takes longer.

Therefore the question arises: how does the speed of the convergence change in terms of time if the local search based technique applied in the method is changed?

Apparently, this is a very important question of applicability, because in real world applications time as a resource is a very important and expensive factor, but the number of generations the algorithm executes does not really matter.

This is the reason why the efficiency in terms of time was chosen to be investigated in this paper.

4 Proposed Encoding Methods

Two types of individual representation (i.e. two encoding methods) are proposed in this paper for the evolutionary techniques.

The first one is based on the permutations themselves, thus the evolutionary operators modify the elements of the permutations directly.

The second encoding method is an indirect, real value based encoding approach, which is an obvious extension of those representations applied for numerical optimization problems. Although the operators modify the values of real valued vectors (arrays) – since the objective function is defined over permutations, the chromosomes represent permutations actually – there is a need to convert the real valued vectors to permutations somehow.

In order to reduce time complexity costs, the chromosomes can be 'mirrored' within the individuals in a manner which makes the modifications caused by the evolutionary operators and the evaluation of the individual more simply performable.

4.1 Permutation-based Encoding

This representation is based on the permutations themselves. Each chromosome holds one single permutation, where the genes represent the jobs and each gene holds an element of the permutation. That is, the chromosome is an integer vector, where the values of the genes are between 1 and n (where n is the number of jobs), additionally, every integer appears once in the chromosome.

Actually, this permutation is not the permutation the objective function gets, i.e. it is not the one defined in Section 2, but its inverse (as was explained by a simple example before). Thus, hereafter this will be called the 'inverse permutation'.

4.2 Real Value-based Encoding

Most often in the case of numerical optimization problems, the individuals have binary or real representations. This means that the chromosomes are binary or real valued vectors (arrays) representing points in the search space, i.e. candidate solutions.

In those most frequent cases when the objective function is defined over R^n (or over a subset of R^n), it is a natural way to encode the individuals in real valued vectors.

This representation can be extended to PFSP tasks easily as follows.

If the number of jobs is n, then the chromosomes are real valued vectors with length of n. Since in the case of PFSP tasks the objective function is defined over permutations, the real valued vectors must be converted to permutations. This can be done by ordering the genes according to the values they have. Because there is exactly one permutation in S_n corresponding to every gene order, the new gene order is equivalent to a permutation.

There seems to be an unnecessary 'overhead' in the previous encoding technique, because one could say that the chromosome should hold the permutation and the operators should modify the permutations directly, instead of changing a real valued vector and the permutation via this vector.

However, despite the computational overhead, this encoding manner is more useful, as will be presented in the next sections.

4.3 Mirroring the Chromosomes

Performing the effects of the evolutionary operators on the individuals can be made computationally cheaper in the following way.

The individuals do not comprise only one single chromosome as usual, but two chromosomes being similar to each other: an original and a mirrored one. The original chromosome contains a vector of real numbers and the inverse permutation or only the inverse permutation (based on the base of the encoding) as discussed above. The mirrored chromosome contains the inverse of the inverse permutation (i.e. the permutation used by the objective function) and in the case of real value based encoding, the adequate permutated order of the real numbers (i.e. the real numbers in an ascending order).

The chromosome and the mirrored chromosome are updated simultaneously in every step during the application of the evolutionary operators. Thus, they are always equivalent in the sense that they always represent the same candidate solution for the problem.

The reasons why this mirroring technique causes the reduction of computational costs will be explained in the next section during the discussion of the certain operators.

5 Established Evolutionary Operators

The different evolutionary operators used by the algorithms are derived back to three 'atomic operators': mutation, gene transfer and local search.

Mutation in GA and bacterial mutation in BEA can be obviously constructed by using the atomic operator mutation, and gene transfer in BEA can be done by using the atomic operator gene transfer.

Crossover in GA can be considered as a sequence of gene transfers from a given position to a given position in a random order.

The atomic operator local search is exactly the same as the local search operator in all three evolutionary techniques.

It is easy to see that if all the atomic operators are defined so that their results are valid individuals (i.e. individuals representing permutations), the constructed operators also results in valid individuals.

The atomic operators are the following.

5.1 Mutation

5.1.1 Permutation-based Encoding

In the permutation-based case, the mutation of a gene means that the value of the gene is set to a random integer value from 1 to n (where n is the number of jobs). This modification would lead out from the search space, because the resulting integer vector would not be a permutation, hence a 'compensation step' is made, i.e. the gene whose value is taken by the mutated gene changes its value to the previous value of the mutated gene. That is, during mutation, the mutated gene changes its value with a random gene. In this way the mutation operator is closed with respect to the search space.

The change is committed both in the chromosome and in the mirrored chromosome.

Since the permutation-based mutation modifies only two values in the permutation, it makes 'local' changes within the chromosome.

5.1.2 Real Value-based Encoding

When a real value based gene is mutated, it is set to a random real value. Thus, the permutation represented by the chromosome changes.

It would be computationally expensive to compute the new corresponding inverse permutation by reordering the whole chromosome. Instead of this, the mirrored chromosome can be applied, where the place of the new value can be found easily by a computationally cheap binary search (since the real values are ordered in the mirrored chromosome). Then, in the mirrored chromosome, the sub-chromosome (i.e. the gene-sequence) between the original and the new place of the mutated gene is shifted one place left (if the new value is higher than the old one) or one place right (if the new value is lower than the old one).

During the shift, the corresponding elements of the inverse permutation are also updated in the chromosome.

Actually, the real value-based mutation means a random direction shift of a random length part of the mirrored chromosome. Thus, it causes not only local effects unlike the permutation based mutation.

After the previous description of the real value-based mutation, one could ask whether an equivalent operator could not be constructed based on only the permutations; i.e. is it possible that there is no difference between the strength of the two different encoding manners?

The answer is no, an equivalent operator could not be constructed based on only the permutations, because although the shift of random length gene-sequences could easily be made, in the case of real value based representation, the distribution of random variables determining the lengths and positions are also developing (implicitly) while the real values are changing. Therefore, the diversity of the real value based encoding is higher.

It was mentioned in the previous section that this representation has a computational overhead, but as it discussed now, it may give higher diversity to the mutation. Thus, certainly there is a difference between the strength of the two different encoding manners; however, it is an open question which one is more efficient, and by how much.

This will also be investigated in Section 6.

5.2 Gene Transfer

5.2.1 Permutation-based Encoding

During gene transfer in the case of permutation based encoding the inverse permutation value of the selected gene in the target individual is set to the inverse permutation value of the corresponding gene of the source individual. Hereafter, a compensation step is made similarly as in the case of mutation.

5.2.2 Real Value-based Encoding

Applying real value-based encoding, the gene transfer is not much different. The real value of the selected gene in the target individual is set to the real value of the corresponding gene of the source individual. Hereafter, a similar shifting in the mirrored chromosome followed by the update of the chromosome is made as in the case of mutation.

5.3 Local Search

The local search is performed the same way in the case of both representations. One iteration cycle of the local search is as follows.

First of all, a random order of the elements of the permutation from the first to the last but one is selected. Then, following this order, the neighboring elements within the mirrored chromosome try to change their values with each other so that if according to the random order the current element is the i^{th}, then it tries to change its value with the $(i+1)^{th}$. After each change between the neighbors, if the resulting permutation is better (i.e. it has a higher fitness value), the change is kept and both the chromosome and the mirrored chromosome are updated according to the modification made. Otherwise, the change is rolled back.

6 Optimization Algorithms Investigated in this Paper

Based on the previous sections ten different evolutionary optimization techniques were constructed and evaluated. These are the following:

- Genetic Algorithm based techniques:
  - GAr: Genetic Algorithm without local search using real value based encoding
  - GAp: Genetic Algorithm without local search using permutation based encoding
  - GMAr: Genetic Algorithm with local search (Memetic Algorithm) using real value based encoding
  - GMAr: Genetic Algorithm with local search (Memetic Algorithm) using permutation based encoding
- Bacterial Evolutionary Algorithm based techniques:
  - BEAr: Bacterial Evolutionary Algorithm without local search using real value based encoding
  - BEAp: Bacterial Evolutionary Algorithm without local search using permutation based encoding
  - BMAr: Bacterial Evolutionary Algorithm with local search (Bacterial Memetic Algorithm) using real value based encoding
  - BMAp: Bacterial Evolutionary Algorithm with local search (Bacterial Memetic Algorithm) using permutation based encoding
- Particle Swarm Optimization based techniques:
  - PSO: Particle Swarm Optimization without local search using real value based encoding

- PMO: Particle Swarm Optimization with local search using real value based encoding

In the remaining part of the paper GAr, GAp, GMAr and GMAp will be referred to as 'genetic' techniques, BEAr, BEAp, BMAr and BMAp will be labeled as 'bacterial' methods, and finally PSO and PMO will be referred to as 'particle swarm' algorithms.

7 Evaluation of the Obtained Techniques

Simulation runs were carried out in order to evaluate and to compare the efficiency of the proposed approaches and the established algorithms. First, the new methods are compared to each other, then the best one is compared to two other heuristics: the well-known Iterated Greedy technique (IG) [12] and a genetic algorithm based memetic method (MA) [13], which is e.g. used in combination with IG in multi-processor systems.

For these purposes, a dozen problems were applied from the well-known Taillard's benchmark set. Exactly one problem from each available problem sizes.

In the simulations, the parameters of the newly proposed methods had the following values, because after a number of test runs these values seemed to be the most suitable.

The number of individuals in a generation was 14 in genetic and 8 in bacterial algorithms, whereas it was 80 in particle swarm methods. In the case of genetic techniques the selection rate was 0.5 and the mutation rate was 0.3; in the case of bacterial techniques, the number of clones was 2 and 1 gene transfer was carried out in each generation. The genetic methods applied the elitist strategy.

In the iterated greedy methods, 4 jobs were selected to remove in each generation and the temperature parameter was 5 (see [12]). The MA technique used 13 individuals as in [13].

The simulation was carried out on a PC with E8500 3.16 GHz Intel Core 2 Duo CPU and using Windows Vista Business 64-bit operating system.

In the case of all ten new algorithms for all benchmark problems 5 runs were carried out. Then the means of the obtained values were taken.

The means of the objective function values of the best individuals during the runs of the new techniques are presented in figures (Figures 1-12) to get a better overview. The horizontal axes show the elapsed computation time in seconds and the vertical axes show the makespan values of the best individuals at the current time.

In the figures, dotted lines show the results of the pure evolutionary algorithms applying permutation based encoding (GAp and BEAp); dashed-dotted lines denote the memetic techniques using permutation-based encoding (GMAp and BMAp); solid lines present the graphs of the pure evolutionary methods applying real value-based encoding (GAr, BEAr and PSO); and dashed lines show the memetic techniques using real value based encoding (GMAr, BMAr and PMO). In each case a dashed horizontal line shows the best known makespan value according to [6].

The means of the resulting values were collected in tables (Tables I-VI). In the tables under the 'Problem' label the 'ID' columns show the identifier of the tasks in Taillard's benchmark problem set [6] and 'Size' denotes the size of the benchmark problem (in the form of "number of jobs times number of machines"). The best known makespan values according to [6] are collected in columns labeled by 'Best known value'. 'Time limit' shows the length of the simulation runs in seconds. The time limits were chosen according to test runs to values, after which the techniques did not show significant improvements (cf. Figures 1-12). Under the algorithm labels the 'Results' columns present the mean of the makespan values produced by the techniques, 'Rel. diff.' shows the mean of the relative difference of these makespan values compared to the known best ones

$$\frac{1}{5}\sum_{i=1}^{5}(\text{Result}_i - \text{Best known value})/\text{Best known value}, \tag{3}$$

whereas 'No. of gen.' denotes the mean of the number of executed generations. The best makespan values for each benchmark problem are bold underlined numbers and the best values of a particular evolutionary algorithm family (genetic, bacterial and particle swarm) for each benchmark problem not being the totally best values are italic underlined numbers.

The results of the runs of the new algorithms and their short explanations follow in the next subsection. After that, the results of the comparison with IG and MA are analyzed. In Subsection 7.3 conclusions will be drawn about the main characteristics of the behavior of the methods.

7.1 Experimental Results for the Established Techniques

The following observations could be made based on the obtained values (see Figures 1-12 and Tables I-V).

Considering the figures and tables probably the most obvious tendency of the results is that bacterial techniques gave better performance in each case than genetic and particle swarm based ones, as they were never outperformed by other methods. Such an unambiguous observation cannot be made between the latter two algorithm families. The superiority of the bacterial algorithms is growing in

terms of the difficulty (i.e. the size) of the optimization task. In the case of easier problems, the difference in efficiency between the algorithm families is not so significant (see Figures 1, 2, 4 and 7); however as the problem size increases, the significance grows (see Figures 3, 5, 6, 8, 9 and 10). Finally, in the case of the most difficult tasks (i.e. the biggest problem sizes) the difference is more than significant.

By looking at Table IV it is clear to see that BMAr performed best during the simulation runs, because in half of the cases it produced better results than the other algorithms, whereas in three more cases it found the known best values for the benchmark problems. This means that BMAr was outperformed by other techniques only in a quarter of the problems.

As can also be observed, memetic algorithms (the methods integrating local search steps) had higher efficiency in most of the cases. Among the genetic techniques, GMAr was the best in 8 problems, whereas GAr was the best in its family only 4 times out of 12 (see Tables I-II). In the case of the bacterial methods, the pure evolutionary techniques gave better results only in two cases, whereas the memetic ones had higher performance in seven cases. PMO was outperformed by PSO only once (see Table V).

One more very important feature is characterized by the results. Except for two cases (‘ta011’ in Table II and ‘ta071’ in Table IV) out of 48, the real value based methods were never worse than the corresponding permutation based ones. Moreover, even in those exceptional cases, the differences are insignificantly tiny.

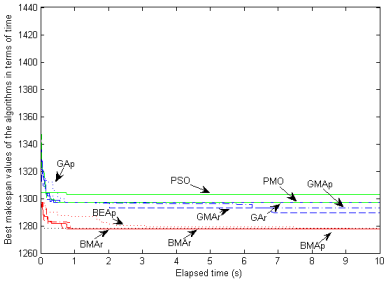


Figure 1
Results for the 20x5 size problem (ta001)

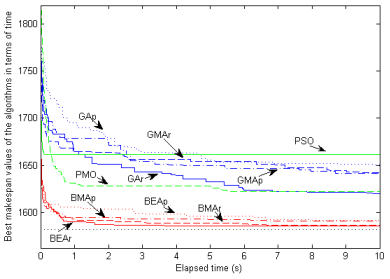


Figure 2
Results for the 20x10 size problem (ta011)

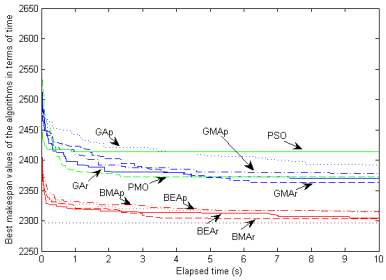


Figure 3
Results for the 20x20 size problem (ta021)

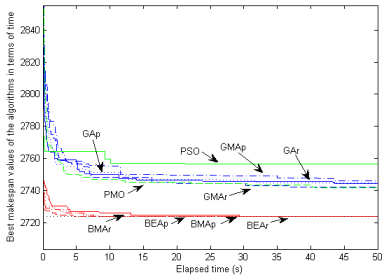


Figure 4
Results for the 50x5 size problem (ta031)

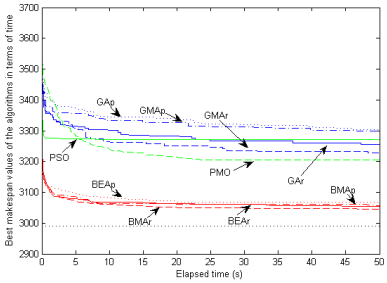


Figure 5
Results for the 50x10 size problem (ta041)

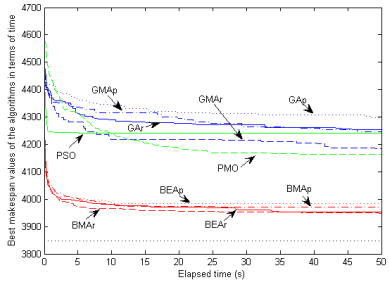


Figure 6
Results for the 50x20 size problem (ta051)

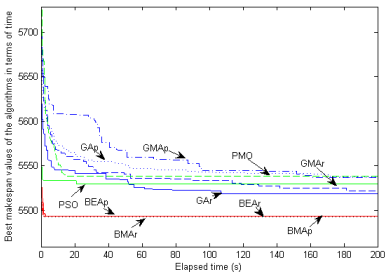


Figure 7
Results for the 100x5 size problem (ta061)

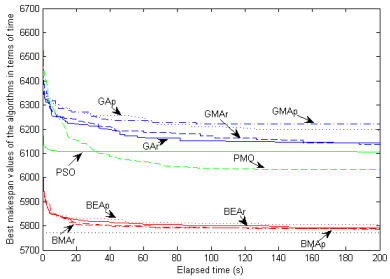


Figure 8
Results for the 100x10 size problem (ta071)

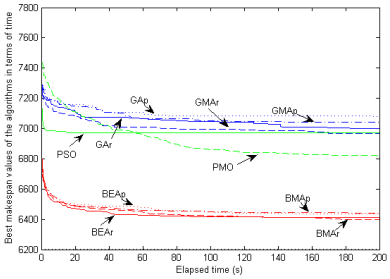


Figure 9
Results for the 100x20 size problem (ta081)

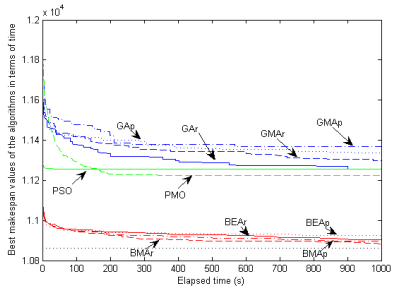


Figure 10
Results for the 200x10 size problem (ta091)

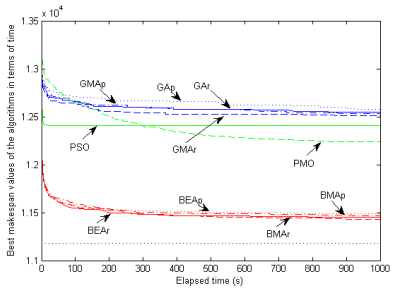


Figure 11
Results for the 200x20 size problem (ta101)

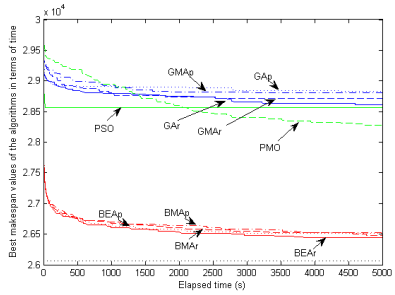


Figure 12
Results for the 500x20 size problem (ta111)

Table I
Results of the pure evolutionary genetic methods

Problem				GAp			GAr		
ID	Size	Best known value	Time limit	Result	Rel. diff.	No. of gen.	Result	Rel. diff.	Num. of gen.
ta001	20x5	1278	10	1297	1.49%	115751.8	1297	1.49%	99583.6
ta011	20x10	1582	10	1650.4	4.32%	85947.4	*1619.8*	2.39%	76262
ta021	20x20	2297	10	2392.6	4.16%	53249.6	2370	3.18%	49463.4
ta031	50x5	2724	50	2745	0.77%	229461.6	2744.4	0.75%	174902.8
ta041	50x10	2991	50	3304.6	10.48%	172645.8	3255.8	8.85%	137602.4
ta051	50x20	3847	50	4298.6	11.74%	111569.2	4254.2	10.58%	96096
ta061	100x5	5493	200	5537.8	0.82%	486640	*5519*	0.47%	301985.4
ta071	100x10	5770	200	6197	7.40%	348813.6	6142.2	6.45%	241455.6
ta081	100x20	6202	200	7077.8	14.12%	224934	7000.4	12.87%	174915.6
ta091	200x10	10862	1000	11336.2	4.37%	858038.4	*11254.4*	3.61%	482466.8
ta101	200x20	11181	1000	12577.6	12.49%	527048.8	12551.2	12.25%	346630.6
ta111	500x20	26059	5000	28827.6	10.62%	1033591.6	*28614.6*	9.81%	490406.2

Table II

Results of the genetic algorithm based memetic techniques

Problem				GMAp			GMAr		
ID	Size	Best known value	Time limit	Result	Rel. diff.	No. of gen.	Result	Rel. diff.	No. of gen.
ta001	20x5	1278	10	1293.2	1.19%	13525.8	*1289.4*	0.89%	13220.6
ta011	20x10	1582	10	1641	3.73%	9520.8	1641.6	3.77%	9422.4
ta021	20x20	2297	10	2378	3.53%	5648.6	*2363.2*	2.88%	5598.2
ta031	50x5	2724	50	2746.2	0.81%	11376.4	*2742*	0.66%	11203.8
ta041	50x10	2991	50	3297.2	10.24%	7934.6	*3228.6*	7.94%	7930.8
ta051	50x20	3847	50	4246.6	10.39%	4911	*4183.6*	8.75%	4831.6
ta061	100x5	5493	200	5537.2	0.80%	11758.4	5522	0.53%	11543.4
ta071	100x10	5770	200	6220.6	7.81%	8079.4	*6136.2*	6.35%	8048
ta081	100x20	6202	200	7040.8	13.52%	5057.4	*6967.8*	12.35%	5014.2
ta091	200x10	10862	1000	11369.6	4.67%	10104.4	11298.4	4.02%	10109.8
ta101	200x20	11181	1000	12532.4	12.09%	5787.8	*12514.4*	11.93%	5857.2
ta111	500x20	26059	5000	28793.6	10.49%	4725.8	28715	10.19%	4642

Table III

Results of the pure evolutionary bacterial methods

Problem				BEAp			BEAr		
ID	Size	Best known value	Time limit	Result	Rel. diff.	No. of gen.	Result	Rel. diff.	No. of gen.
ta001	20x5	1278	10	***1278***	0.00%	5943.4	***1278***	0.00%	5495
ta011	20x10	1582	10	1591.6	0.61%	4320.4	***1585.2***	0.20%	4048.8
ta021	20x20	2297	10	2316	0.83%	2549	2305	0.35%	2452.8
ta031	50x5	2724	50	***2724***	0.00%	5050	***2724***	0.00%	4720
ta041	50x10	2991	50	3067	2.54%	3684	3055.2	2.15%	3508.4
ta051	50x20	3847	50	3983.2	3.54%	2190	3951.6	2.72%	2140
ta061	100x5	5493	200	5493.4	0.01%	4998	***5493***	0.00%	4759
ta071	100x10	5770	200	5804.8	0.60%	3695.6	5791	0.36%	3597.6
ta081	100x20	6202	200	6436	3.77%	2205	6411	3.37%	2199
ta091	200x10	10862	1000	10923.4	0.57%	4497.8	10905.4	0.40%	4274.4
ta101	200x20	11181	1000	11491.4	2.78%	2698	11448.8	2.40%	2597
ta111	500x20	26059	5000	26516.4	1.76%	2159.8	***26440***	1.46%	2097.4

Table IV

Results of the bacterial evolutionary algorithm based memetic techniques

Problem				BMAp			BMAr		
ID	Size	Best known value	Time limit	Result	Rel. diff.	No. of gen.	Result	Rel. diff.	No. of gen.
ta001	20x5	1278	10	***1278***	0.00%	4030.4	***1278***	0.00%	3818
ta011	20x10	1582	10	1591	0.57%	2944.4	1586.4	0.28%	2770.2
ta021	20x20	2297	10	2314.4	0.76%	1728.8	***2303.8***	0.30%	1685.8
ta031	50x5	2724	50	***2724***	0.00%	3340	***2724***	0.00%	3200
ta041	50x10	2991	50	3055.6	2.16%	2455	***3045.6***	1.83%	2365.8
ta051	50x20	3847	50	3970.2	3.20%	1460	***3945.6***	2.56%	1440
ta061	100x5	5493	200	***5493***	0.00%	3332.6	***5493***	0.00%	3194
ta071	100x10	5770	200	***5787.6***	0.31%	2427.4	5788.2	0.32%	2379.6
ta081	100x20	6202	200	6438.6	3.81%	1482.6	***6392.4***	3.07%	1464
ta091	200x10	10862	1000	10897.6	0.33%	2994.6	***10882.2***	0.19%	2884.6
ta101	200x20	11181	1000	11466.2	2.55%	1774.8	***11432.4***	2.25%	1733.6
ta111	500x20	26059	5000	26516.4	1.76%	1393.8	26476.4	1.60%	1377.8

Table V
Results of the particle swarm methods

Problem				PSO			PMO		
ID	Size	Best known value	Time limit	Result	Rel. diff.	No. of gen.	Result	Rel. diff.	No. of gen.
ta001	20x5	1278	10	1302.8	1.94%	12817.4	*1297*	1.49%	1250.6
ta011	20x10	1582	10	1661.6	5.03%	10142.4	*1622.2*	2.54%	904.4
ta021	20x20	2297	10	2413.4	5.07%	7750.6	*2372.8*	3.30%	563.6
ta031	50x5	2724	50	2756.6	1.20%	22930	*2741.6*	0.65%	1070
ta041	50x10	2991	50	3271.2	9.37%	18066	*3205.4*	7.17%	766
ta051	50x20	3847	50	4240.8	10.24%	13640	*4164.6*	8.26%	470
ta061	100x5	5493	200	*5530.4*	0.68%	35867.4	5538.6	0.83%	1053.6
ta071	100x10	5770	200	6104.8	5.80%	31259	*6031*	4.52%	776.2
ta081	100x20	6202	200	6972	12.42%	24472.4	*6819*	9.95%	478
ta091	200x10	10862	1000	11255.6	3.62%	57462.2	*11222.4*	3.32%	959
ta101	200x20	11181	1000	12409.6	10.99%	47035	*12240*	9.47%	569.8
ta111	500x20	26059	5000	28578	9.67%	58591.6	*28274.2*	8.50%	436.4

Now, the question that arose in Section 5 is answered: it is worth applying real value based representation, because despite the computational overhead, the methods based on it are more efficient than the ones using permutation based encoding.

The observed behavior of the different algorithms matches with the results of our previous works comparing evolutionary algorithms on general optimization benchmark problems, and particularly on fuzzy rule based supervised machine learning problems (cf. e.g. [14], [15]).

7.2 Comparison to other Methods

Since BMAr appeared to be the most efficient algorithm, this technique is involved in further investigations: this method is compared to the Iterated Greedy heuristic and to the genetic algorithm based memetic method.

Table VI shows the results of the comparison of BMAr, IG and MA, where the heightened results are the best makespan values.

Table VI
Comparison of BMAr, IG and MA

Problem				BMAr		IG		MA	
ID	Size	Best known value	Time limit	Result	Rel. diff.	Result	Rel. diff.	Result	Rel. diff.
ta001	20x5	1278	10	***1278***	0,00%	***1278***	0,00%	***1278***	0,00%
ta011	20x10	1582	10	1586,4	0,28%	***1583,2***	0,08%	1585,4	0,21%
ta021	20x20	2297	10	2303,8	0,30%	***2301,6***	0,20%	2304,4	0,32%
ta031	50x5	2724	50	***2724***	0,00%	***2724***	0,00%	***2724***	0,00%
ta041	50x10	2991	50	3045,6	1,83%	***3035,2***	1,48%	3062,2	2,38%
ta051	50x20	3847	50	3945,6	2,56%	***3925***	2,03%	3958,4	2,90%
ta061	100x5	5493	200	***5493***	0,00%	***5493***	0,00%	***5493***	0,00%
ta071	100x10	5770	200	5788,2	0,32%	***5786,8***	0,29%	5797,8	0,48%
ta081	100x20	6202	200	6392,4	3,07%	***6350***	2,39%	6387,8	3,00%
ta091	200x10	10862	1000	***10882,2***	0,19%	10888,6	0,24%	10885,2	0,21%
ta101	200x20	11181	1000	11432,4	2,25%	***11392,4***	1,89%	11434,6	2,27%

As can be observed, BMAr was more efficient than MA, because 6 times out of 11 BMAr gave lower makespan values, whereas MA was better only 2 times. However, the most obvious fact appearing in the table is that IG significantly outperformed both other methods.

This result leads to two consequences. First, even the best established chromosome based technique cannot be a rival for one of the state-of-the-art methods, the Iterated Greedy heuristic. Second, although it cannot be a rival, it can be a "helpmate" of IG. In further research it would be worth constructing and evaluating hybrid methods based on BMAr and IG. A reason for this is that in the case of multi-processor systems, the combination of MA and IG resulted in a better technique than approaches applying only parallel IG threads [13].

However, such further investigations are beyond the scope of this paper.

7.3 Summary of the Main Observations

In short, the main observations made can be summarized as follows:

- Generally, bacterial techniques clearly outperformed the genetic and particle swarm ones.
- Usually, memetic methods (i.e. algorithms comprising local search steps as additional evolutionary operators) showed better performance than pure evolutionary approaches.
- Except in 2 cases out of 48, the methods applying real value based encoding technique were better than the ones using permutation based individual representation.
- BMAr seemed to be the overall best chromosome based evolutionary optimization heuristic for the PFSP problems.
- Although, the best constructed method was more efficient than a genetic algorithm based memetic technique applied in multi-processor systems, it was outperformed by one of the state-of-the-art heuristics, the Iterated Greedy method.

Conclusions

Our work proposed approaches for adapting chromosome based evolutionary methods to the Permutation Flow Shop Problem. The proposal included two types of individual representation (i.e. encoding method): a permutation and a real value based one. They were applied on three different chromosome based evolutionary techniques, namely the Genetic Algorithm, the Bacterial Evolutionary Algorithm and the Particle Swarm Optimization method. Both representations were applied on the two former methods, whereas the real value-based one was used for the latter optimization technique. Each mentioned algorithm was involved without and with local search steps as one of its evolutionary operators. Since the

evolutionary operators of each technique were established according to the applied representation, this paper investigated a total number of ten different chromosome based evolutionary methods.

The obtained techniques were evaluated via simulation runs carried out on the well-know Taillard's benchmark problem set. Based on the experiments the following observations could be made.

The real value based representation seemed to be better than the permutation based encoding technique. The algorithms applying local search performed better than the corresponding pure evolutionary methods, whereas bacterial techniques outperformed both genetic and particle swarm algorithms overwhelmingly.

Therefore, BMAr appeared to be the best established chromosome based evolutionary optimization method for the PFSP problem.

Although, the best constructed method was more efficient than a genetic algorithm based memetic technique applied in multi-processor systems, it was outperformed by one of the state-of-the-art heuristics, the Iterated Greedy method.

Ongoing research aims to combine the BMAr technique with IG and to establish new hybrid methods more efficient than either of them. That work considers single- as well as multi-threaded algorithms.

Since among chromosome based evolutionary algorithms bacterial methods performed best, in further research, slightly modified bacterial techniques, such as the Bacterial Memetic Algorithm with Modified Operator Execution Order [16], might also be involved.

Future work may also aim to compare the investigated techniques with other state-of-the-art methods published for the PFSP task and to combine the best one among them with the chromosome based evolutionary techniques, thus establishing a promising hybrid algorithm.

Finally, an additional research direction could be the extension of the proposed approaches to other scheduling tasks, such as scheduling problems considering setup times or involving concurrent processing of batches of jobs (see e.g. [17]).

Acknowledgement

The research is supported by the National Development Agency and the European Union within the frame of the project TÁMOP-4.2.2-08/1-2008-0021 at the Széchenyi István University entitled "Simulation and Optimization – basic research in numerical mathematics".

References

[1] S. M. Johnson: Optimal Two- and Three-Stage Production Schedules with Setup Times Included, Naval Research Logistics Quarterly 1, 1954, pp. 61-68

[2] A. H. G. Rinnooy Kan: Machine Scheduling Problems: Classification, Complexity and Computations, Martinus Nijhoff, The Hague, The Netherlands, 1976

[3] E. Taillard: Some Efficient Heuristic Methods for the Flow Shop Sequencing Problem, European Journal of Operational Research, Amsterdam, 1990, pp. 65-74

[4] A. Juan, A. Guix, R. Ruiz, P. Fonseca, F. Adelantado: Using Simulation to Provide Alternative Solutions to the Flowshop Sequencing Problem, 14th ASIM Dedicated Conf. on Simulation in Production and Logistic, Karlsruhe, Germany, 2010, pp. 349-356

[5] Z. Horváth, P. Pusztai, T. Hajba, Ch. Kiss-Tóth: Mathematical Methods and Parallel Codes for Production Line Optimization, Factory Automation 2011 Conference, Győr, Hungary, 2011

[6] E. Taillard, Benchmarks for Basic Scheduling Problems, European Journal of Operational Research 64 (2), 1993, pp. 278-285

[7] J. H. Holland: Adaption in Natural and Artificial Systems, The MIT Press, Cambridge, Massachusetts, 1992

[8] N. E. Nawa and T. Furuhashi: Fuzzy System Parameters Discovery by Bacterial Evolutionary Algorithm, IEEE Transactions on Fuzzy Systems, 7(5), 1999, pp. 608-616

[9] J. Kennedy and R. Eberhart: Particle Swarm Optimization, Proceedings of the IEEE International Conference on Neural Networks (ICNN '95), 4, Perth, WA, Australia, 1995, pp. 1942-1948

[10] N. E. Nawa, T. Hashiyama, T. Furuhashi, and Y. Uchikawa: Fuzzy Logic Controllers Generated by Pseudo-Bacterial Genetic Algorithm, Proceedings of the IEEE International Conference on Neural Networks, ICNN'97, Houston, 1997, pp. 2408-2413

[11] P. Moscato: On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms, Technical Report Caltech Concurrent Computation Program, Report. 826, California Institute of Technology, Pasadena, USA, 1989

[12] R. Ruiz, T. Stützle: A Simple and Effective Iterated Greedy Algorithm for the Permutation Flowshop Scheduling Problem, European Journal of Operational Research, 177, 2007, pp. 2033-2049

[13] M. G. Ravetti, C. Riveros, A. Mendes, M. G. C. Resende, and P. M. Pardalos: Parallel Hybrid Heuristics for The Permutation Flow Shop Problem, AT&T Labs Research Technical Report, 2010, p. 14

[14] K. Balázs, J. Botzheim, L. T. Kóczy: Comparison of Various Evolutionary and Memetic Algorithms, Proceedings of the International Symposium on

Integrated Uncertainty Management and Applications, IUM 2010, Ishikawa, Japan, 2010, pp. 431-442

[15] K. Balázs, J. Botzheim, L. T. Kóczy: Comparative Analysis of Interpolative and Non-interpolative Fuzzy Rule-based Machine Learning Systems Applying Various Numerical Optimization Methods, World Congress on Computational Intelligence, WCCI 2010, Barcelona, Spain, 2010

[16] R. Lovassy, L. T. Kóczy, L. Gál: Function Approximation Performance of Fuzzy Neural Networks, Acta Polytechnica Hungarica, Vol. 7, No. 4, 2010, pp. 25-38

[17] E. Kodeekha: Case Studies for Improving FMS Scheduling by Lot Streaming in Flow-Shop Systems, Acta Polytechnica Hungarica, Vol. 5, No. 4, 2008, pp. 125-143

Computation of Thermal and Hydraulic Performances of Minichannel Heat Sink with an Impinging Air Jet for Computer Cooling

M'hamed Beriache

Université des sciences et technologies Lille1, Polytech'Lille, USTL, LML, UMR 8107, 59655 Villeneuve d'Ascq, France
Université Hassiba Benbouali de Chlef, BP 151, Hay Essalem, Chlef (02000), Algérie; E-mail: m.beriache@gmail.com

Hassan Naji

Laboratoire Génie Civil et géo-Environnement (LGCgE- EA 4525), F-62400 Béthune, Université d'Artois, France Université Lille Nord de France, F-59000 Lille, France; E-mail: hassane.naji@univ-artois.fr

Ahmed Bettahar, Leila Mokhtar Saïdia

Département de mécanique, Université Hassiba Benbouali, BP 151, Hay Essalem, Chlef (02000), Algérie; E-mail: bettahara@yahoo.fr, ms_layla80@yahoo.fr

Abstract: This research work presents a numerical simulation of the CPU (Central Processor Unit) heat sink with a parallel plate fin and impingement air cooling. The governing equations are discretized by using the finite difference technique. The objective of this article is to investigate the thermal and hydraulic performances of a heat sink with impinging air flow to evaluate the possibility of improving heat sink performances. The numerical simulations are done by a personal C++ developed code. The thermal and hydraulic characterization of a heat sink under air-forced convection cooling condition is studied. The hydraulic and thermal parameters, including velocity profiles, the distribution of static pressure, pressure drop and temperature distributions through the fins, the base heat sink and the heat sink body through the heat sink are analyzed and presented schematically. The results show that the heat transferred by the heat sink increases with impinging Reynolds number. The performance of the proposed model computed by the numerical calculation is high compared with literature results.

Keywords: electronics cooling; heat sink; air flow; impingement jet; thermal resistance; pressure drop

1 Introduction

As electronic equipment becomes smaller and more advanced, it necessitates higher circuit integration per unit area, which in turn contributes to a rapid increase in heat generation. Thus, the effective removal of heat dissipations and maintaining the chip at a safe operating temperature have played important roles in insuring a reliable operation of electronic components [1]. There are many methods for electronics cooling, such as jet impingement cooling and heat pipe [2]. Conventional electronics cooling has normally used an impinging jet with a heat sink, showing superiority in terms of unit price, weight and reliability. Therefore, the most common way to enhance air-cooling is through the utilization of an impingement air jet on a mini or micro channel heat sink. In order to design an effective heat sink, some criteria such as thermal resistance, a low pressure drop and a simple structure should be considered.

When the literature is surveyed, a number of scholars have examined the air jet impingement on the heat sink geometry. Hilbert *et al*. [3] reported a novel laminar flow heat sink with two sets of triangular or trapezoidal shaped fins on the two inclined faces or a base. This design is efficient because the downward flow increases the air speed near the base of the fins, where the fin temperatures are highest. By having the cool air enter at the center of the heat sink and exit at the sides, the length of the fins in the flow direction is reduced so that the heat transfer coefficient is increased. Jang and Kim [4] conducted an experimental study of a plate fin heat sink subject to an impinging air jet. The geometry of a heat sink in impingement flow is similar to that shown in Figure 1. In this flow arrangement, the air enters at the top and exits at the sides. Based on experimental results, a correlation for the pressure drop and a correlation for its thermal resistance are suggested. They show that the cooling performance of an optimized microchannel heat sink subject to an impinging jet is enhanced by about 21% compared to that of the optimized microchannel heat sink with a parallel flow. Saini and Webb [5] developed a numerical model for predicting the pressure drop and thermal performance for impingement air flow in plate fin heat sinks. The model was than validated by experiments. The predicted pressure drop is 31% lower than the experimental data. El-Sheikh and Garimella [6] experimentally investigated the heat transfer enhancement of air jet impingement by using pin-fin heat sinks. In their study, the heat transfer coefficient, for both pinned and unpinned heat sinks, is only modestly dependent on the nozzle-to-target plate spacing (H/d). They also found that the heat transfer coefficient increases as the nozzle diameter decreases at a fixed flow rate. Biber [7] carried out a numerical study to determine the pressure drop of single isothermal channel with variable width impingement flow. Many different combinations of channel parameters are studied, and a correlation for the average static pressure loss coefficient across the slot jet and for average Nusselt number from an isothermal channel was presented. Duan and Muzychka [8] developed an impingement flow thermal resistance model. The simple model

is suitable for heat sink parametric design studies. The analytical model is developed for the low Reynolds number laminar flow and heat transfer in the inter fin channels of impingement flow plate fin heat sinks. The accuracy of the predicted thermal resistance was found to be within 20% of the experimental data at channel Reynolds numbers less than 1200. Duan and Muzychka [9] proposed a simple impingement flow pressure drop model based on developing laminar flow in a rectangular-channels heat sink. The validity of the model is verified by an experimental test. Measurements of the pressure drop were performed with heat sinks of various impingement inlet widths, fin spacing, fin heights and airflow velocities. It was found that the predictions agreed with experimental data within 20% at a channel Reynolds number less than 1200.

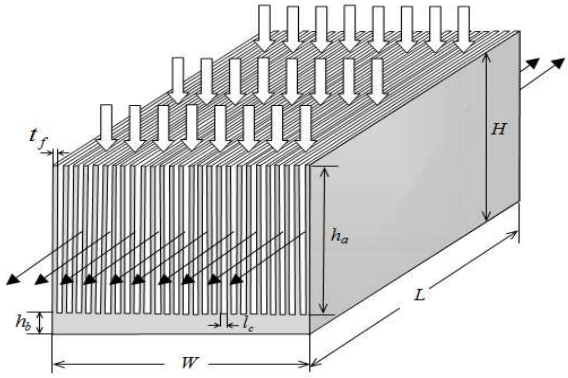


Figure 1
Schematic diagram showing the physical model

Duan and Muzychka [10] performed an experimental investigation of the thermal performance with four heat sinks of various impingement inlet widths, fin spacing, fin heights and airflow velocities. They developed a heat transfer model to predict the thermal performance of impingement air cooled plate fin heat sinks for design purposes. Shah *et al.* [11] demonstrated the results of a numerical analysis of the performance of an impingement heat sink designed for use with a specific blower as a single unit. The effects of the shape of the heat sink fins, particularly near the center of the heat sink, were examined. It was found that the removal of fin material from the central region of the heat sink enhances the thermal as well as the hydraulic performance of the sink. Shah *et al.* [12] extended the previous work by investigating the effect of the removal of fin material from the end fins, the total number of fins, and the reduction in the size of the hub fan. The reduction in the size of the hub of the fan is found to be a more uniform distribution of the air inside the heat sink, particularly near the center of the module. In general, there are many testing processes for heat sinks which must be introduced in an effort to obtain the thermal and hydraulic performance of heat sinks. If we take advantage of the numerical simulation to obtain some probable optimal design parameters before running experiments, the cost and research time can be reduced.

In the works cited above, some analytical and empirical models were developed and verified experimentally; others are based on CFD modeling, and they were mainly based on analyses of individual channels subjected to different boundary conditions. The literature survey has not revealed any articles that deal in detail with the flow field and heat transfer in heat sinks. In order to better reflect the flow field and temperature distribution in the channels and the body of the heat sink respectively, the present work complements the detailed flow field, pressure drop predictions and temperature distribution in the plate-fin heat sink with impingement air flow. An important focus of this study is to examine the capability of Navier-stokes equations treated by our developed code based on finite difference method (FDM) to effectively represent the fluid flow and heat transfer behavior in mini-channel heat sinks. In this paper, the numerical simulation of plate-fin heat sinks with confined impingement cooling in thermal-fluid characteristics will be investigated. The objective of this study is focused on the impingement flow plate-fin geometry. The research objectives are to develop a simple model for predicting thermal and hydraulic performances of a plate-fin heat sink for impingement air cooling.

2 Physical and Mathematical Model

2.1 Physical Model

The physical model of this study is illustrated in Fig. 1. In the configuration shown above, a heat sink with rectangular mini-channels with hydraulic diameter $D_h=0.0029\ m$ is heated from the bottom with the power (Q) generated by the CPU, which is absorbed by the sink and released to the atmosphere through the fins topped by an axial fan that blows air in impingement flow to dissipate heat in the atmosphere with a constant flow rate $\dot{V}$ (m^3/s). The sink is made of aluminium of $\lambda = 237 W/m.K$. The power dissipation from the P4 CPU is set to 80W [13]. The following assumptions are made in order to model the heat transfer and fluid flow in the heat sink (2D fluid flow and 3D heat transfer) [14]: (1) steady state; (2) the fluid is assumed to be incompressible (Although air is a compressible fluid, for the range of flow velocities considered in this study, incompressible flow assumption is valid as long as the Mach number is smaller than 0.3 [15]); (3) laminar flow, because of the small fin spacing and low air flow rates; (4) constant fluid and solid properties; (5) negligible viscous dissipation; (6) negligible radiation heat transfer. The radiation effect is important in the natural convection system; it can easily be ignored when convection becomes a significant term. The dimensions of the heat sink considered in this work are listed in Table 1. Only a quarter of the heat sink was included in the computational domain, in view of the symmetry conditions.

Table 1
Geometry of the heat sink used in the computation

$H(m)$	$h_a(m)$	$h_b(m)$	$L(m)$	$l_c(m)$	N_f	$t_f(m)$	$W(m)$
0.036	0.032	0.004	0.082	0.0015	27	0.001	0.066

2.2 Mathematical Model

The governing equations of continuity, momentum and energy are solved numerically using a finite difference scheme with boundary conditions as follows. The velocity is zero on all wall boundaries, and as flow is assumed to be hydrodynamically developed [9].

(1) Continuity equation

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0 \tag{1}$$

(2) Momentum equation

- *for the fluid flow:*

$$\rho\left(u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}\right)=-\frac{\partial p}{\partial x}+\mu\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right) \tag{2}$$

$$\rho\left(u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}\right)=-\frac{\partial p}{\partial y}+\mu\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right) \tag{3}$$

- *for the solid :*

$$u=v=0 \tag{4}$$

(3) Energy equation

- *for the fluid:*

$$\rho.c_p\left(u\frac{\partial T}{\partial x}+v\frac{\partial T}{\partial y}\right)=\lambda_f\left(\frac{\partial^2 T}{\partial^2 x}+\frac{\partial^2 T}{\partial^2 y}+\frac{\partial^2 T}{\partial^2 z}\right) \tag{5}$$

- *for the solid :*

$$\lambda_s\left(\frac{\partial^2 T}{\partial^2 x}+\frac{\partial^2 T}{\partial^2 y}+\frac{\partial^2 T}{\partial^2 z}\right)=0 \tag{6}$$

The assumption of constant physical properties for the solid and the air allows decoupling the hydrodynamic equations of heat equations. Thus, solving the energy equation may be conducted after the convergence of computing hydrodynamics. In addition, the energy equation is linear; it converges faster than the Navier-Stokes. The equations were solved using a C++ developed code.

The developed code is based upon the finite differences approach. The discretized equations are solved iteratively using the Gauss-Seidel method.

The Reynolds number of the impingement jet is defined as:

$$\mathrm{Re} = \frac{|V_0|.D_h}{\nu} \tag{7}$$

Where $D_h = \frac{4A}{P_m} = \frac{4.h_a.l_c}{2h_a + l_c}$ (8)

The average convection heat transfer coefficient h is calculated by

$$h = \frac{Q}{A_h(T_s - T_a)} \tag{9}$$

The average Nusselt number *Nu* is calculated by

$$Nu = \frac{h.D_h}{k_a} \tag{10}$$

For very small distances from the rectangular duct inlet, the effect of curvature on the boundary layer development is negligible. Thus, it should approach the classical isothermal flat plate solution for developing flow in the entrance region of rectangular ducts ($L^* \rightarrow 0$) [10].

$$L^* = (L/2)/(D_h \,\mathrm{Re}\,\mathrm{Pr}) \tag{11}$$

$$Nu = 0.664\,\mathrm{Re}^{1/2}\,\mathrm{Pr}^{1/3} \tag{12}$$

Assuming Pr=0.707 for air, [16].

The thermal resistance of the heat sink is calculated by

$$R_{th} = \frac{T_{\max} - T_\infty}{Q} \tag{13}$$

Note that the boundary conditions of this problem are stated in Figures 2 and 3.

This enables us to set properties in the solid and the fluid regions appropriately and to solve the conjugate conduction convection problem [16]. The iterations

were terminated when the residuals for the continuity, momentum equations were 1% of the characteristic flow rate and were 0.1% for the energy, the solution is converged.

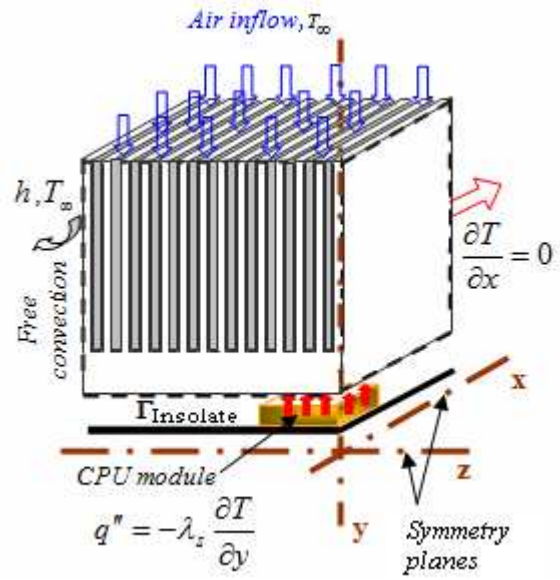


Figure 2
Thermal boundary conditions

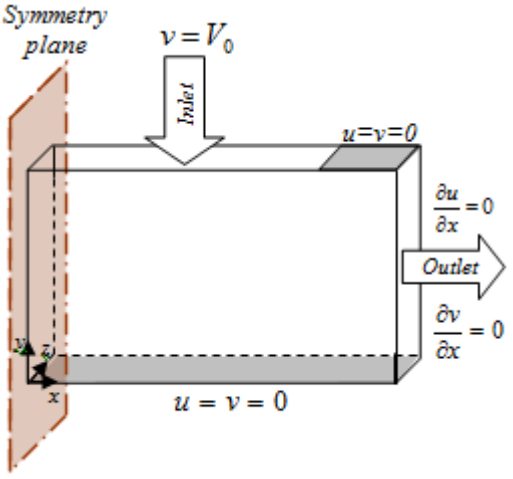


Figure 3
Hydrodynamics boundary conditions

3 Results and Discussion

The model is validated with the numerical results taken on the studied heat sink with the operating parameters listed in Table 2.

Table 2
Operating parameters of the model, [13]

Parameter	Value	Description
V_0	5 m/s	inlet velocity (m/s)
h	10 W/m².K	Natural convection Coefficient (both sides of heat sink), [17]
T_∞	27 $^\circ C$	ambient temperature.
Q	80 W	power of heat source (W)
A_{Die}	14x16 mm²	Die Surface of CPU.
A_{CPU}	31x 31 mm²	Surface of CPU.

Figure 4 shows the distribution of static pressure through the channel of the heat sink (XY plane). The static pressure increases gradually up to the stagnation zone at the center where the velocity is the lowest, then decreases in the direction of the air outlet. A low pressure drop is located at the exit channel where the velocity is highest. The numerical results (Figures 4 and 5) show that the pressure drops lie in a range of 5-62 Pascal, which is not so large. It indicates proper fin spacing. This is an advantage that improves heat transfer with less energy pumping. The numerical results of the pressure drop are compared with experiment data for a similar geometry and flow velocities for impingement flow in the laminar regime obtained by Duan and Muzychka, [9] and Saini and Webb [5] as shown in Figure 5. Overall, the trend is very good. The results are in good agreement in view of the simplicity of the model.

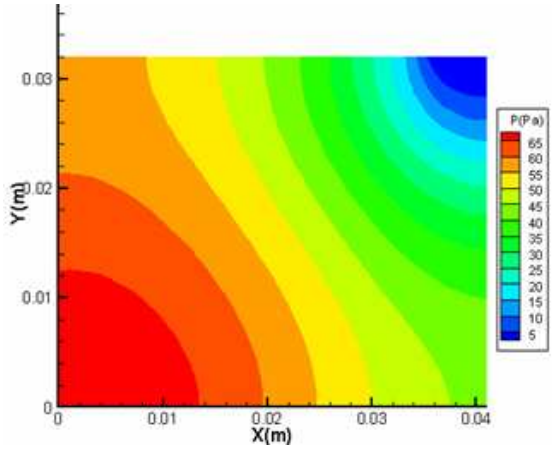


Figure 4
Static pressure contours through the channel

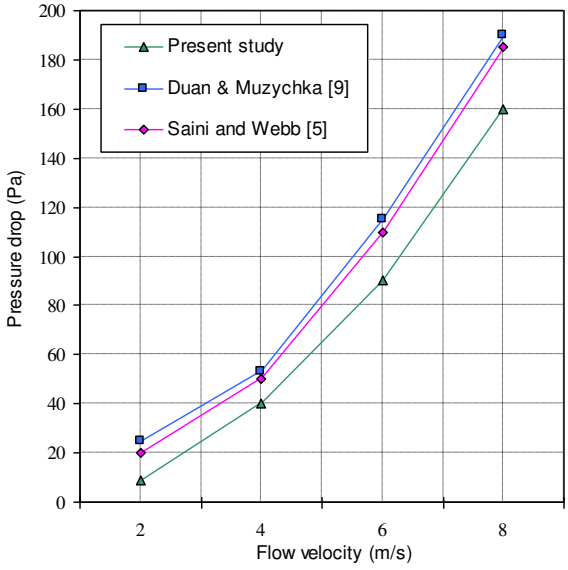


Figure 5

Pressure drop through the mini-channel as a function of cooling velocity (V_0)

Figure 6 shows the average velocity field in the channel of the heat sink. The results show clearly that the velocity slows down by going to the base until a breakpoint in the center. This velocity increases rapidly in the direction of the exit channel and reaches higher values than those at the entrance; it is the boundary layer phenomenon. These conclusions are in excellent agreement with those of Biber [7].

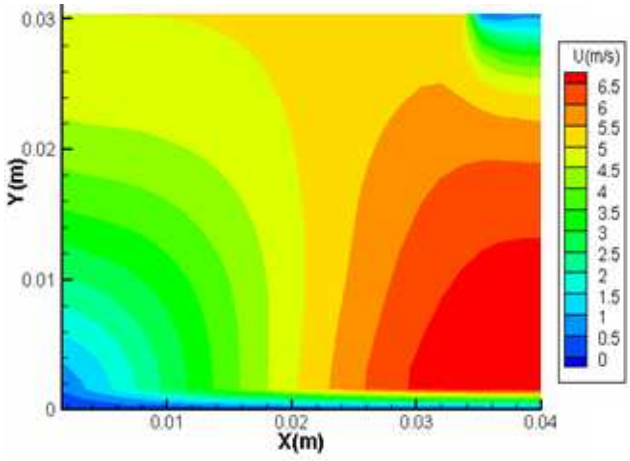


Figure 6

Velocity contours through the channel

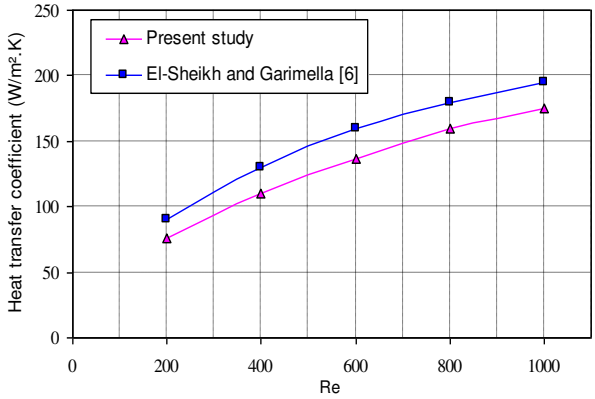


Figure 7

Variation of convective heat coefficient as a function of cooling velocity (V_0)

From Figure 7, the heat transfer coefficient is seen to increase as the jet velocity increases. Consequently, heat dissipation from heat sinks may be enhanced by increasing the cooling velocity. The obtained results are in good agreement with those of El-Sheikh and Garimella [6]. On the other hand, increasing jet velocity affects negatively the pressure drop and consequently flow by pass results.

The results of the temperature profile in Figures 8, 9 and 10 show a peak (T_{max}) at the heat source, which is obvious, but also a high temperature gradient in the center of the sink, where it should be a low airflow following, while the corner points have a minimum temperature. Figure 10 depicts the contours of temperature in the sink at the central cross-section. The maximum temperature is always observed in the heat source; this temperature under the mentioned conditions is about 46°c, which is less than the temperature restricted by manufacturer.

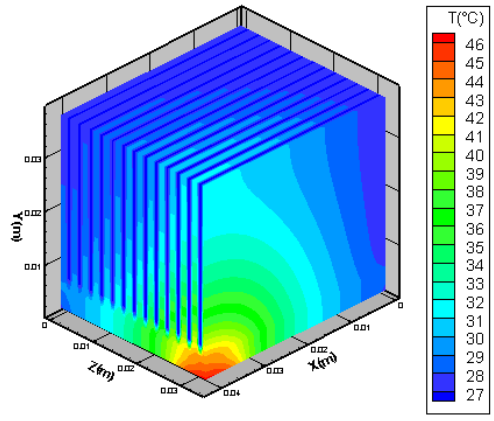


Figure 8

Temperature distribution through the heat sink, (3D)

The temperature distribution through the base of the sink is presented in Figure 9. The results are obtained by the developed code. The hottest spot ($T_{max}=46°c$) is located in the center of the base of the sink, as well as the pressure drop. This is due primarily to the presence of the heat source at this level, and as well because the cooler fluid has the lowest traffic there (stagnation zone).

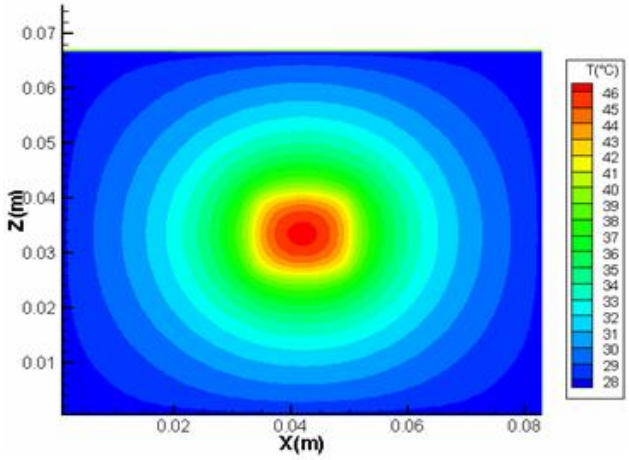


Figure 9
Contours of temperature distribution on the base of the heat sink

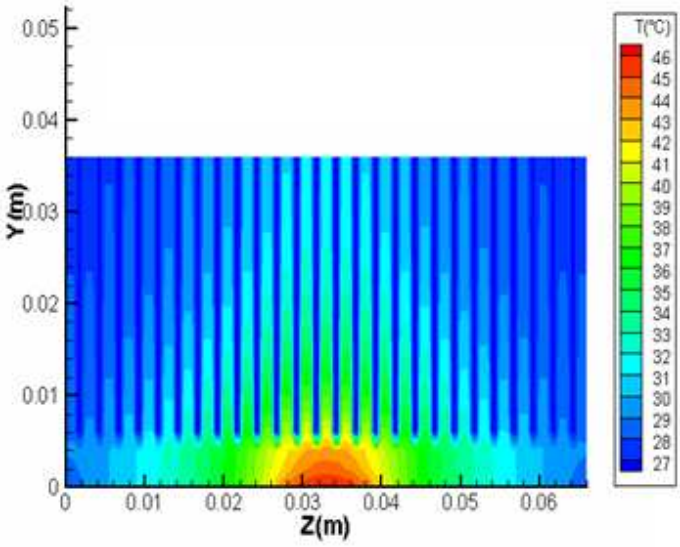


Figure 10
Temperature distribution in the sink at the central cross-section

In addition, heat is dissipated in both longitudinal and transverse directions, where it appears that the length of fin beyond a limit of 38 mm does not contribute to a decrease in temperature, and so fin length can be reduced. However, as regards width, the temperature can be dissipated through greater width. These conclusions, drawn in present work, are found to be in good agreement with conclusions drawn by Shah et al. [11] and Jang and Kim [4]. The cooling performance of the mini-channel heat sink subject to an impinging jet can be evaluated in terms of the thermal resistance [4]. The temperature difference source-air is about 19 °C, so the thermal resistance R_{th} is calculated from Eq. (13) based on numerical results, it is of 0.2375 °C/W under fixed operating parameters of the model summarized in Table 2.

Figure 11 shows that thermal resistance is decreased if the flow velocity is increased. To check the validity of the results for thermal resistance of the mini-channel heat sink subject to impinging jet, a comparison is made with Saini and Webb's [5] experimental data, and Duan and Muzychka's [9] analytical model. Overall, the trend is very good.

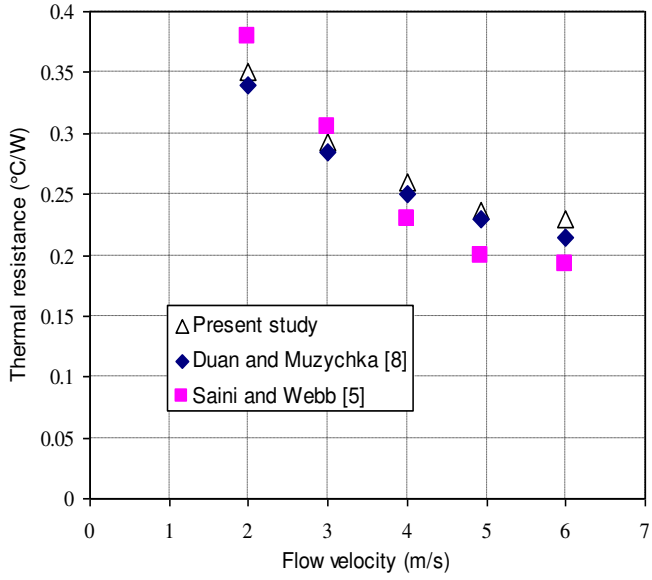


Figure 11
Thermal resistance comparison

The obtained numerical results revealed a detailed picture of the flow and temperature distributions in the entire domain. They have provided a larger amount of data than an experiment can generally provide. Therefore, these results also show that there is potential for optimizing the geometric design, specifically for improving the low heat transfer and flow within the stagnation zone, as mentioned above.

Conclusion

In this research work, thermal and hydraulic characterization have been carried out on a parallel plate heat sink, considering real operating parameters. The developed code was validated with available results, and the obtained results show that impingement air flow on mini-channel heat sink appears to have advantages for heat transfer. Thermal performances numerically carried out show that the chip delivers up to 80 W, which is adequately cooled by the studied heat sink as recommended by the manufacturer. The increase in the Reynolds number improves the convective heat transfer coefficient, which provides better thermal performance. The results also show that there is potential for optimizing the geometric design. This study should serve designers involved in the cooling of electronic components. In other words, it can be used to select or design heat sinks for effective thermal management in electronic assemblies. The obtained results have been favourably compared with available results in literature.

Nomenclature

A	- section of the channel (m²)
A_h	- heat transfer area (m²)
C_P	- specific heat coefficient (J/kg K)
D_h	- hydraulic diameter (m)
h	- convective heat transfer coefficient (w/m².K)
h_b	- height of the base (m)
H	- height of a minichannel fin (m)
ha	- height of a minichannel fin (m)
λ	- thermal conductivity of a aluminium (W/m.K)
Ka	- thermal conductivity of air (W/m.K)
L	- length of the base of a minichannel heat sink (m)
l_c	- channel width (m)
L^*	- dimensionless thermal developing flow length
$\dot{m}$	- mass flow rate (kg/s)
N_f	- number of fins
n	- number of channels
Nu	- Nusselt number
p	- pressure (Pa)
Q	- power of heat source (W)
q''	- heat flux (W/m²)
R_e	- Reynolds number
R_{th}	- thermal resistance (°C/W)
t_f	- fin thickness (m)
T	- temperature (°C)
u	- velocity in the x-direction (m/s)
V	- velocity in the y-direction (m/s)

V_0 - inlet velocity (m/s)
$\dot{V}$ - air flow rate (m^3/s).
W - width of the base (m)
w_C - channel width (m)
ρ - air density (kg/m^3)
μ - dynamic viscosity (Ns/m²)
x, y, z - space variables
$\Delta x, \Delta y$ - cell widths

Acknowledgment

The authors are thankful to the Algerian ministry of higher education and research for providing financial support to carry out this research.

References

[1] Y. T. Yang and H.S. Peng, "Numerical Study of Pin-Fin Heat Sink with Un-Uniform Fin Height Design", Int. J. Heat and Mass Transfer, 2008, Vol. 51, pp. 4788-4796

[2] K. Nishino, M. Samada, K. Kasuya, and K. Torii, "Turbulence Statistics in the Stagnation Region of an Axisymmetric Impingement Jet Flow", Int. J. Heat Fluid Flow, 1996, Vol. 17, pp. 193-201

[3] C. Hilbert, S. Sommerfeldt, O. Gupta and D. J. Herrell, "High Performance Micro-Channel Air Cooling", Proceedings of the 6th Annual IEEE Semiconductor Thermal and Temperature Measurement Symposium, Feb. 6-8, Scottsdale, Arizona, 1990, pp. 108-113

[4] S. P. Jang and S. J. Kim, "Fluid Flow and Thermal Characteristics of a Microchannel Heat Sink Subject to an Impinging Air Jet", J. Heat Transf., 2005, Vol. 127, pp. 770-777

[5] M. Saini and R.L. Webb, "Validation of Models for Air-cooled Plane fin Heat Sinks Used in Computer Cooling", Proc. 8th Intersoc. Conf. Therm. Thermomech. Phenom. Electronic Syst., (ICTTPES'02), Pennsylvania State Univ., University Park, PA, USA, 2002, pp. 243-250

[6] H. A. El-Sheikh and S. V. Garimella, "Enhancement of Air Impingement Heat Transfer Using Pin-Fin Heat Sinks", IEEE Trans. Compon. Pack. Technol, 2000, Vol. 23, No. 2, pp. 300-328

[7] C. R. Biber, "Pressure Drop and Heat Transfer in an Isothermal Channel with Impinging Flow", IEEE Transac. On Comp. and Packag. Tech. – Part A, 1997, Vol. 20, No. 4, pp. 458-462

[8] Z. P. Duan, and Y. S. Muzychka, "Impingement Air-cooled Plate Fin Heat Sinks Part II-Thermal Resistance Model", Proceedings of 9th Int. Soc. Conf. Therm. Phenom. Electronic Syst., 2004, pp. 436-443

[9] Z. P. Duan, and Y. S. Muzychka, "Impingement Air-cooled Plate Fin Heat Sinks Part I-Pressure Drop Model", Proceedings of 9th Int. Soc. Conf. Therm. Phenom. Electronic Syst., 2004, pp. 429-435

[10] Z. P. Duan and Y. S. Muzychka, "Experimental Investigation of Heat Transfer in Impingement Air-cooled Plate Fin Heat Sinks", J. Electron. Packag, 2006, Vol. 128, pp. 412-418

[11] A. Shah, B. G. Sammakia, K. Srihari and K. Ramakrishna, "A Numerical Study of the Thermal Performance of an Impingement Heat Sink Fin Shape Optimisation", IEEE Trans. Compon. Packag. Technol., 2004, Vol. 27, No. 4, pp. 710-717

[12] A. Shah, B. G. Sammakia, K. Srihari and K. Ramakrishna, "Optimization Study for a Parallel Plate Fin Impingement Heat Sink", J. Electron. Packag., 2006, Vol. 128, pp. 311-318

[13] Intel® Corporation, Intel® Pentium® 4 processor on 90 nm process thermal and mechanical design guidelines, Design Guide, 2004

[14] P. S. Lee and S. V. Garimella,: Thermally Developing Flow and Heat Transfer in Rectangular Microchannels of Different Aspect Ratios, Proc. Int. J. Heat Mass Transf., 2006, Vol. 49, pp. 3060-3067

[15] R. L. Panton,: Incompressible Flow. John Wiley and Sons, 1984

[16] S. V. Patankar,: Numerical Heat Transfer and Fluid Flow. Washington: Hemisphere, 1980

[17] J. P. Holman,: Heat Transfer. 8th SI Metric Edition. New York: McGraw-Hill Book Co., 1996

[18] Builder C++,: User's Guide. Borland Software Corporation, 2002

Performance Analysis of Equally weighted Portfolios: USA and Hungary

András Urbán, Mihály Ormos

Department of Finance, Budapest University of Technology and Economics,
Magyar tudósok körútja 2, H-1117 Budapest, Hungary
urban@finance.bme.hu
ormos@finance.bme.hu

Abstract: Investigating U.S. equally weighted portfolios, one can measure positive abnormal returns (Jensen alphas) according to the classical equilibrium models. Applying the Carhart four-factor model, we show that excess returns generated by the equally weighted multi-period investment strategy are neither caused by the small-firm effect, nor by the book-to-market equity, nor even by persistence. We document that this phenomenon cannot be observed in the Hungarian stock market, where the equally weighted rebalancing strategy neither achieves significant abnormal return, nor outperforms the value weighted index in terms of mean return. This latter result suggests that, from this point of view, the Hungarian capital market exhibits a higher level of efficiency than its US counterpart.

Keywords: equally weighted portfolio, performance measure, market efficiency

1 Introduction

We investigate a simple multi-period investment strategy using equally weighted portfolios by comparing the performance of a value weighted market index to equally weighted portfolios. An equally weighted portfolio takes every asset into account with the same weight, while in a value weighted portfolio, the market capitalization determines the weight of a stock. In the case of the U.S. stock market, we document positive abnormal returns for the equally weighted portfolios using the Capital Asset Pricing Model (CAPM) by Sharpe (1964), Lintner (1965), and Mossin (1966), and the Carhart (1997) Four-Factor Model. We argue that the excess return is neither due to the small firm effect documented by Banz (1981) and Reinganum (1980, 1981), nor the book to market equity factor documented by Basu (1983). This phenomenon cannot be observed in the Hungarian stock market; furthermore, the equally weighted portfolio does not outperform the value weighted market index in terms of mean return.

We even state that the negative autocorrelation caused by the mean reverting behavior of stock returns (see French and Roll (1986), Fama and French (1988), Poterba and Summers (1989) or De Bondt and Thaler (1985, 1987)) has no effect on the return of the equally weighted portfolios. Rather, holding a portfolio compiled by rebalancing different random processes gains higher returns by the nature of the stochastic processes. Opposite to the Budapest Stock Exchange (BSE), on equally weighted portfolios formed from U.S. stocks, one can measure much higher returns than that of a value weighted portfolio. A large number of explanations can be found in the literature which try to give some theoretical background for the significant difference. If an equally weighted portfolio is investigated, the first argument is connected to small firm effect. As Roll (1981) states, "a value weighted index such as the S&P 500 is obviously more heavily invested in large firms than is an equally weighted index. Thus, comparing the behavior of two such indexes will enable us to study, with very little effort, the size effect." In other words he argues that the difference between the returns of similar risky portfolios in that the behavior comes from the size differences. Roll argues that the small firm effect is the result of a measurement problem and trading infrequency seems to be a powerful cause of bias in risk assessments with short-interval data. Rather horrendous bias is induced in daily data and the bias is still large and significant with returns measured over intervals as long as one month. In our analysis, we use monthly returns instead of daily ones. This argument is similar to Banz's (1981) results.

The other reasoning according to the higher return is concentrated on the autocorrelation in the process; however, the autocorrelation of data has a time varying behavior (see Li and Yen, 2011). In the short run, one can measure a mean reverting price behavior, which in turn means negative serial autocorrelation; see e.g. Dyl and Maxfield (1987), Bremer and Sweeney (1988) or Brown et al. (1988). On longer time intervals, for weekly returns, Howe (1986) or Lehmann (1988) measures also negative autocorrelations. Similarly, for monthly returns, Rosenberg et al. (1985), Jagedeesh (1990), Brown and Harlow (1988) measure negative autocorrelation. For even longer intervals, for twelve months, Jagedeesh (1990) documents positive autocorrelation. However, investigating much longer intervals, e.g. DeBondt and Thaler (1985, 1987), Poterba and Summers (1989) or Fama and French (1988) report again a negative serial correlation in market returns over observation intervals of three to five years. In the case of an equally weighted portfolio, like e.g. the S&P equally weighted index (S&P EWI), which is compiled from the largest 500 U.S. stocks, these autocorrelations have a high impact on the return. However, the studies state that the abnormal return of the S&P EWI is due to the small firm effect, neglecting the findings according to the autocorrelations. The S&P EWI is a quarterly rebalanced portfolio; therefore, the negative autocorrelation measured for this interval has a positive effect on the return because the equally weighted portfolio increases the weights of the "past losers" and at the same time decreases the weights of the "past winners", where the past means three months' performance. However, if the process is mean-reverting, the

return generated by this contrarian investment strategy must be higher than the return of the value weighted index. Of course, because of its nature, it is not difficult to see that it would be helpful to have such processes to achieve better performance. However, regarding Mulvey and Kim (2008), the truth is that mean-reversion is not necessary for the fixed mix to accomplish superior performance. Stein et al. (2009) investigate the diversification and rebalancing of Emerging Market countries' portfolios. They show that even though Emerging Markets suffer high transaction costs and unreliable information, pragmatic portfolio implementations such as equally weighted rebalancing with relatively little trading still promise excess performance. However, our findings on BSE stock portfolio does not support this issue.

The remaining question is whether the difference in the return, if it exists, can be explained by the standard equilibrium models (the CAPM and Four-Factor Model) or not. If not, i.e., significant positive alpha can be measured especially by the four factor model, this means that the strategy promises excess return, above the equilibrium, where the small-firm-effect, the book-to-market equity effect and the effect of persistency is already managed. In fact, our results for survivorship biased dataset using the components of the Standard and Poor's 500 index components in the rebalancing strategy gains significant positive or non-significant but positive alphas. The same strategy formed from BUX index (the main Hungarian equity index) components only provide non-significant positive alpha.

2 Stock Market Model

The model of stock market investigated in this section is the one considered, among others, by Luenberger (1998), Mulvey and Kim (2008). Consider a market of d assets whose mean return vector is $\mathbf{r}$ where $\mathbf{r} \in R^d$. Let $\mathbf{S} \in R^{d \times d}$ be a covariance matrix. Assuming normality for the return's joint distribution, a static portfolio's return according to the portfolio vector $\mathbf{w} \in R^d$ is also normal with mean $\langle \mathbf{w}, \mathbf{r} \rangle$ and variance $\langle \mathbf{w}, \mathbf{S}\mathbf{w} \rangle$, where $\langle \bullet, \bullet \rangle$ denotes inner product. Let us investigate a constantly rebalanced portfolio made of the same stocks and rebalanced in each instantaneous moment according to $\mathbf{w}$. We can write the following stochastic differential equation for the price process of asset i as

$$\frac{dp_t}{p_t} = (r_i + \frac{1}{2}\sigma_i^2)dt + dB_i^t, \tag{1}$$

where r_i and σ_i^2 are the return and variance of asset i, respectively. B is a geometric Brownian motion. Thus, for the value of a constantly rebalanced portfolio we have

$$\frac{dP_t}{P_t} = \sum_{i=1}^{d} w_i \frac{dp_t}{p_t} = \sum_{i=1}^{d} w_i \left\{ (r_i + \frac{1}{2}\sigma_i^2)dt + dB_i^t \right\}. \tag{2}$$

The portfolio's growth rate is the weighted average of the individual asset's rate, that is, we can write

$$\frac{dP_t}{P_t} = \left(\langle \mathbf{w}, \mathbf{r} \rangle + \frac{1}{2} \langle \mathbf{w}, \sigma^2 \rangle \right) dt + \sigma_P^2 \, dW^t \tag{3}$$

for the portfolio's growth rate, where σ_P^2 denotes the portfolio's variance and W^t is an element of a standardized Wiener process. Thus, for the constantly rebalanced portfolio's mean return for a unit period we have

$$r_P = \langle \mathbf{w}, \mathbf{r} \rangle + \frac{1}{2} \langle \mathbf{w}, \sigma^2 \rangle - \frac{1}{2} \sigma_P^2 = \langle \mathbf{w}, \mathbf{r} \rangle + \frac{1}{2} \langle \mathbf{w}, \sigma^2 \rangle - \frac{1}{2} \langle \mathbf{w}, \mathbf{S}\mathbf{w} \rangle, \tag{4}$$

that is, the constantly rebalanced portfolio's mean is larger than the static case by the factor of ½($\langle \mathbf{w}, \sigma^2 \rangle$-$\langle \mathbf{w}, \mathbf{S}\mathbf{w} \rangle$), which is the so-called *rebalancing gain* (see Mulvey and Kim (2008)). Since $\langle \mathbf{w}, \sigma^2 \rangle$ is the weighted sum of the portfolio constituents' variances, its value is equal to the portfolio's variance ($\langle \mathbf{w}, \mathbf{S}\mathbf{w} \rangle$) if and only if the constituents are absolutely correlated. In any other case, the constantly rebalanced portfolio outperforms its static counterpart in terms of mean return respect to that (1) both portfolios consist of the same stocks and (2) the static portfolio's initial capital allocation vector is identical to the rebalancing strategies' **w**. Furthermore, the constantly rebalanced portfolio's returns are also normal with the same variance as the static portfolio ($\langle \mathbf{w}, \mathbf{S}\mathbf{w} \rangle$).

In the next section we investigate two types of equally weighted portfolios to get empirical evidence whether these attractive theoretical properties are manifested in abnormal returns in terms of an equilibrium model or not, and whether these excess returns infer higher risks.

3 Empirical Results

3.1 Portfolio Construction

We investigate 25 portfolios which are formed on the basis of Standard & Poor's 500 large-cap index for 10-year-long periods. We launch a new portfolio at the beginning of each year from 1975 to 1999 in the following way: The portfolios are reviewed each month to map exactly the actual Standard & Poor's 500 constituents. Stocks included or excluded from the index not at the beginning of a month are considered for the whole month. These portfolios are not free of survivorship bias (NFB) since they follow the performance of the actually largest companies. To ease the notation of these portfolios, we refer to them as S&P500

EW. We rebalance all portfolios on the first trading day of each month according to a weight vector which divides the accumulated wealth equally among constituents. For each portfolio a 10-year-long holding period is investigated. The first portfolio is launched in January 1975 and ends in December 1984. Similarly, new portfolios are formed at the beginning of each year until January 1999. We use data of U.S. stock returns from Center for Research in Security Prices (CRSP) database. The returns are merged to Standard & Poor's 500 constituents list from the Compustat North America dataset.

Similarly we form an equally weighted portfolio on the basis of the BUX index. Due to the limited availability of data, five-year-long periods are investigated; that is, we launch a new portfolio at the beginning of each year from 1999 to 2005 in the following way: The portfolios are reviewed each month to map exactly the actual BUX constituents. Although the BSE reopened in 1990, the maturity of the market and the limited availability of consistent data provide the facility to form portfolios in the above-mentioned way. For better comparison, we launch each portfolio in exactly the same way as that of the S&P500; however, in statistical terms, the outcome of this analysis is not representative. Before the analysis we modify by splits, dividends and we recalculate the returns in U.S. dollars, by which we get results that are comparable to the U.S. market. The applied methodology, as in the previous case, is also not free of survivorship bias. The BSE equally weighted portfolio is referred as BUX EW. In this case the capitalization weighted BUX index is used for the comparison of the equally and capitalization weighted portfolios.

3.2 Analysis of Past Performance

In Figures 1 and 2 we present the wealth levels of the introduced S&P500 EW and BUX EW strategies against time. The solid lines are the EW portfolios' wealth. The capitalization weighted indices' wealth are captured by dashed lines. On the U.S. market one can see that both types of the proposed portfolios outperform the capitalization weighted S&P500 index in the sense of final wealth (and almost always in sense of any intertemporal wealth level); however, the equally weighted strategies are more volatile.

The more volatile return induces higher expected return in an equilibrium setting; therefore, if one would like to compare the two styles of portfolio creation, the difference between the returns should be extended with systematic risk measures. Formalizing the method of performance measurement, two equilibrium models are constructed, the classical Capital Asset Pricing Model (CAPM) (see Sharpe 1964, Linter 1965, Mossin 1966), and a Four-Factor Model (see Carhart [6]). More precisely, one can estimate the return of the strategies in the following ways sequentially:

$$r_l^t - r_f^t = \alpha_l + \beta_l (r_m^t - r_f^t) + \varepsilon_l^t \tag{5}$$

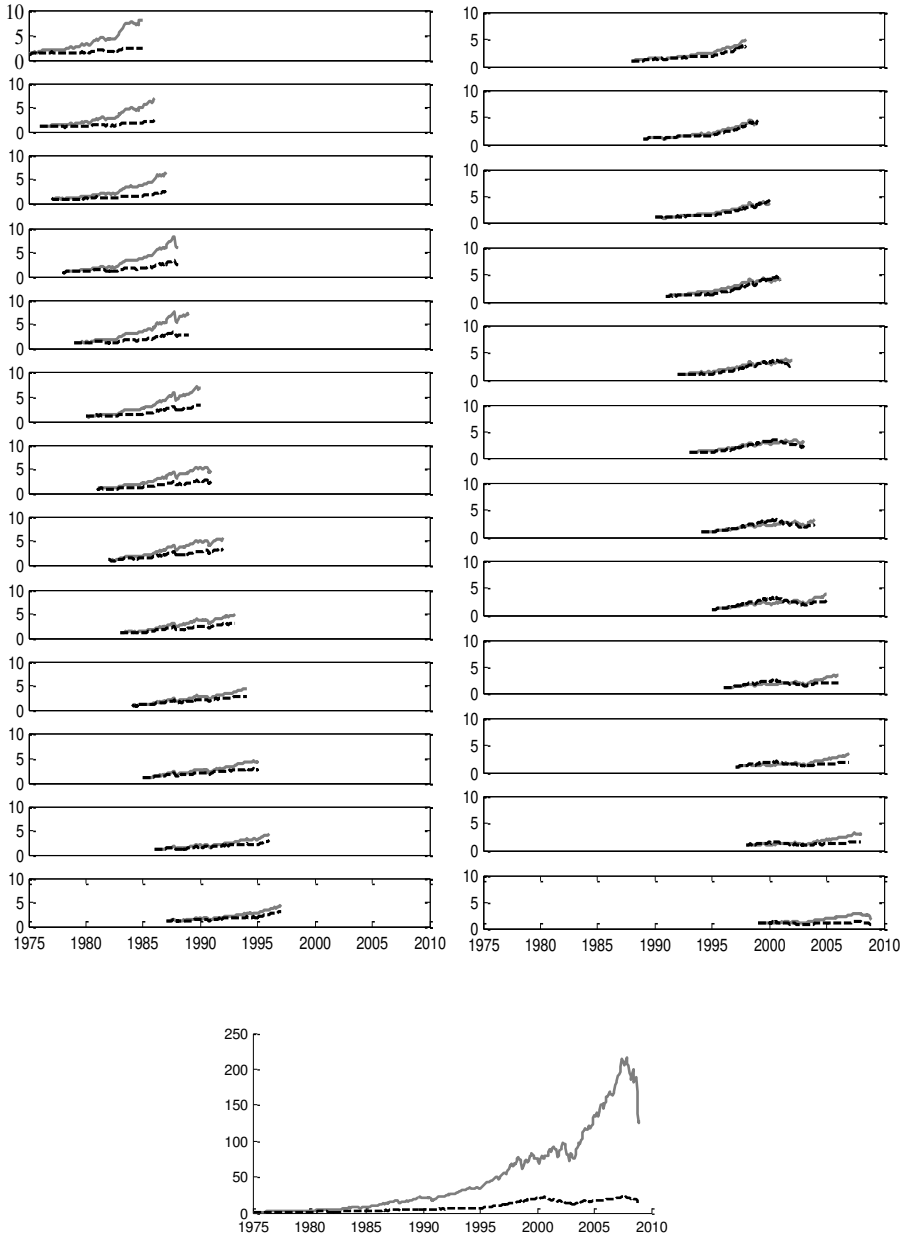


Figure 1
Wealth levels accumulated by the S&P500

Notes: EW portfolios and the capitalization weighted S&P500 market proxy. EW portfolios are launched each January from 1975 until 1999, and rebalanced on the first trading day of each month according to a weight vector which divides the accumulated wealth equally among Standard & Poor's

500 index constituents which appear in the index anytime in the month for the 10-year long period. CRSP-VW is a capitalization based index and needs no rebalancing.

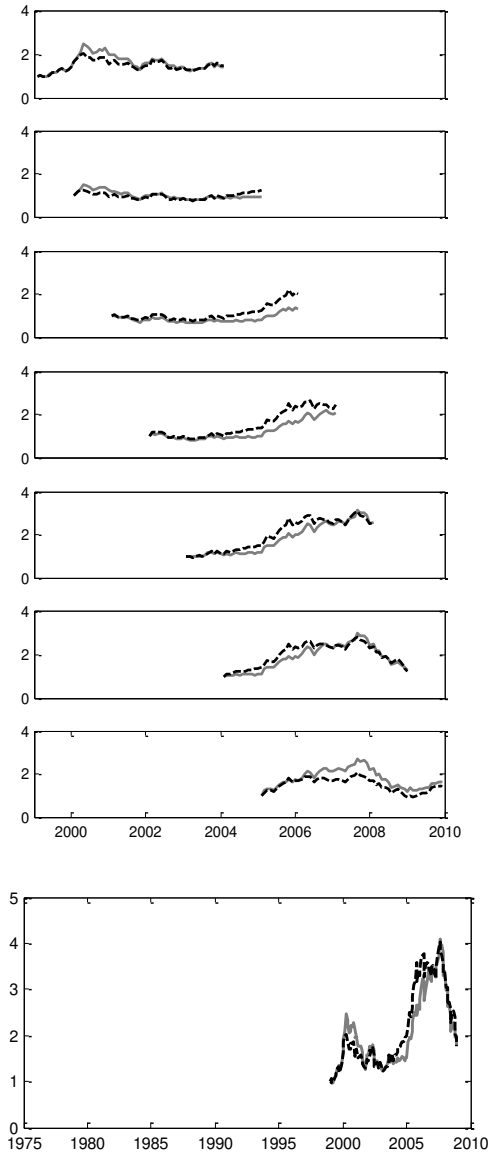


Figure 2
Wealth level accumulated by the BUX

Notes: Figure 2 shows the EW and the capitalization weighted BUX portfolio. BUX index portfolio is launched in January 1999, and rebalanced on the first trading day of each month according to a weight vector which divides the accumulated wealth equally among BUX index constituents which appear in the index anytime in the month for the 10-year long period. The pure BUX index is a capitalization weighted and needs no rebalancing.

$$r_l^t - r_f^t = \alpha_l + \beta_l (r_m^t - r_f^t) + s_l SMB^t + h_l HML^t + mom_l MOM^t + \varepsilon_l^t, \tag{6}$$

where l, t, r_f, $(r_m{}^t - r_f^t)$, ε stand for asset l, time, risk free rate, market premium and estimation residuals, respectively. The risk free rate is the rate of return of the one-month Treasury-bill obtained from Ibbotson and Associates. We use the capitalization weighted S&P500 index as market proxy. According to Fama and French (1993) *SMB* (small-minus-big) measures the average return difference between small and large capitalization assets, while *HML* (high-minus-low) is the average return difference between high and low book-to-market equity (*B/M*) companies. *MOM* is the one-year momentum factor (see Carhart 1997), which shows the average excess return of the last one year's winners above the return of last one year's loser securities[1]. The regression coefficients α, β, *SMB*, *HML* and *MOM* were estimated based on equation (5) and (6).

The alpha parameter measures the excess return of an investment above or below the risk adjusted equilibrium value. According to the Efficient Market Hypothesis of Fama (1965, 1970) and Jensen's (1968) research on abnormal performance of mutual funds, one can achieve repeatedly positive alpha only by chance. Thus, a significant alpha in the model is against the market equilibrium assumptions since it implies systematic abnormal performance in the past returns. We investigate 10-year-long periods in the case of the S&P500 and 5-year-long periods for the Budapest Stock Exchange BUX, because on the one hand, the return anomalies in this long periods are diminishing, i.e. these intervals are as long that the probability of beating the equilibrium return only by chance is small, and on the other hand, we have the opportunity to run statistical analyses on the returns comparing the periods investigated.

Tables 1 and 2 present the results of the regression analysis based on equations (5) and (6) for all EW portfolios and for various periods. Although each period covered different economic environments, the coefficients do not show remarkable differences in the various periods for the U.S. market except in the case of the small-firm factor loading, which is much smaller in the more recent periods. As the tables show, the majority of sensitivity coefficients are significant at 0.05 level; that is, each factor has unambiguous loading on the EW

[1] *SMB*, *HML* and *MOM* factor portfolio returns are obtained from Kenneth French's homepage

performance. This is not the case in the Hungarian market, where only the β is significant. Regarding our analysis, however, the most important observations are the values of the alphas, which are not always significant, but their significant values are always positive. The monthly abnormal returns in the CAPM are between -0.09% and 0.4%. Regarding the Four-Factor Model their values in the U.S. scatter from monthly 0.04% to 0.29%, where the latter is 3.48% in annum. An interesting fact is that the R^2-s are consistently lower on the closer investing intervals as the portfolios excess returns are also lower for the more recent periods.

Table 1
S&P500 EW and BUX EW Summary Statistics for the CAPM

CAPM for USA						
Period	$r_{SP500\ EW}$-r_f	Std Dev	β	α	R^2	adj R^2
1975-1984	1.10%	5.19%	1.1	**0.40%**	0.928	0.928
1976-1985	0.93%	4.76%	1.06	**0.32%**	0.941	0.941
1977-1986	0.80%	4.77%	1.05	**0.28%**	0.95	0.95
1978-1987	0.90%	5.56%	1.07	**0.29%**	0.959	0.959
1979-1988	0.98%	5.37%	1.06	**0.30%**	0.957	0.957
1980-1989	0.97%	5.27%	1.06	**0.26%**	0.956	0.956
1981-1990	0.69%	5.40%	1.11	**0.26%**	0.968	0.967
1982-1991	0.98%	5.47%	1.12	0.16%	0.974	0.974
1983-1992	0.92%	5.19%	1.13	0.12%	0.975	0.974
1984-1993	0.85%	5.13%	1.14	0.09%	0.977	0.977
1985-1994	0.87%	5.02%	1.13	0.09%	0.976	0.975
1986-1995	0.88%	4.91%	1.13	0.06%	0.973	0.973
1987-1996	0.90%	4.72%	1.12	0.02%	0.97	0.97
1988-1997	1.02%	3.81%	1.1	-0.02%	0.954	0.953
1989-1998	1.01%	4.16%	1.05	-0.08%	0.944	0.943
1990-1999	0.94%	4.18%	1.01	-0.09%	0.909	0.908
1991-2000	1.14%	3.93%	0.86	0.28%	0.76	0.758
1992-2001	0.91%	4.08%	0.85	0.34%	0.766	0.764
1993-2002	0.66%	4.60%	0.9	0.28%	0.796	0.794
1994-2003	0.87%	4.82%	0.92	0.32%	0.803	0.801
1995-2004	1.02%	4.78%	0.92	0.35%	0.8	0.799
1996-2005	0.87%	4.82%	0.93	0.36%	0.804	0.802
1997-2006	0.84%	4.76%	0.92	0.36%	0.799	0.798
1998-2007	0.63%	4.68%	0.93	0.35%	0.789	0.788
1999-2008	0.14%	5.08%	0.99	0.32%	0.817	0.815
Average	0.87%	4.82%	1.03	0.22%	0.907	0.906

CAPM for Hungary						
Period	$r_{BUX\ EW}$-r_f	Std Dev	β	α	R^2	adj R^2
1999-2008	0.74%	7.30%	0.23	0.62%	0.181	0.172
1999-2003	0.82%	7.50%	0.22	0.60%	0.239	0.232
2000-2004	0.11%	6.93%	0.14	-0.01%	0.2	0.193
2001-2005	0.69%	6.82%	0.59	0.21%	0.181	0.172
2002-2006	1.44%	6.52%	0.5	1.20%	0.1	0.091
2003-2007	1.80%	6.36%	0.6	1.63%	0.046	0.037
2004-2008	0.66%	7.15%	0.29	0.64%	0.118	0.109
2005-2009	1.18%	7.55%	0.2	1.15%	0.085	0.076

Notes: Table 1 presents the S&P500 EW and BUX EW Summary Statistics for the CAPM. Portfolios are launched each January from 1975 until 1999. We rebalance the portfolios on the first trading day of each month according to a weight vector which divides the accumulated wealth equally among Standard & Poor's 500 and BUX index constituents which appear in the index anytime in the month. $r_{SP500\,EW} - r_f$ and $r_{BUX\,EW} - r_f$ are the average U.S. Dollar denominated return of the S&P EW and BUX EW portfolios in excess of the one-month U.S. Treasury-bill return. Std Dev refers to the standard deviation of the excess return. β and α are parameters of the OLS regression model (5). Parameter α measures the average abnormal return (significant alphas at 0.05 level are in bold). R^2 is the coefficient of determination. The model selection criteria is the adjusted R^2.

Table 2

S&P500 EW and BUX EW Summary Statistics for the Four-Factor Model

Four-Factor Model for USA									
Period	$r_{SP500\,EW}$-r_f	Std Dev	β	SMB	HML	MOM	α	R^2	adj R^2
1975-1984	1.10%	5.19%	1.08	0.27	0.17	-0.15	0.18%	0.971	0.97
1976-1985	0.93%	4.76%	1.07	0.23	0.14	-0.13	**0.18%**	0.967	0.966
1977-1986	0.80%	4.77%	1.09	0.18	0.12	-0.13	**0.23%**	0.968	0.966
1978-1987	0.90%	5.56%	1.09	0.19	0.09	-0.12	**0.27%**	0.972	0.971
1979-1988	0.98%	5.37%	1.08	0.17	0.08	-0.12	**0.29%**	0.97	0.969
1980-1989	0.97%	5.27%	1.09	0.16	0.08	-0.12	**0.29%**	0.969	0.968
1981-1990	0.69%	5.40%	1.13	0.19	0.11	-0.06	**0.25%**	0.977	0.976
1982-1991	0.98%	5.47%	1.13	0.15	0.09	-0.06	**0.18%**	0.981	0.98
1983-1992	0.92%	5.19%	1.14	0.14	0.08	-0.09	**0.15%**	0.983	0.983
1984-1993	0.85%	5.13%	1.15	0.13	0.07	-0.09	**0.15%**	0.985	0.984
1985-1994	0.87%	5.02%	1.15	0.13	0.07	-0.1	**0.16%**	0.984	0.984
1986-1995	0.88%	4.91%	1.15	0.13	0.07	-0.12	**0.15%**	0.984	0.983
1987-1996	0.90%	4.72%	1.14	0.12	0.09	-0.12	0.10%	0.981	0.98
1988-1997	1.02%	3.81%	1.13	0.1	0.12	-0.13	0.04%	0.974	0.973
1989-1998	1.01%	4.16%	1.1	0.08	0.18	-0.18	0.08%	0.972	0.971
1990-1999	0.94%	4.18%	1.09	0.08	0.22	-0.22	0.10%	0.966	0.965
1991-2000	1.14%	3.93%	1.09	0.04	0.34	-0.24	**0.22%**	0.933	0.93
1992-2001	0.91%	4.08%	1.05	0.06	0.4	-0.19	0.19%	0.932	0.93
1993-2002	0.66%	4.60%	1.05	0.06	0.42	-0.2	**0.26%**	0.95	0.948
1994-2003	0.87%	4.82%	1.06	0.07	0.44	-0.19	**0.24%**	0.953	0.952
1995-2004	1.02%	4.78%	1.06	0.08	0.45	-0.19	0.20%	0.953	0.951
1996-2005	0.87%	4.82%	1.07	0.08	0.45	-0.19	**0.20%**	0.954	0.952
1997-2006	0.84%	4.76%	1.06	0.08	0.45	-0.19	0.15%	0.95	0.949
1998-2007	0.63%	4.68%	1.06	0.08	0.46	-0.19	**0.21%**	0.951	0.949
1999-2008	0.14%	5.08%	1.05	0.06	0.43	-0.17	**0.24%**	0.955	0.953
Average	0.87%	4.82%	1.09	0.12	0.22	-0.15	0.19%	0.965	0.964

Four-Factor Model for Hungary									
Period	r_{BUXEW}-r_f	Std Dev	β	SMB	HML	MOM	α	R^2	adj R^2
1999-2008	0.74%	7.30%	0.64	0.02	-0.08	-0.01	0.89%	0.182	0.174
1999-2003	0.82%	7.50%	0.61	0.05	-0.11	-0.05	0.93%	0.246	0.239
2000-2004	0.11%	6.93%	0.53	-0.02	-0.11	-0.08	0.41%	0.208	0.2
2001-2005	0.69%	6.82%	0.64	0.32	-0.27	0.15	0.51%	0.211	0.203
2002-2006	1.44%	6.52%	0.67	0.17	-0.05	0.22	0.97%	0.123	0.114
2003-2007	1.80%	6.36%	0.45	0.28	0.16	0.34	1.18%	0.085	0.076
2004-2008	0.66%	7.15%	0.7	-0.28	0.32	0.2	0.61%	0.133	0.124
2005-2009	1.18%	7.55%	0.56	-0.14	-0.29	0.03	1.23%	0.1	0.091

Notes: Table 2 shows the S&P500 EW and BUX EW Summary Statistics for the Four-Factor Model. Portfolios are launched each January from 1975 until 1999. We rebalance the portfolios on the first trading day of each month according to a weight vector which divides the accumulated wealth equally among Standard & Poor's 500 and BUX index constituents which appear in the index anytime in the month. $r_{SP500\,EW} - r_f$ and $r_{BUX\,EW} - r_f$ are the average U.S. Dollar denominated return of the S&P EW and BUX EW portfolios in excess of the one-month U.S. Treasury-bill return. Std Dev refers to the standard deviation of the excess return. β, *SMB*, *HML*, *MOM* and α are parameters of the OLS regression model (6). Parameter α measures the average abnormal return (significant alphas at 0.05 level are in bold). R^2 is the coefficient of determination. The model selection criteria is the adjusted R^2.

The betas are slightly higher than one for each portfolio; that is, over-weighting relatively smaller firms against the about 50 giants which dominated the value weighted index appreciably raised the portfolio risk, although the CAPM's betas are slightly lower in more recent periods. According to the model selection criteria (adjusted R^2) the additional factors raise the explanatory power of model (6), especially in the more recent dates. It is worth noting, however, that these portfolios consisted of large-cap firms; the SMB factor had a small, but significantly positive loading on EW premia. Positive HML coefficients imply that over-weighting smaller companies also supports investing in high book-to-market equity stocks. The negative loading on return momentum is the natural attendant of an equally weighted strategy, which gives relatively larger portions for stocks which performed below the average in the previous investment period. First and last, investors who had preferred an equally weighted mixture of U.S. large-cap stocks could achieve extra yield on every 10-year-long period in contrast to the value weighted index, which is not the case in the Hungarian market. For the BUX EW portfolio, both models have very low explanatory power, the factor loadings are insignificant, and the equally weighted index slightly underperforms the capitalization weighted BUX index in terms of final wealth. On the one hand, we argue that the Hungarian capital market exhibits a higher level of market efficiency from this point of view, as the rebalancing strategy gains no significant abnormal returns. This result may contradict its US counterpart; however it confirms the weak form of market efficiency. On the other hand, as the explanatory power of the equilibrium model is very low, one would suggest that besides the BUX value weighted returns, other parameters should have been used to proxy the market. However, if we accept the arbitrage pricing theory by Ross (1976) all efficient portfolio can be used as a reference.

Conclusions and Further Research Directions

We show that the equally weighted portfolios' higher return compared to the capitalization weighted market index cannot be explained by the well-known equilibrium model. We use survivorship biased portfolio setting, and these significantly over-perform the equilibrium models. Using the Four-Factor Model, we show that the excess return generated by the U.S. equally weighted multi-period investment strategy is neither caused by the small-firm effect, nor by the book-to-market equity, nor by the persistence. Contrary to the results for the U.S. markets, on the Budapest Stock Exchange we cannot measure excess return with

the periodically equally weighting rebalancing strategy. Although the explanatory power of the Four-Factor model is low, we state that, from this point of view, a very high level of market efficiency can be measured on the Hungarian capital market.

There are several directions from which the equally rebalanced portfolio strategy can be further investigated. On the one hand, one can argue that the variance or standard deviation or any other risk parameter which is compiled by using these measures is not adequate. One may use the variance of the log returns for calculating the variance, because in this setup the "penalty", i.e. the increase in the variance in the case of a positive trigger, is lower than that of a negative one. This assumption is even more reasonable if, instead of utility maximization, one would use the loss-aversion approach. On the other hand the equilibrium model set up suggests a one-period world. The goal of the one-period portfolio theory (Markowitz 1952) and the rival equilibrium models (like CAPM, APT, Fama-French, or Carhart model), is the optimization of the asset allocation in order to achieve the optimal trade-off between expected one-period return and risk. This supposes a world where the investors optimize their consumptions and investment strategies for that given one period. However, most of the mean-variance analysis handles only static models, contrary to the expected utility models, whose literature is rich in multi-period models, supposing an individual with longer interval than simply one-period thinking. One could suppose that a Data Envelopment Analysis (see Gökgöz, 2010) or fractal analyses (see Bohdalová and Greguš, 2010) would increase the accuracy of our estimations. In the multi-period models the investors are allowed to rebalance their portfolios in each trading period, and therefore their investments may be characterized in different ways in one and multiple periods due to the multiplicative effect of consecutive reinvestments. There is a third direction which seems to be worth a closer look; this is the volatility. The question arises as to what is the mathematical connection in a discrete world between the volatility of the single securities and the return of the portfolio.

Acknowledgments

We are thankful to the anonymous reviewer for the valuable comments and suggestions.

References

[1] Bohdalová, M and Greguš, M.: Fractal Analysis of Forward Exchange Rates, *Acta Polytechnica Hungarica*, Vol. 7, 2010, pp. 57-69

[2] Bremer, M. A. and Sweeney R. J.: The Reversal of Large Stock-Price Decreases, *Journal of Finance*, Vol. 46, 1991, pp. 747-754

[3] Brown, K. C., Harlow, W. V. and Tinic S. M.: Risk Aversion, Uncertain Information, and Market Efficiency, *Journal of Financial Economics*, Vol. 22, 1988, pp. 355-385

[4] Brown, K. C. and Harlow W. V.: Market Overreaction: Magnitude and Intensity, *Journal of Portfolio Management*, Vol. 14, 1988, pp. 6-13

[5] Banz, R.: The Relationship between Return and Market Value of Common Stocks, *Journal of Financial Economics*, Vol. 9, 1981, pp. 3-18

[6] Basu, S.: The Relationship between Earnings Yield, Market Value, and Return for NYSE Common Stocks: Further evidence, *Journal of Financial Economics*, Vol. 12, 1983, pp. 129-156

[7] Carhart, M. M.: On Persistence in Mutual Fund Performance, *Journal of Finance*, Vol. 52, 1997, pp. 57-82

[8] De Bondt, W. F. M. and Thaler R. H.: Does the Stock Market Overreact, *Journal of Finance*, Vol. 40, 1985, pp. 793-805

[9] De Bondt, W. F. M. and Thaler R. H.: Further Evidence on Investor Overreaction and Stock Market Seasonality, *Journal Finance*, Vol. 42, 1987, pp. 557-581

[10] Dyl, E. A. and Maxfield K.: Does the Stock Market Overreact? Additional Evidence, working paper, University of Arizona, 1987, June

[11] Fama, E. F.: The Behavior of Stock Market Prices, *Journal of Business*, Vol. 38, 1964, pp. 34-105

[12] Fama, E. F.: Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, Vol. 25, 1970, pp. 383-417

[13] Fama, E. F., and French K. R.: Permanent and Temporary Components of Stock Prices, *Journal of Political Economy*, Vol. 98, 1988, pp. 246-74

[14] Fama, E. F. and French, K. R.: Common Risk Factors in the Returns on Stocks and Bonds, *Journal of Financial Economics*, Vol. 33, 1993, pp. 3-56

[15] French, K. R., and Roll R.: Stock Return Variances: The Arrival of Information and the Reaction of Traders, *Journal of Financial Economics*, Vol. 17, 1986, pp. 5-26

[16] Gökgöz, F.: Measuring the Financial Efficiencies and Performances of Turkish funds, *Acta Oeconomica*, Vol. 60, 2010, pp. 295-320

[17] Howe, J. S.: Evidence on Stock Market Overreaction, *Financial Analysts Journal*, Vol. 42, 1986, pp. 74-77

[18] Jegadeesh N.: Evidence of Predictable Behavior of Security Returns, *Journal of Finance*, Vol. 45, 1990, pp. 881-898

[19] Jegadeesh, N. and Titman, S.: Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *Journal of Finance*, Vol. 48, 1993, pp. 65-91

[20] Jensen, M.: The Performance of Mutual Funds in the Period 1945-1964, *Journal of Finance*, Vol. 23, 1968, pp. 389-416

[21] Lehmann, B. N.: Fads, Martingales, and Market Efficiency, *Quarterly Journal of Economics*, Vol. 105, 1990, pp. 1-28

[22] Li, M. Y. L. and Yen, S. M. F.: Re-Examining Covariance Risk Dynamics in International Stock Markets Using Quantile Regression Analysis, *Acta Oeconomica*, Vol. 61, 2011, pp. 33-59

[23] Lintner, J.: The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, *Review of Economics and Statistics*, Vol. 47, 1965, pp. 13-37

[24] Markowitz, H.: Portfolio Selection, *Journal of Finance*, Vol. 7, 1952, pp. 77-91

[25] Merton, R. C.: Portfolio Selection under Uncertainty: The Continuous-Time Case, *Review of Economics and Statistics*, Vol. 51, 1969, pp. 247-257

[26] Merton, R. C.: An Intertemporal Capital Asset Pricing Model, *Econometrica*, Vol. 41, 1973, pp. 867-887

[27] Meir, K.: Financial Institutions and Markets, Oxford University Press, USA; 2 edition, 2003

[28] Mossin, J.: Equilibrium in a Capital Asset Market, *Econometrica*, Vol. 34, 1966, pp. 468-483

[29] Mulvey, J. M. and Kim, W. C.: Constantly Rebalanced Portfolios - Is Mean-Reverting Necessary?, working paper Princeton University, Princeton, NJ, USA, 2008

[30] Poterba, J. M. and Summers L. H.: Mean Reversion in Stock Prices: Evidence and Implications, *Journal of Financial Economics*, Vol. 22, 1988, pp. 27-59

[31] Reinganum, M. R.: A Simple Test of the Arbitrage Pricing Theory, Graduate School of Business, University of Southern California, 1980

[32] Reinganum, M. R.: Misspecification of Capital Asset Pricing: Empirical Anomalies Based on Earnings' Yields and Market Values, *Journal of Financial Economics*, Vol. 9, 1981, pp. 19-46

[33] Rosenberg, B., Reid K., and Lanstein R.: Persuasive Evidence of Market Inefficiency, *Journal of Portfolio Management*, Vol. 11, 1985, pp. 9-16

[34] Ross, S.: The Arbitrage Theory of Capital Asset Pricing, *Journal of Economic Theory*, Vol. 13, 1976, pp. 341-360

[35] Stein, D. M., Nemtchinov, V. and Pittman S.: Diversifying and Rebalancing: Emerging Market Countries, *Journal of Wealth Management*, Vol. 11, 2009, pp. 79-88

[36] Sharpe, W. F.: Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, *Journal of Finance*, Vol. 19, 1964, pp. 425-442

Nucleation of a Crack under Inner Compression of Cylindrical Bodies

Ebrahim Zolgharnein, Vagif M. Mirsalimov

Institute of Mathematics and Mechanics of NAS of Azerbaijan
Baku, Azerbaijan
E-mail: ezolgharnein@yahoo.com, mir-vagif@mail.ru

Abstract: The problem of fracture mechanics of crack nucleation in plunger pair bushing is considered. It is assumed that under the repeated reciprocating motion of a plunger there happens crack nucleation and a failure of materials of pair elements. Crack nucleuses are simulated by a bridged prefracture zone that is considered as areas of weakened interparticle bonds of the material. It is assumed that the inner boundary of the bushing is close to the annular one and has rough surfaces.

Keywords: contact pair; nucleation of a crack; bonds between surfaces; prefracture zone; cohesive forces; rough surfaces

1 Introduction

The bushing-plunger friction pair operates in conditions of a complex stress state. Experience in using a plunger pair shows that the initiation of cracks and the fracture of the materials of the components of the friction pair occur during repeated reciprocating motion. To control the friction and wear processes in the friction pair, the investigation of material fracture and the friction caused by contact interaction and accompanied by the joint action of contact pressure and friction force are necessary. It is therefore necessary in the planning stage of the construction of sliding pairs to take into account the possibility of the occurrence of cracks and to carry out a limit analysis of the components of the contact pair. It should be taken into account that the bushing internal contour and the plunger external contour are nearly circular. As is known, real treated surfaces are never absolutely smooth but always have micro- or macroscopic irregularities (of a technological character) forming the rough surface. Despite the extremely small sizes of such irregularities, they affect the different service properties of tribo-conjugation [1, 2].

2 Formulation of the Problem

The contact deformation of cylindrical bodies of close radii under inner compression is considered. It is assumed that the surfaces of the bodies in the contact area are rough.

We assume that the outer cylinder (bushing) is an unrestricted plate with a hole close to circular, into which is inserted elastic cylinder (shaft). A concentrated force P is pressing into the hole's boundary and concentrated pair whose moment is determined from the cylinder's limit equilibrium condition under the action of Coulomb friction forces is applied to the center of the shaft (Fig. 1).

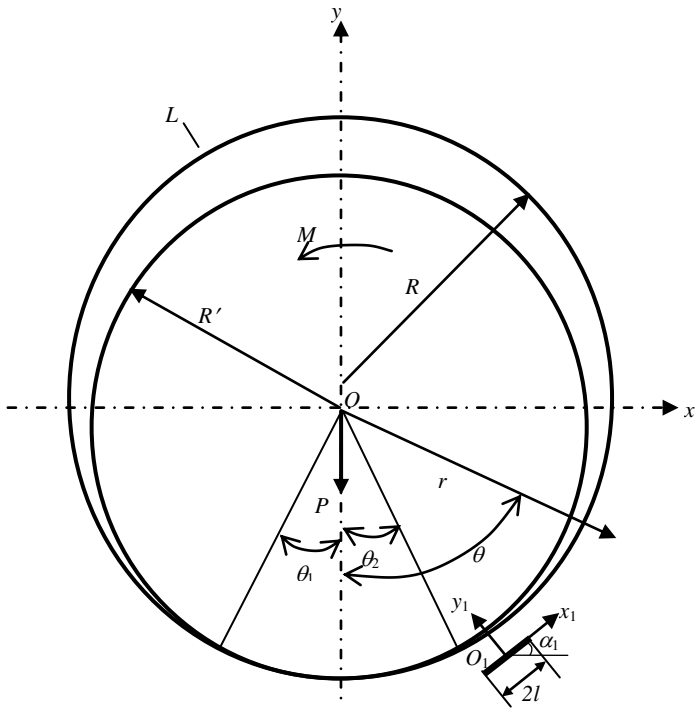


Figure 1
Computational diagram of a problem of the contact fracture mechanics

To determine the contact pressure, it is necessary to consider [3, 4] the contact problem of the pressing of a plunger into the surface of the bushing involving wear.

Let some unknown part shaft with mechanical characteristics G_1 and μ_1 be retained against the internal surface of the bushing with mechanical characteristics G (shear modulus) and μ (Poisson ratio). The condition relating the displacements of the bushing and the plunger is written in the form [3, 4].

$$v_1 + v_2 = \delta(\theta) \qquad \theta_1 \le \theta \le \theta_2 \tag{1}$$

Here $\delta(\theta)$ is sag of the point on the surface of the bushing and the plunger, which is determined by the form of the inner surface of the bushing and the plunger surface, and, also by the magnitude of the pressing force P. $\theta_2 - \theta_1$ is the magnitude of the contact angle area.

In the contact area, in addition to the contact pressure, there is a tangential stress $\tau_{r\theta}$ which is related to the contact pressure $p(\theta,t)$ by the Coulomb law

$$\tau_{r\theta}(\theta,t) = fp(\theta,t) \tag{2}$$

where f is the coefficient of friction of the "bushing-plunger" pair.

We refer the bushing of the contact pair to the polar system of coordinates $r\theta$; for that we choose an origin of coordinates at the center of circle L of radius R.

We will assume that the inner contour of the bushing and the external contour of the plunger are close to annular one.

Represent the boundary L' of inner contour of the bushing in the form

$$r = \rho(\theta), \quad \rho(\theta) = R + \varepsilon H(\theta)$$

where $\varepsilon = R_{\max}/R$ is a small parameter; $R_{\max}$ is the greatest height of the bulge (cavity) of the unevenness of the friction surface.

The coefficients of the Fourier series for the function $H(\theta)$:

$$H(\theta) = \sum_{k=0}^{n} (a_k^0 \cos k\theta + b_k^0 \sin k\theta)$$

are found by means of a profilogram of the treated surface of the bushing which describes each inner profile of the bushing. In a similar way, the plunger contour may be represented as

$$\rho_1(\theta) = R' + \varepsilon H_1(\theta), \qquad H_1(\theta) = \sum_{k=0}^{n} (a_k^1 \cos k\theta + b_k^1 \sin k\theta)$$

It is assumed that the bushing and plunger wear is of an abrasive character. For displacements of the points of surface of the bushing we have

$$v_1 = v_{1e} + v_{1r} + v_{1w} \tag{3}$$

where v_{1e} are elastic displacements of bushing's contact surface; v_{1r}, v_{1w} are displacements caused by the removal of the micro-bulges and by the bushing surface wear, respectively.

Similarly, for displacements of plunger's contact surface we have

$$v_2 = v_{2e} + v_{2r} + v_{2w} \tag{4}$$

The rate of change of the displacements of the surface in the course of the bushing and the plunger wear will be [4]

$$\frac{dv_{jи}}{dt} = K^{(j)} p(\theta,t) \qquad (j=1,2) \tag{5}$$

where $K^{(j)}$ is the wear coefficient of the bushing and the plunger material (j=1,2), respectively.

Prefracture zones will arise in proportion to the loading of the bushing during the operation of the friction pair with force load, and these zones are modelled as domains in which the interparticle bonds in the material have been weakened. The interaction of the surfaces of these domains is modelled by the introduction of a prefracture zone of the bonds, which have a specified deformation pattern. The physical nature of these and the dimensions of the prefracture zone depend on the form the material. Since the above-mentioned zones (layers) are small compared with the remaining part of the bushing, they can be conceptually eliminated by replacing them with cuts, the surfaces of which interact with one another according to a certain law that the corresponds to the action of the material which has been removed. Taking account of these effects in fracture mechanics is an important but exceedingly difficult problem. In the case being investigated, the occurrence of a crack nucleus involves the transition of a prefracture zone into domain where there are ruptured bonds between the surfaces of the material.

Investigations [5-7] of occurrence domains with a disrupted structured of the material show that, during the initial stage, the prefracture zones have the form of a narrow elongated layer, and then, when the load is increased, a secondary system of zones suddenly appears, and these zones contain material with partially ruptured bonds.

We will assume that the prefracture zone is oriented in the direction of maximal tensile stresses arising in the bushing.

Let us consider the prefracture zone of length $2l$ allocated on the segment $|x_1| \le l, \quad y_1 = 0$. At the center of the prefracture zone, we located an origin of local system of coordinates $x_1 O_1 y_1$ whose axis x_1 coincides with the line of the zone and makes an angle with the axis x $(\theta = 0)$. The surfaces of the prefracture zone interact in such a way that this interaction (the bonds between the surfaces) restrains the formation of a crack.

For a mathematical description of the interaction of the surfaces prefracture zones, it is assumed that between them for which the law of deformation is known. Under the action of external loads on the bushing, normal $q_{y_1}(x_1)$ and tangential $q_{x_1 y_1}(x_1)$ tractions will arise in the bonds joining the surfaces prefracture zones.

Consequently, the normal and tangential stresses numerically equal $q_{y_1}(x_1)$ and $q_{x_1y_1}(x_1)$ respectively, will be applied to the surfaces prefracture zones. The quantities of these stresses are not known beforehand and are to be determined when solving the boundary value problem of fracture mechanics.

For determining the displacements v_{1e} and v_{1r} it is necessary to solve the following problem of elasticity theory for a bushing

$$\sigma_n = -p(\theta); \quad \tau_{nt} = -fp(\theta) \quad \text{for} \quad r = \rho \quad \text{in the contact area} \tag{6}$$

$$\sigma_n = 0; \quad \tau_{nt} = 0 \quad \text{for} \quad r = \rho \quad \text{out of the contact area}$$

on surfaces prefracture zone

$$\sigma_{y1} = q_{y1}(x_1); \quad \tau_{x_1y_1} = q_{x_1y_1}(x_1) \quad \text{for} \quad |x_1| \le l, \tag{7}$$

n, *t* are natural coordinates; σ_n, σ_t and τ_{nt} are stress tensor components.

In a similar way, we state the problem of elasticity theory for determining displacements v_{2e} and v_{2r} of the contact surface of the plunger

$$\sigma_n = -p(\theta); \quad \tau_{nt} = -fp(\theta) \quad \text{for} \quad r = \rho \quad \text{in the contact area} \tag{8}$$

$$\sigma_n = 0; \quad \tau_{nt} = 0 \quad \text{for} \quad r = \rho \quad \text{out of the contact area}$$

The magnitudes of θ_1 and θ_2, that is, of the ends of the segment over which the plunger and the bushing are in contact, are unknown. In order to determine them, we will use a condition which expresses the continuous fall of the pressure $p(\theta)$ to zero when then point θ falls outside the segment where the surface touch

$$p(\theta_1) = 0, \qquad p(\theta_2) = 0$$

The equations and conditions (1)-(8) have to be supplemented with a relation between the expansion of the prefracture zone and bond tractions. Without loss of generality, we will represent this relation in the form

$$(v^+ - v^-) - i(u^+ - u^-) = C(x_1, \sigma_1)\left[q_{y_1}(x_1) - iq_{x_1y_1}(x_1)\right], \qquad \sigma_1 = \sqrt{q_{y_1}^2 + q_{x_1y_1}^2} \tag{9}$$

Here the function $C(x_1, \sigma_1)$ may be considered as an effective compliance of the bonds, which depends on their tension; σ_1 is the modulus of the vector of the bond tractions; $(u^+ - u^-)$ is the tangential, $(v^+ - v^-)$ is the normal component of the expansion of the prefracture zone.

In order to determine the value of the external load (the contact pressure) at which a crack is initiated, it is necessary to supplement the formulation of the problem with a condition (criterion) for the appearance of a crack (the rupture of the

interparticle bonds in the material). As such a condition, we will adopt the criterion for the critical expansion of the prefracture zone

$$\left|(v^+ - v^-) - i(u^+ - u^-)\right| = \delta_{cr}$$

where δ_{cr} is a characteristic of the fracture toughness of the bushing material.

The additional condition enables us to determine the parameters of the contact pair for which a crack appears in the bushing.

3 The Method of the Boundary-Value Problem Solution

Using the perturbation method, we find the boundary conditions at each approximation:

for zero approximation of the problem

$$\sigma_r^{(0)} = -p^{(0)}(\theta); \quad \tau_{n\theta}^{(0)} = -fp^{(0)}(\theta) \quad \text{for } r=R \text{ in the contact area} \tag{10}$$

$$\sigma_r^{(0)} = 0; \quad \tau_{n\theta}^{(0)} = 0 \quad \text{for } r=R \text{ out of the contact area}$$

on surfaces prefracture zone

$$\sigma_{y_1}^{(0)} = q_{y_1}^{(0)}(x_1); \quad \tau_{x_1y_1}^{(0)} = q_{x_1y_1}^{(0)}(x_1) \quad \text{for } |x_1| \le l, \tag{11}$$

for the first approximation of the problem

$$\sigma_r^{(1)} = N - p^{(1)}(\theta); \quad \tau_{n\theta}^{(1)} = T - fp^{(1)}(\theta) \quad \text{for } r=R \text{ in the contact area} \tag{12}$$

$$\sigma_r^{(1)} = N; \quad \tau_{n\theta}^{(1)} = T \quad \text{for } r=R \text{ out of the contact area}$$

on surfaces prefracture zone

$$\sigma_{y_1}^{(1)} = q_{y_1}^{(1)}(x_1); \quad \tau_{x_1y_1}^{(1)} = q_{x_1y_1}^{(1)}(x_1) \quad \text{for } |x_1| \le l, \tag{13}$$

Here $$N = -H(\theta)\frac{\partial \sigma_r^{(0)}}{\partial r} + 2\tau_{r\theta}^{(0)}\frac{1}{R}\frac{dH}{d\theta}; \quad \text{for } r=R$$

$$T = (\sigma_\theta^{(0)} - \sigma_r^{(0)})\frac{1}{R}\frac{dH}{d\theta} - H(\theta)\frac{\partial \sigma_r^{(0)}}{\partial r}$$

Similarly we can write the boundary conditions at each approximation for the plunger. Additional relation (9) accepts the following form:

at the zero approximation

$$(v^{(0)+}(x_1,0)-v^{(0)-}(x_1,0))-i(u^{(0)+}(x_1,0)-u^{(0)-}(x_1,0))= \tag{14}$$
$$=C(x_1,\sigma_1^{(0)})\left[q_{y_1}^{(0)}(x_1)-iq_{x_1y_1}^{(0)}(x_1)\right]$$

at the first approximation

$$(v^{(1)+}(x_1,0)-v^{(1)-}(x_1,0))-i(u^{(1)+}(x_1,0)-u^{(1)-}(x_1,0))= \tag{15}$$
$$=C(x_1,\sigma_1^{(1)})\left[q_{y_1}^{(1)}(x_1)-iq_{x_1y_1}^{(1)}(x_1)\right]$$

By means of the Kolosov-Muskheleshvili formulas [8], we write the boundary conditions of the problem at zero approximation (10)-(11) for complex potentials $\Phi^{(0)}(z)$ and $\Psi^{(0)}(z)$. On annular boundaries of the bushing they will be of the form

$$\Phi^{(0)}(z)+\overline{\Phi^{(0)}(z)}-e^{2i\theta}\left[\bar{z}\Phi^{(0)\prime}(z)+\Psi^{(0)}(z)\right]=X^{(0)}(\theta) \tag{16}$$

$$z=Re^{i\theta};\quad X^{(0)}(\theta)=\begin{cases}-(1-if)p^{(0)}(\theta) & \text{on the contact area}\\ 0 & \text{out the contact area}\end{cases}$$

Boundary conditions on the surfaces prefracture zone will be written as:

$$\Phi^{(0)}(z)+\overline{\Phi^{(0)}(z)}+\bar{t}\Phi^{(0)\prime}(t)+\Psi^{(0)}(t)=q_{y_1}^{(0)}+iq_{x_1y_1}^{(0)} \tag{17}$$

where t is affix of points of the surfaces prefracture zone.

We look for the potentials $\Phi^{(0)}(z)$, $\Psi^{(0)}(z)$, $\Phi_1^{(0)}(z)$, $\Psi_1^{(0)}(z)$, $\Phi_2^{(0)}(z)$, $\Psi_2^{(0)}(z)$ and in the form

$$\Phi^{(0)}(z)=\sum_{k=0}^{2}\Phi_k^{(0)}(z),\qquad \Psi^{(0)}(z)=\sum_{k=0}^{2}\Psi_k^{(0)}(z) \tag{18}$$

$$\Phi_1^{(0)}(z)=\frac{1}{2\pi}\int_{-l}^{l}\frac{g_k^0(t)dt}{t-z_1},\qquad \Psi_1^{(0)}(z)=\frac{1}{2\pi}e^{-2i\alpha}\int_{-l}^{l}\left[\frac{\overline{g_k^0(t)}}{t-z_1}-\frac{\overline{T}_k e^{i\alpha}}{(t-z_1)^2}g^0(t)\right]dt \tag{19}$$

$$T_1=te^{i\alpha}+z_1^0;\qquad z_1=e^{-i\alpha}(z-z_1^0)$$

$$\Phi_2^{(0)}(z)=\frac{1}{2\pi}\int_{-l}^{l}\left[\left(-\frac{1}{z}-\frac{\overline{T}_1}{z-\overline{T}_1}\right)e^{i\alpha}g^0(t)+\frac{1-T_1\overline{T}_1}{\overline{T}_1(1-z\overline{T}_1)^2}e^{-i\alpha}\overline{g^0(t)}\right]dt, \tag{20}$$

$$\Psi_2^{(0)}(z)=\frac{1}{2\pi z}\int_{-l}^{l}\left[\left(\frac{1}{zT_1}-\frac{2}{z^2}-\frac{\overline{T}_1}{z(1-z\overline{T}_1)}-\frac{\overline{T_1^2}}{(1-z\overline{T}_1)^2}\right)e^{i\alpha}g^0(t)+\right.$$

$$\left.+\left(-\frac{1}{1-zT_1}+\frac{1-T_1\overline{T}_1}{z\overline{T}_1(1-z\overline{T}_1)^2}-\frac{2(1-T_1\overline{T}_1)}{(1-z\overline{T}_1)^3}\right)e^{-i\alpha}\overline{g^0(t)}\right]dt$$

Here $g^0(t)$ is the required function, which characterizes the expansion of the prefracture zone.

For defining the potentials $\Phi_0^{(0)}(z)$ and $\Psi_0^{(0)}(z)$ we use the N. I. Muskheleshvili method [8]

$$\Phi_0^{(0)}(z) = -\frac{1}{2\pi z}\int_L \frac{X^{(0)}(\sigma)d\sigma}{\sigma - z}, \qquad \sigma = e^{i\theta} \tag{21}$$

$$\Psi_0^{(0)}(z) = \frac{1}{z^2}\Phi_0^{(0)}(z) + \frac{1}{z^2}\overline{\Phi}_0^{(0)}(z) - \frac{1}{z}\Phi_0^{(0)'}(z)$$

Satisfying the boundary condition on the surfaces prefracture zone by the functions (18)-(20), we find singular integral equation with respect to the function $g^0(x_1)$:

$$\int_{-l}^{l}\left[R(t,x_1)g^0(x_1) + S(t,x_1)\overline{g^0(x_1)}\right]dt = \pi\left[q_{y_1}^{(0)} - iq_{x_1y_1}^{(0)} + f^0(x_1)\right] \qquad |x_1| \le l \tag{22}$$

$$f^{(0)}(x_1) = -\left[\Phi_0^{(0)}(x_1) + \overline{\Phi_0^{(0)}(x_1)} + x_1\overline{\Phi_0^{(0)'}(x_1)} + \overline{\Psi_0^{(0)}(x_1)}\right]$$

To the singular integral equation for the inner prefracture zone at zero approximation, we should add equality

$$\int_{-l}^{l} g^{(0)}(t)dt = 0 \tag{23}$$

Using the procedure for converting to an algebraic form [10, 11], the singular integral equation (22) with condition (23) reduced to the system of M complex algebraic equations for determining M unknowns $g^{(0)}(t_m) = v^0(t_m) - iu^0(t_m)$ (m=1,2,…,M)

$$\frac{1}{M}\sum_{m=1}^{M} l\left[g^{(0)}(t_m)R(lt_m, lx_r) + \overline{g^{(0)}(t_m)}S(lt_m, lx_r)\right] = \tag{24}$$

$$= q_{y_1}^{(0)}(x_r) - iq_{x_1y_1}^{(0)}(x_r) + f^{(0)}(x_r)$$

$$\sum_{m=1}^{M} g^{(0)}(t_m) = 0, \qquad r=1,2,\ldots,M\text{-}1$$

where $t_m = \cos\frac{2m-1}{2M}\pi$; $x_r = \cos\frac{\pi r}{M}$.

If in (24) we go over to complexly conjugated values, we get M algebraic equations more. The right hand side of (24) contains unknown values of the forces $q_{y_1}^{(0)}(x_r)$ and $q_{x_1y_1}^{(0)}(x_r)$ in bonds.

The additional relation (14) at zero approximation is the condition determining forces in the bonds arising on the surfaces prefracture zone

$$g^{(0)}(x_1) = \frac{2G}{i(i+k_b)} \frac{d}{dx_1} \left[C(x_1, \sigma_1^{(0)}) (q_{y_1}^{(0)}(x_1) - i q_{x_1 y_1}^{(0)}(x_1) \right] \tag{25}$$

where $k_b = 3 - 4\mu$ for plane strain, $k_b = (3-\mu)/(1+\mu)$ for plane stress state.

For constructing the missing algebraic equations for finding the approximate values of the forces $q_{y_1}^{(0)}(x_r)$ and $q_{x_1 y_1}^{(0)}(x_r)$ at the nodal points, we require the conditions (25) to be fulfilled at the nodal points. For that, we use the finite differences method.

We need two complex equations determining the dimensions of the prefracture zone for closeness of the obtained system. Writing the stress finiteness conditions, we find two missing equations more in the following form:

$$\sum_{m=1}^{M} (-1)^m g^{(0)}(t_m) \cot \frac{2m-1}{4M} \pi = 0 \tag{26}$$

$$\sum_{m=1}^{M} (-1)^{M+m} g^{(0)}(t_m) \tan \frac{2m-1}{4M} \pi = 0$$

By means of complex potentials (18)-(20) and the Kolosov-Muskheleshvili formulae [8] and integration of the kinetic equation (5) wear of bushing's material at zero approximation, we find the displacements $v_1^{(0)}$ of the bushing's contact surface. In a similar way, we find the solution of the elasticity theory problem for the shaft in the first approximation. Using the solution and kinetic equation of shaft's material wear at zero approximation, we find the displacements $v_2^{(0)}$ of the shaft's contact surface.

We substitute the found quantities $v_1^{(0)}$ and $v_2^{(0)}$ into the basic contact equation (1) at zero approximation

$$p^{(0)}(\theta, t) = p_0^0(\theta) + t p_1^0(\theta) + \ldots; \tag{27}$$

$$p_0^0(\theta) = \sum_{k=0}^{\infty} (\alpha_k^0 \cos k\theta + \beta_k^0 \sin k\theta),$$

$$p_1^0(\theta) = \sum_{k=0}^{\infty} (\alpha_k^1 \cos k\theta + \beta_k^1 \sin k\theta)$$

For the algebraization of the basic contact equation, the unknown functions of the contact pressure at zero approximation are found in the form of expansions. Substituting the relation in the basic contact equation at zero approximation, we get the functional equations for the sequential determination of $p_0^0(\theta)$, $p_1^0(\theta)$, etc. For constructing the algebraic system for finding α_k, β_k, we equate the

coefficients for the same trigonometric functions in the left and right hand sides of the functional of the contact problem. We get an infinite algebraic system with respect to α_k^0 (k=0,1,2,…), β_k^0 (k=1,2,…) and α_k^1, β_k^1, etc.

On account of the unknown quantities θ_1, θ_2 and l_1, the joint system of equations is nonlinear even in the case of linear elastic bonds. To determine the quantities θ_1 and θ_2 ($\theta_1 = \theta_1^0 + \varepsilon\theta_1^1 + \ldots; \theta_2 = \theta_2^0 + \varepsilon\theta_2^1 + \ldots$), we have the condition:

for the zero approximation $\quad p^{(0)}(\theta_1^0) = 0; \qquad p^{(0)}(\theta_2^0) = 0;$

for the first approximation $\quad p^{(1)}(\theta_1^1) = 0; \qquad p^{(1)}(\theta_2^1) = 0.$

The right hand sides of infinite algebraic systems with respect to α_k, β_k contain integrals of unknown function $q^{(0)}(x_1)$. Thus, the infinite algebraic systems with respect to α_k, β_k and finite systems with respect to $q^{(0)}(x_1)$, $q_{y_1}^{(0)}(x_r)$, $q_{x_1y_1}^{(0)}(x_r)$ and l are connected between themselves and they must be solved jointly. The combined system equations even for linear-elastic bonds become nonlinear because of unknown quantities θ_1, θ_2, l. For its solution at zero approximation, the reduction and successive approximations methods were used [10].

In the case of the nonlinear law of deformation of bonds for determining forces in bonds, we also use the iteration algorithm similar to the method of elastic solutions [11]. Nonlinear part of the bonds deformation curve is represented in the form of bilinear dependence, whose outgoing section corresponds to the elastic deformation of bonds ($0 < \mathrm{V}(x_1) < \mathrm{V}_*$) with maximal tension of bonds. For $\mathrm{V}(x_1) < \mathrm{V}_*$, the deformation law was described by a nonlinear dependence determined by two points (V_*, σ) and $(\delta_{cr}, \sigma_{cr})$; moreover, for $\sigma_{cr} \geq \sigma_*$ we have increasing linear dependence (linear hardening corresponds to the elastoplastic deformation of the bonds).

After defining the quantities of the desired zero approximation, we can construct the solution of the problem at the first approximation N and T determined on the base of obtained solution for r=R. The boundary conditions (12), (13) may be written in the form of a boundary value problem for finding complex potentials $\Phi^{(1)}(z)$ and $\Psi^{(1)}(z)$ that we seek in the form of (18), with obvious changes. The further course of the solution is at the zero approximation. The obtained complex integral equation with respect to $g^{(1)}(t)$, $\overline{g^{(1)}(t)}$ under additional condition of type (23) by means of the algebraization system is reduced to the system of M algebraic equations for determining $N_0 \times M$ unknowns $g^{(1)}(t)$ (m=1,2,..M).

The desired expansion coefficients of the contact pressure $p^{(1)}(\theta)$ and the unknown values of forces in bonds $q_{y_1}^{(1)}(x_1)$ and $q_{x_1y_1}^{(1)}(x_1)$ are contained in the right hand side of this system.

The construction of the missing equations for determining the unknown forces at the nodal points and prefracture zone sizes are realized as in the zero approximation. The problem of the theory of elasticity for a shaft at the first approximation is solved in some way. The algebraization of solving the equation of the contact problem at the first approximation is carried out similar to the zero approximation. For that, the desired functions of the contact pressure are represented in the form:

$$p^{(1)}(\theta,t) = p_0^1(\theta) + tp_1^1(\theta) + \ldots; \tag{28}$$

$$p_0^1(\theta) = \alpha_{0.0}^1 + \sum_{k=0}^{\infty}(\alpha_{k,0}^1 \cos k\theta + \beta_{k,0}^1 \sin k\theta);$$

$$p_1^1(\theta) = \alpha_{0.1}^1 + \sum_{k=0}^{\infty}(\alpha_{k,1}^1 \cos k\theta + \beta_{k,1}^1 \sin k\theta);$$

As the result we get infinite linear algebraic systems with respect to $\alpha_{0,0}^1$, $\alpha_{k,0}^1$, $\beta_{k,0}^1$ and $\alpha_{0,1}^1$, $\alpha_{k,1}^1$, $\beta_{k,1}^1$ (k=1,2,…), etc.

4 Analysis of the Simulation Results

The system of equations becomes nonlinear because of the unknown quantities θ_1^1 and θ_2^1. The constructed combined system of equations is closed and under the given functions $H(\theta)$ and $H_1(\theta)$ allows us to find the contact pressure, forces in the bonds, the prefracture zone sizes, the stress-strain state, and the bushing and contact pair wear by numerical calculations. The functions $H(\theta)$ and $H_1(\theta)$, describing the roughness of the internal surface of the bushing and the plunger, were considered as the determined totality of unevenness of contours profile and also stationary random function with zero mean value and known variance.

As a rule, the greatest values of contact pressure depend on the angle of contact and the friction coefficient. The presence of friction forces in the contact zone leads to displacements of the graph contact pressure distribution to the contrary action of the moment.

The numerical calculations were carried out for the bushing of a U8-6MA2 double-stoke slush pump for a velocity of the plunger of 0,2 *m/sec*. As constants, we used the following values of the parameters: $2R = 57mm$, $2R' = 56.7mm$, f=0.2, $E = 1.8 \cdot 10^5 MPa$, $\mu = 0.25$, $V_* = 10^{-6} m$, $\sigma_* = 75MPa$, $\sigma_{cr}/\sigma_* = 2$, $\delta_{cr} = 2.5 \cdot 10^{-6} m$, $K^{(1)} = 2 \cdot 10^{-8}$, $K^{(2)} = 2 \cdot 10^{-9}$, $C_b = 2 \cdot 10^{-7} m/MPa$ (C_b is the effective compliance of the bonds).

Using the solution of the problem to calculate displacements on surfaces prefracture zone:

$$-\frac{1+k_b}{2G}\int_{-l}^{l} g(x_1)dx_1 = v(x_1,0) - iu(x_1,0)$$

Assuming $x_1 = x_0$ applying change of variable, changing the integral by the sum, we find displacement vector on the surfaces prefracture zone for $x_1 = x_0$

$$V_0 = \sqrt{u^2+v^2} = \frac{1+k_b}{2G}\frac{\pi d}{M}\sqrt{A^2+B^2} \tag{29}$$

$$A = \sum_{m=1}^{M}\left(v^0(t_m) + \varepsilon v^1(t_m)\right), \qquad B = \sum_{m=1}^{M_1}\left(u^0(t_m) + \varepsilon u^1(t_m)\right)$$

Here M_1 is the number of nodal points contained in the interval $(-l, x_0)$.

In the place of crack nucleation condition, we accept the criterion of critical opening of surfaces prefracture zone. Considering relation (9) we can write the limit condition in the form

$$C(x_0, \sigma(x_0))\sigma(x_0) = \delta_{cr} \tag{30}$$

The joint solution of the combined algerbraic system and conditions (30) makes it possible to determine the ultimate size of the external load (contact pressure), the size of surfaces prefracture zone for the limiting equilibrium state under which a crack arises under the given characteristics of the crack resistance of the material.

The graphs of the length of the prefracture zone $\lambda = l/R$ for the bushing borehole pump against the dimensionless values of the contact pressure p_0/σ_* are shown in Fig. 2 (Curve 1 refers to the smooth surface, curve 2 refers to the rough surface).

The distributions of the normal force q_{y1}/p_0 in the bonds between the surfaces prefracture zone as a function of the dimensionless coordinate x_1/l are shown in Fig. 3. Curve 1 corresponds to the linear bond and curve 2 to the bilinear bond. The dependence of the critical load p_{cr}/σ_* on the dimensionless length of the prefracture zone is shown in Fig. 4.

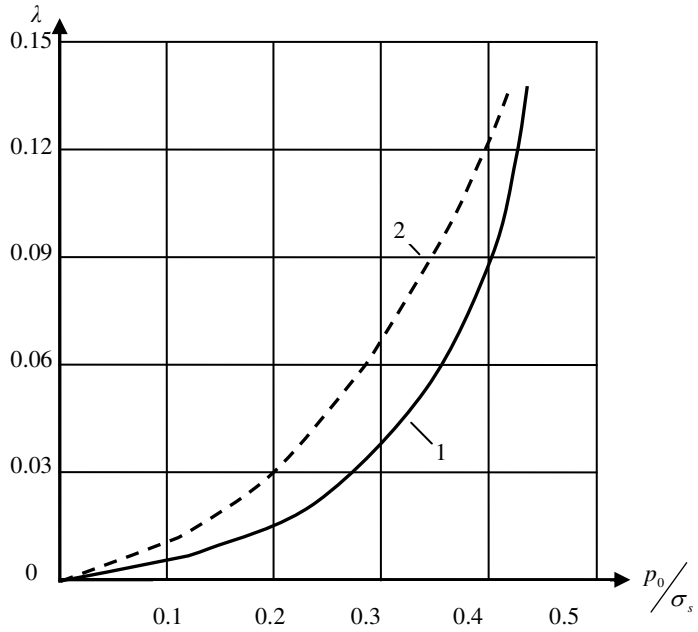


Figure 2
Dependence of length of the prefracture zone $\lambda = l/R$ for the bushing borehole pump on dimensionless contact pressure p_0/σ_*

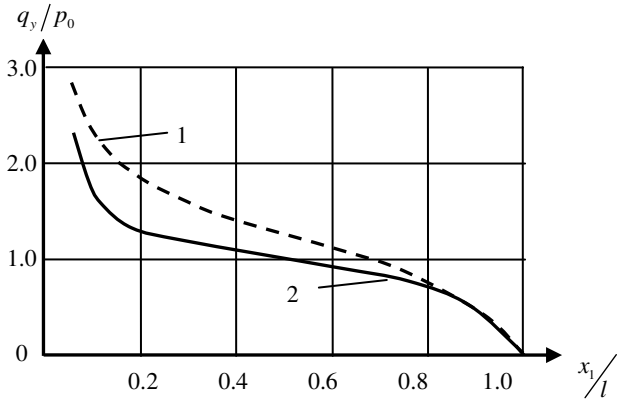


Figure 3
The distributions of the normal force q_{y_1}/p_0 in the bonds between the surfaces prefracture zone as function of the dimensionless coordinate x_1/l

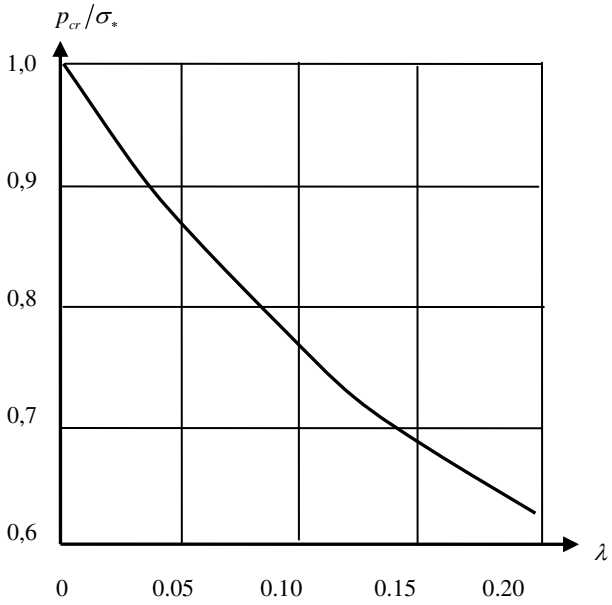


Figure 4
The dependence of the critical load p_{cr}/σ_* on the dimensionless length of the prefracture zone

Conclusions

An analysis of the critical equilibrium state of the bushing contact pair at which the crack appears reduces to a parametric study of the resolving algebraic system (24), (25), (26), etc. and the criterion of crack emergence (30) with different laws of bond deformation, physical and elastic constants of material, and geometric characteristics of the latter. The forces in the bonds and the opening of the prefracture zone are found directly by solving the resultant algebraic systems in each approximation.

An effective algorithm for solving contact fracture mechanics problems on crack nucleation in a bushing friction pair is proposed. This algorithm allows the solution to be constructed in a single manner in each approximation by the method of perturbations.

The model of the prefracture zone with bonds between its faces makes it possible to do the following: to study the basic features of the distribution of forces in the bonds with different deformation laws; to analyze the ultimate equilibrium state of the prefracture zone with allowance for the determination condition of fracture; to estimate the critical external load and crack resistance of the material; and to determine the conditions of equilibrium and growth of the prefracture zone size, as well as conditions of crack nucleation based on the analysis of the ultimate

equilibrium state with allowance for mechanical parameters of the bonds. This model allows us to account for not only each specific realization of the roughness profile (deterministic approach), but also to carry out the statistic description of the roughness of bushing and plunger surfaces by realization of stationary random function. The results of the present work allow us to choose the class of roughness of friction pairs, providing the loading ability of conjugation, optimal in strength and stiffness.

References

[1] Thomas T. R.: 'Rough surfaces' Longman, London, 1982

[2] Aykut Ş.: 'Surface Roughness Prediction in Machining Castamide Material Using ANN', Acta Polytechnica Hungarica, Vol. 8, No. 2, pp. 21-32, 2011

[3] Galin L. A.: 'Contact Problem of Theory of Elasticity and Viscollasticity' Moscow: Nauka (in Russian), 1980

[4] Goryacheva L. G.: 'Contact Mechanics Tribology' Kluwer Acad. Publ. Dordreht, 1998

[5] Budianscky B., Evans A. G., Hutchinson J. W.: 'Fiber-Matrix de Bonding Effects on Cracking in Aligned Fiver Ceramic Composite', Int. J. Solid structures, Vol. 32, No. 3-4, pp. 315-328, 1995

[6] Ji H., de Gennes P. G.: 'Adhesion via Connector Molecules: The Many-Stitch Problem', Macromolecules, Vol. 26, pp. 520-525, 1993

[7] Cox B. N., Marshall D. B.: 'Concepts for Bridged Cracks Fracture and Fatigue', Acta Met. Mater., Vol. 42, No. 2, pp. 341-363, 1994

[8] Muskhelishvili N. I.: 'Some Basic Problems of Mathematical Theory of Elasticity' Amsterdam: Kluwer, 1977

[9] Panasyuk V. V., Savruk M. P. and Datsyshyn A. P.: 'A General Method of Solution of Two-Dimensional Problems in the Theory of Cracks', Eng. Fract. Mech., Vol. 9, No. 2, pp. 481-497, 1977

[10] Mirsalimov V. M.: 'Non-One-Dimensional Elastoplastic Problems' Moscow: Nauka (in Russian), 1987

[11] Il'yushin A. A.: 'Plasticity' Moscow and Leningrad; Gostexhizdat (in Russian), 1948

Pseudo-Isochromatic Plates to Measure Colour Discrimination

Klára Wenzel, Krisztián Samu

Budapest University of Technology and Economics
Department of Mechatronics, Optics and Information Engineering
Bertalan L. u. 4-6, 1111 Budapest, Hungary
E-mail: wenzel@mogi.bme.hu; samuk@mogi.bme.hu

Abstract: We have developed 3 series of pseudoisochromatic plates for colour vision testing. The plates are arranged in order of increasing difficulty. In the first (red/green) series, a red Landolt C is shown in front of a green background. This series is used to determine the severity of colour vision deficiency. In the second series, colours are located on the protan confusion line, whereas in the third series, on the deutan confusion line. The plates were printed by a calibrated colour printer, then bound in a book. The plates were used to test 320 persons with colour vision deficiency and 20 ones with normal colour vision. Our results showed a 96.25% efficiency in separating colour anomals and colour normals as verified by an anomaloscope. The test book gives prompt results and it is fun to use. A test takes about 5 minutes so it is suitable for mass tests and moreover, it may also be used to test the colour vision of children.

Keywords: colour vision deficiency; anomaloscope; ishihara test; D15 test

1 Introduction

1.1 An Optical Explanation of Colour Vision Deficiency

Daytime vision is made possible by the approximately 6.8 million photoreceptors (also known as the cones) found in the retina – the interior part of the eye. Some of the photoreceptors are sensitive to the colour red, others to green, and a third group is sensitive to blue. A person can distinguish between and identify more than a million different colours through the degree of stimulation of the three spectrally sensitive receptor groups. The English terms for the receptors sensitive to red, green and blue colours are: long wave, middle wave and short wave sensitive receptor, or L, M and S, for short. Wave-length determined spectral sensitivity is indicated by $l(\lambda)$, $m(\lambda)$ and $s(\lambda)$ [12], (Fig. 1). In medical literature, these receptors are named protos, deuteros and tritos. Colour vision relying on the

three types of receptor is called trichromatic vision. The most common forms of colour vision deficiency are protanomaly (the anomaly of the L receptor) and deuteranomaly (the anomaly of the M receptor), [12].

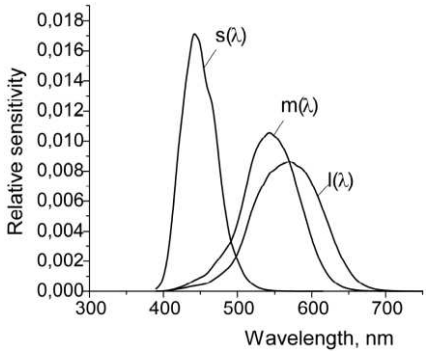


Figure 1

Spectral sensitivity curves of the three daytime receptors of those with normal colour vision, as function of wavelength [Gegenfurtner, Sharpe, 1999]. In the picture l(λ) means the spectral sensitivity of the long wave, m(λ) of the middle wave and s(λ) of the short wave sensitive receptors adapted for white light.

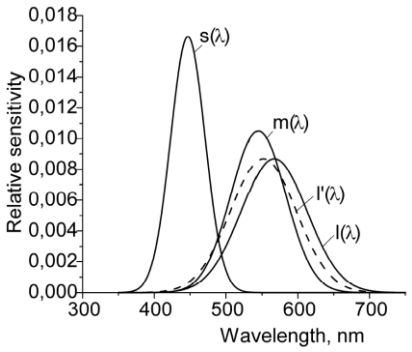


Figure 2

The diagram illustrates the spectral sensitivity disorder in people with protanomaly. The continuous lines show the spectral sensitivity curves of daytime receptors in the case of normal colour vision (l(λ), m(λ) and s(λ)), while the interrupted line shows the spectral sensitivity curve of a person

We developed our colour boosting eyeglasses for these types of colour vision deficiency.

In protanomaly and deuteranomaly, the spectral sensitivity curves of the L and M receptors, respectively are different from those seen in people with normal colour vision. The difference is a shift of the curves along the wavelength axis. The cause of this difference are genetic: different L and M photopigment alleles code for different amino acid sequences, and some differences in the amino acid sequences of the photopigments result in differences in their peak sensitivities.

The diagram in Fig. 2 illustrates the altered spectral sensitivity in people with protanomaly. The continuous lines show the spectral sensitivity curves of daytime receptors in the case of normal colour vision (l(λ), m(λ) and s(λ)), while the dashed lines show the spectral sensitivity curve of a person with protanomaly l'(λ). We can see that protanomaly is caused by the protos spectral sensitivity being shifted toward shorter wavelengths and thus being closer to the sensitivity of the deuteros than happens in subjects with normal colour vision.

Fig. 3 illustrates the spectral sensitivity of the deuteranomalous receptor. In this case, the spectral sensitivity of the deuteros is shifted toward longer wavelengths, and is found closer to the sensitivity of the protos than in subjects with normal colour vision.

The ability to discriminate between hues in the red-green segment of the spectrum is due to the different sensitivities of the L and M pigments. In protanomaly, as well as in deuteranomaly, the spectral distance between the L and M pigments is reduced compared to people with normal colour vision. Therefore, the ability to distinguish between red and green hues is impaired in both cases; this is the reason for sub-normal red-green colour vision. Correspondingly, colour identification is also impaired: the anomalous L and/or the anomalous M pigments are not sufficiently sensitive to red-green differences. Instead, the sensitivity to yellow hues dominates in the middle-to-long end of the spectrum.

The impairment characteristic of protanomaly is shown by the protan confusion line, whereas that of deuteranomalous people by the deutan confusion line in the CIE xyY system (Fig. 4).

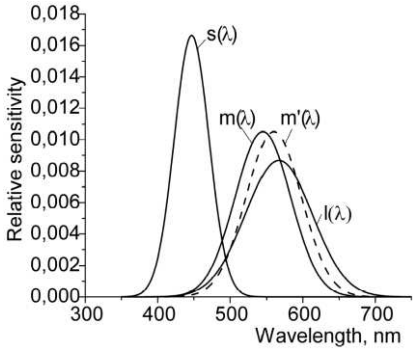


Figure 3
The diagram illustrates the spectral sensitivity disorder in people with deuteranomaly. The continuous lines show the spectral sensitivity curves of daytime receptors in the case of normal colour vision (l(λ), m(λ) and s(λ)), while the interrupted line shows the spectral sensitivity curve of a person with deuteranomaly m'(λ).

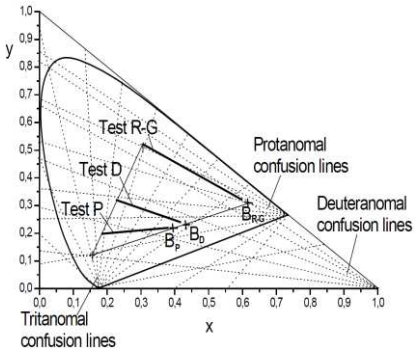


Figure 4
Points of the pseudoisochromatic plates in the CIE xyY chromaticity space. Test $_{R\text{-}G}$, Test $_{P}$, Test $_{D}$ and are the points of the plates in the three (R-G, P and D) series, B R-G, BP and BD are the points of the background in the series.

Colour vision deficiency is tested in most cases by anomaloscopes, the original form of which was constructed by the well-known mathematician Lord Rayleigh, and by different types of pseudo-isochromatic plates.

1.2 Our New Colour Vision Test

Our objective was to develop a colour vision test as simple and effective as the Ishihara test yet as accurate as an anomaloscope, providing quantifiable results. We also aimed at developing a prompt, simple method, also suitable for testing children.

There are some excellent colour vision tests that comply with these criteria [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. However, we wished to develop a test that could be able to distinguish 15 levels according to the severity of colour vision deficiency.

Our test was composed series of pseudoisochromatic plates. These plates do not show numbers or letters but Landolt C-s in various positions. The method was tested on 320 persons with colour vision deficiency and 20 ones with normal colour vision. These test persons had first been tested by the Ishihara test and anomaloscope. The criteria for diagnosis were developed according to the histograms of the measured results. The efficiency of separating patients with colour vision deficiency from those with normal colour vision was 96.25%.

2 Methods

The pseudoisochromatic plates were designed using the principles described in the book of J. Birch [12].

The pseudoisochromatic plates were designed in the colour system of CIE Lab, by a software that utilises confusion lines. The coloured dots on the plates are circular, and their density and sizes are similar to those in the Ishihara images. Both the Landolt-C and the background are composed of dots of 3 shades of the same colour.

The plates are arranged in order of increasing difficulty. In the first (R/G) series, a red Landolt C is shown in front of a green background. This series measures the ability to discriminate green and red colours. In the second (P) series, colours are located on the protan confusion line whereas in the third (D) series, on the deutan confusion line. Plates in the R/G series are coded as 300, 280, 260, etc. down to 60, 40, 30, 20. Plates, and in the P and D series they are coded as 200, 180, 160, etc. down to 60, 40, 30, 20.

In each series, the first plate is readily identifiable; that is, there is a pronounced difference between the average colour of the Landolt-C and the average colour of the background. This difference was determined in accordance with previous experiences, in a way that anomalous trichromats could identify the plate while dichromats could not ($\Delta E^*_{a,b}$=60…80). The other plates in the series are arranged in order of increasing difficulty. While the colour of the background remains the same, the colour of the Landolt-C gradually merges into the background. The last, most difficult plate may be identified only by those with excellent colour vision ($\Delta E^*_{a,b}$=8).

The average lightness of the Landolt C and its background is the same in every image.

The test was initially developed for a colour computer screen [13, 14, 15, 16, 17, 18, 19, 20] and eventually we switched to a printed version [21].

Plates were printed by a Canon iX5000 inkjet colour printer. While printing, the consistency of colour stimuli was provided by means of the ICC Color Management system. Printing as well as using the plates is defined for a CIE D65 standard illuminant.

The examination must be conducted with a standard CIE D65 illuminant. Lighting must be diffuse, neither too dark, nor blindingly bright; the ideal is 400…800 lux. There should not be a blinding light source directed at the subject, and neither must the light be in such a way that a glare could disturb the subjects when observing the images [22, 23].

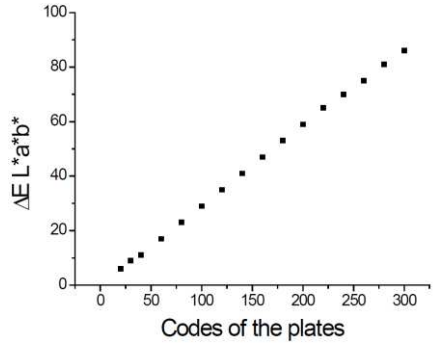


Figure 5
Difference between the average colour of the Landolt-C and the average colour of the background in the series R-G

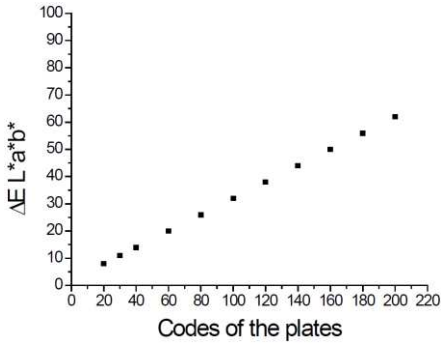


Figure 6
Difference between the average colour of the Landolt-C and the average colour of the

The plates are bound in a book. The book is designed in a way that the test person is able to see one plate at a time only, while the white backside of the next page provides white adaptation for the test person.

Once printed, the colours of plates were verified by Datacolor Microflash 45 (SN: Z151634, White reference: Techkon MF 45.812001). In each plate, 5 ones of the largest dots were measured from the groups of light, medium and dark dots each, then the colours of the background and the Landolt-C were determined from their averages in the CIE xyY system (Fig. 4). As is illustrated by the figure, the colour dots on the plates are near the confusion lines. We also determined the $\Delta E^*_{a,b}$ difference between the colours of the Landolt-C and the background for each plate (Figs. 5, 6 and 7). This difference gradually decreases from plate to plate.

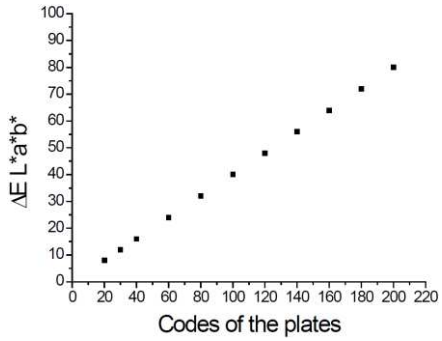


Figure 7
Difference between the average colour of the Landolt-C and the average colour of the

3 Tests

The colour vision of 320 persons with colour vision deficiency and 20 ones with normal colour vision was tested by various methods: anomaloscopy, Ishihara test and the new pseudoisochromatic plates.

The test persons were all males, between 8 and 59 years; the average age of the test persons was 29.67 year. The group of persons with colour vision deficiency was composed of 158 protanomalous and 162 deuteranomalous persons. The test persons cannot be considered as a representative sample of the colour blind as the tests were completed on volunteers.

A standard CIE D65 light source of 600-800 lux was applying for performing the tests.

3.1 Instruments Used

1) Oculus HMC anomaloscope (Typ. 47700, SN 24119901, Germany).

2) Ishihara Tables (ISHIHARA'S TESTS FOR COLOUR DEFICIENCY, 24 Plates Edition, 1999, Kanehara &CO., Tokyo, Japan). The plates were in good condition.

3) Color Vision Test, III. Edition, Printed in 2009.

3.2 Measured Results

A measured result is defined as the code of the first plate the test person was not able to identify in the series of plates of increasing difficulty.

Measured results are first given in the form of histograms. As anomaloscope and the Ishihara test confirmed all the 320 test persons as colour blind, the histograms show the frequency of the results of the new test only.

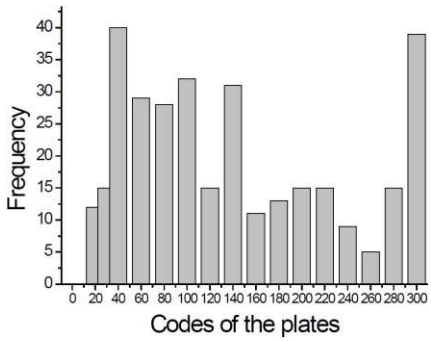


Figure 8
Frequency of the results of colour anomalous people using series R-G

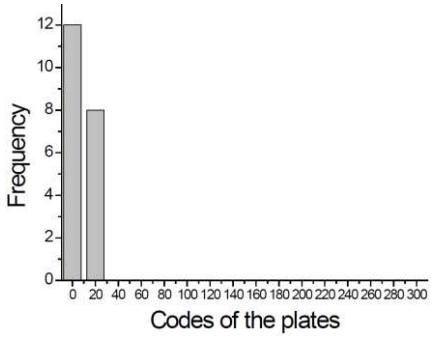


Figure 9
Frequency of the results of people with normal colour vision using series R-G

Fig. 8 shows the results of the R/G series of the test, divided as the measured results for those with colour vision deficiency and Fig. 9 shows the results for control group. On the horizontal axis, the codes of the plates are displayed whereas the vertical axis gives the number of test persons who were not able to identify the orientation of the Landolt-C in the given plate.

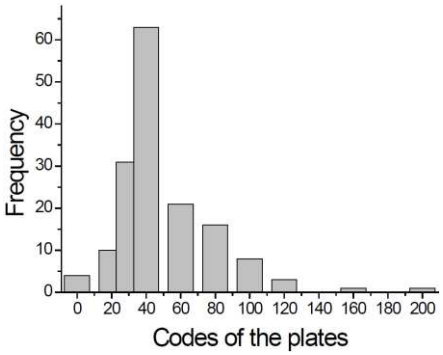


Figure 10
Frequency of the results of protanomalous people using series P

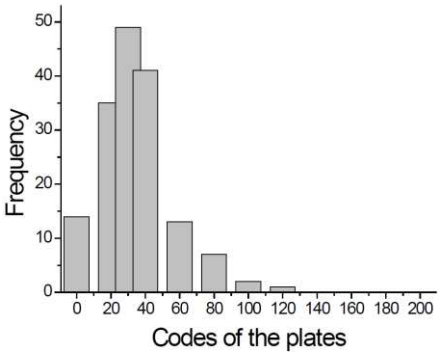


Figure 11
Frequency of the results of deuteranomalous people using series P

Fig. 10 shows the results of the P series of the test for protanomals, Fig. 11 for deuteranomals and Fig. 12 those with a normal colour vision. Protanomals regularly scored lower than deuteranomals in this series.

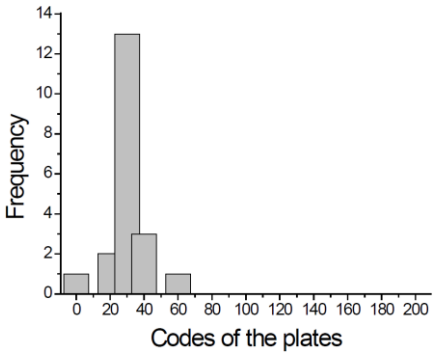


Figure 12
Frequency of the results of people with normal colour vision using series P

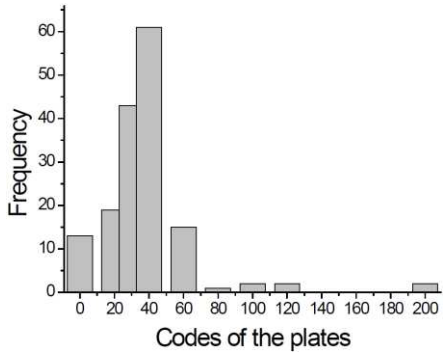


Figure 13
Frequency of the results of protanomalous people using series D

Fig. 13 shows the results of the D series of the test for protanomals, Fig. 14 for deuteranomals and Fig. 15 those with a normal colour vision. Protanomals regularly scored better than deuteranomals in this series.

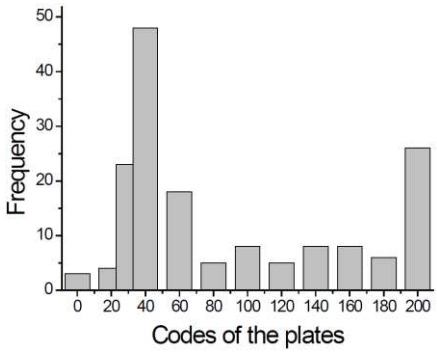


Figure 14
Frequency of the results of deuteranomalous people using series D

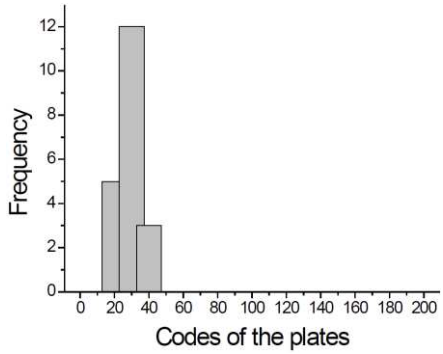


Figure 15
Frequency of the results of people with normal colour vision using series D

3.3 Distinguishing Persons with Colour Vision Deficiency from Those with Normal Colour Vision

Persons with colour vision deficiency are distinguished from those with normal colour vision using the R/G series. Out of the 20 persons with normal colour vision, 12 ones successfully identified all the plates whereas 8 persons could not identify plate 20. Plates with higher code numbers (30, 40, etc.) were readily identified by everybody in the control group.

Accordingly, colour vision deficiency is defined as the inability to identify the plate coded as R/G=30. However, 12 persons out of the 320 test persons with colour vision deficiency were able to identify plate R/G=30, as Fig. 8 clearly illustrates. It means that the results of the test differed from those measured by anomaloscopy in 12 cases; that is the probability of anomaloscopy and the new test yielding the same result is P=((320-12)/320)x100=96.25%.

It should be noted, however, that the ability of discriminate colours deteriorates with age. Our tests were completed on young persons, mostly under 30. Older people or those with impaired vision will not be able to identify plate R/G=40, maybe even plate R/G=60. A possible solution to this problem is to record the codes of the first plates the test person was not able to identify in both the P and D series and consider these values when determining colour vision deficiency using the R/G series. For protanomalous persons, consider the lowest D value, whereas for deuteranomalous ones, the lowest P value.

3.4 Distinguishing Protanomals and Deuteranomals

The recommended criteria for distinguishing protanomals and deuteranomals are as follows:

A person should be considered protanomalous, if P/D>0.9, if not than deuteranomalous.

Our results indicated 124 persons as protanomalous, out of the 158 protanomals as verified by anomloscopy. It means anomaloscopy and the new test yielded the same result in 78.48% of the cases.

A person should be considered deuteranomalous, if P<D during the tests.

Our results indicated 116 persons as protanomalous, out of the 162 protanomals as verified by anomloscopy. It means anomaloscopy and the new test yielded the same result in 71.60% of the cases.

Average of results of the measures is 75%

4 Results

The new R/G series of the pseudoisochromatic plates yielded the same results as anomaloscope in 96.25% of the cases. The test was able to distinguish 16 levels according to the severity of colour vision deficiency.

The test was able to distinguish protanomaly from deuteranomaly in 75% of those cases that were verified by anomaloscope. This difference is probably due to the fact that anomaloscope relies on monochromatic light whereas the new test applies

colours in a broad range, thus the individual variances in the curves of cone-sensitivity of the test person manifest in a different way.

Conclusions

The new test detects colour vision deficiencies with an efficiency of 96.25%, distinguishes protanomaly from deuteranomaly at 75% confidence and is able to distinguish 15 levels according to the severity of colour vision deficiency.

Completing the test takes about 5 minutes while analyzing the results takes only a minute. Thus, the method is suitable for mass tests.

The test is not exhausting; on the contrary, it is fun to use, and moreover the method is suitable to test illiterate children.

Acknowledgement

This work is connected to the scientific program of the " Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project. This project is supported by the New Széchenyi Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

References

[1] City University Online Colour Vision Test, web-based colour vision test

[2] City University Online Colour Vision Test, in print version

[3] Dvorine Pseudoisochromatic Plates, The Psychological Corporation, Harcourt

[4] Farnsworth D15 Test

[5] Hardy, Rand and Rittler: HRR Test, American Optical Co.

[6] Ishihara Tests for Colour Deficiency, Kanehara & Co, Ltd. Tokyo, Japan, 1985

[7] Neitz, J., and Neitz, M.: Neitz Test of Colour Vision, Western Psychological Services, 1988

[8] Rabkin, E. B.: Polichromatitseskie Tablicü dlja issledovania svetooshushenia, Medicina, Moscow, 1971

[9] Velhagen, K. and Broschmann, D: Tafeln und Prüfung des Farbensinnes 12031 Wilshire Blvd, Los Angeles,CA 90025-1251, 2001

[10] Regan BC, Reffin JP & Mollon JD. Luminance Noise and the Rapid Determination of Discrimination Ellipses in Colour Deficiency. Vision Research, 34, 1994; pp. 1279-1299

[11] J D Mollon, J P Refin: A Computer-controlled Colour Vision Test that Combines the Principles of Chibret and of Stilling, R.C.S. MEETING, 1989

[12] Birch J.: Diagnosis of Defective Colour Vision. Brace & Company, San Antonio, 1993

[13] K. Wenzel, I. Kucsera: Chromatic Adaptation Testing with a Computer Graphics System, 10th International Conference of Women Engineers and Scientists, Bp. 1996

[14] Ladunga K., Wenzel K.: New Colour Vision Test on Monitor, XVth ICVS Symposium,Göttingen, 1999

[15] Ladunga K., Kucsera I., Wenzel K.: If I were Colour Blind, CIE Symposium '99, Budapest, 1999

[16] Wenzel K., Ladunga K., Ábrahám Gy., Kovács G., Kucsera I., Samu K.: Measuring Colour Resolution of the Eye by Using Colour Monitors, Colour and Visual Scales Conference, London, 2000

[17] M Piazol, N A Zanc: Medical Tourism - A Case Study for the USA and India, Germany and Hungary, Acta Polytechnica Hungarica Vol. 8, No. 1, 2011

[18] Samu K, Wenzel K, Ladunga K: Colour and Luminance Contrast Sensitivity Function of People with Anomalous Colour Vision, AIC Conference, Rochester, 2001 June 24-29, Proceedings of SPIE Volume 4421

[19] K. Wenzel, K. Ladunga, K. Samu: Measurement of Colour Defective and Normal Colour Vision Subject's Colour and Luminance Contrast Threshold Functions on CRT, Periodica Polytechnica, Vol. 45, No. 1, pp. 103-108, 2001

[20] K Samu, K Wenzel: Presenting Surface Colours on Computer-controlled CRT Displays, Facta Universitatis (NIS), Ser.: Elec. Energ. Vol. 16, 2003, pp. 177-183

[21] Wenzel K, Ladunga K, Samu K, Langer I and Szőke F. Pseudo-Isochromatic Plates for Measuring the Ability to Discriminate Colours. 21st Symposium of the International Colour Vision Society, 2011

[22] Pokorny J, Smith VC, Verriest G & Pinckers AJL. Congenital and Acquired Colour Vision Defects. Grüne & Stratton, New York, 1979

[23] Balázs Vince Nagy, György Ábrahám: Spectral Test Instrument for Colour Vision Measurement, Journal of Bionics Engineering, Vol. 2, pp. 75-79, Issue 2, 2005

Comparison of Plasma and Laser Beam Welding of Steel Sheets Treated by Nitrooxidation

Ivan Michalec, Milan Marônek

Department of Welding
Faculty of Materials Science and Technology in Trnava
J. Bottu 25, 917 24 Trnava, Slovakia
e-mail: ivan.michalec@stuba.sk, milan.maronek@stuba.sk

Abstract: Steel sheets treated by nitrooxidation in comparison to material without surface treatment are characterized by increased mechanical properties and enhanced corrosion resistance. The paper deals with the comparison of solid-state laser beam welding and plasma arc welding in order to reduce the high initial costs of laser beam equipment and to find an adequate counterpart from the arc welding sphere. Results prove solid-state laser beam welding is the most suitable welding method for welding of this type of treated steels, although plasma arc welding, especially after parameters optimizing, can be an adequate alternative to laser beam welding.

Keywords: nitrooxidation; plasma arc welding; laser beam welding

1 Introduction

In view of the positive influence on the steel sheet, a surface treatment is one of the most monitored parts in the industry [1, 2]. The process of nitrooxidation, consisting of a surface nitridation with subsequent oxidation, is a part of non-conventional surface treatment methods, by which a radical mechanical properties increase (Tensile Strength, Yield Strength), together with an increase in the corrosion resistance up to level 10, can be achieved [3]. However, it is not always possible to apply the treatment as the final operation, and materials should be welded after the treatment. In such cases, the knowledge of suitable welding methods that have the least deterioration effect on the surface is essential [1, 2, 4].

In previous outcomes [3, 4, 5, 6], various arc and beam welding methods were examined. In almost every welding method, a high level of porosity together with spatter and weld bead irregularities were observed [7]. Only the joints welded by solid-state laser beam welding were defects-free, and the joints had superior

quality and good consistency [3, 9]. Nevertheless, the high initial cost of the laser equipment turned attention towards an appropriate arc method. The only arc method not tested was plasma arc welding and was supposed as an adequate option to laser beam welding.

2 Materials and Methods Used for Experiments

For all the experiments, low carbon deep drawing steel DC 01 EN 10130/91 of 1 mm in thickness was used. The chemical composition of steel DC 01 is referred to in Table 1.

Table 1
Chemical composition of steel DC 01 EN 10130/91

EN designation	C [%]	Mn [%]	P [%]	S [%]	Si [%]	Al [%]
DC 01 10130/91	**0.10**	**0.45**	**0.03**	**0.03**	**0.01**	**-**

2.1 Thermo-Chemical Treatment

The material was consequently treated by the process of nitrooxidation in a fluidized bed. The nitridation fluid environment was Al_2O_3 with granularity of 120 μm. The fluid environment was wafted by gaseous ammonia. Oxidation was subsequently carried out in the vapours of distilled water. The process parameters are presented in Table 2.

Table 2
Process of nitrooxidation parameters

	Nitridation	Oxidation
Time [min]	**45**	**5**
Temperature [°C]	**580**	**380**

2.2 Methods Used for Experiments

Specimens welded by plasma arc welding (PAW) were made with a Fronius MagicWave 2200 machine with integrated PlasmaModule 10 in Fronius Slovakia, Trnava. Argon with a purity of 99.996 % was used as the shielding gas as well as the plasma gas. The samples were welded in continual and pulse mode, respectively. To provide the constant gap between welded materials, steel sheets were stitched together by GTAW. The fixation of the welded materials during welding is shown in Fig. 1.

Figure 1
Plasma arc welding configuration

Specimens welded by a solid-state laser beam welding (LBW) were welded with a TruDisk 1000 laser machine in PGS Automation, Trnava. The welding parameters are presented in Table 3. The welding process is shown in Fig. 2.

Table 3
TruDisk 1000 laser source characteristics

Source type	TruDisk 1000
Power	1000 W
Optics	ϕ 35 mm
Focal length	200 mm
Focusing plane	surface
Collimation length	100 mm
Spot size	600 µm
Wavelength	1030 nm
Welding speed	20 mm/s
Shielding gas	Argon (10 l/min)
Optical cable	Step Index Φ 300 µm

Figure 2
Solid-state laser beam welding process

Scanning electron microscopy analysis, microhardness measurements, the Erichsen cupping test, and tensile tests were performed in order to obtain the complex information about the properties of the material treated by nitrooxidation.

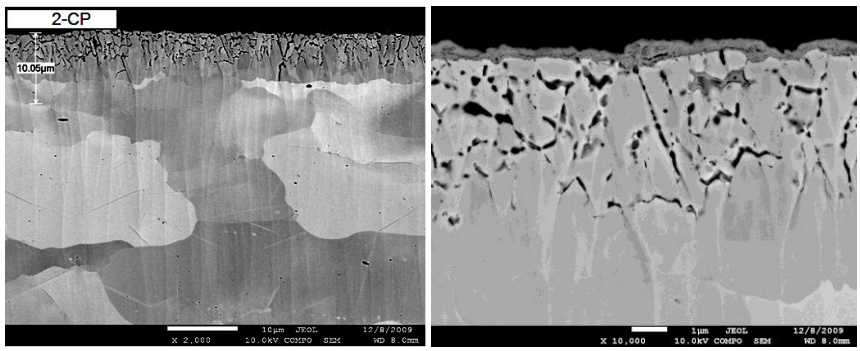
The weld joints were evaluated by macroscopic and microscopic analysis, microhardness measurements, the Erichsen cupping test and tensile tests. The results of both welding methods were compared and assessed.

3 Results

3.1 Material Properties

The key parameters during the nitridation and oxidation phase of the process are temperature and time. The final material properties thus depend on them, and they indirectly effect the welding process stability and weld joint quality. The knowledge of the material properties was therefore seen as essential.

3.1.1 Microscopic Analysis



a) b)

Figure 3

Surface layer of the base material after the nitrooxidation process
a) overall view; b) close up view on the oxide and ε-phase layers

Microscopic analysis was performed with a JEOL 7600-F scanning electron microscope. The microstructure of the base material was ferritic with the dominant orientation of the grains as a consequence of the rolling process. The nitrooxidized surface layer (Fig. 3a) consisted of a very thin oxide layer with thickness up to 700 nm (Fig. 3b). Under the oxide layer, a continuous layer of the ε-phase (Fig. 3b), composed of $Fe_{2\text{-}3}N$, was observed. This layer had a thickness of

approx. 10 μm. Beneath the ε-phase layer, a 60 μm thick compound layer was identified. It consisted of a ferritic matrix and precipitated needle shaped nitrides Fe_4N.

3.1.2 Tensile Tests

The mechanical properties of the material were obtained by tensile tests in accordance to STN EN 10002-1. The average results from five measurements are documented in Table 4.

Table 4
Results obtained by tensile tests

DC 01 EN 10130/91	Yield Strength [MPa]	Tensile Strength [MPa]
Before nitrooxidation	**200**	**270**
After nitrooxidation	**310**	**380**

After the process of nitrooxidation, increases in Yield Strength by 55% and in Tensile Strength by 40% were observed.

3.1.3 Corrosion Test

The corrosion test was carried out in a condensation chamber KB 300 type 43096101 in 100% moisture (environment of distilled water). The testing samples were consequently analysed after 16, 48, 72, 144 and 240 hours and assessed by gravimetric analysis. The results are documented in Table 5.

Table 5
The gravimetric analysis results

Material DC 01 EN 10130/91	Exposure in condensation chamber [h]				
	16	48	72	144	240
	Mass increase [g/m^2]				
Before nitrooxidation	**0.051**	**0.180**	**0.638**	**6.992**	**8.490**
After nitrooxidation	**0.076**	**0.083**	**0.109**	**0.124**	**0.128**

Based on the results, material DC 01 after the process of nitrooxidation was classified as having the maximal (level 10) resistance to atmospheric corrosion. Only 0.128 g/m^2 of mass increase was observed after 240 hours in the condensation chamber. This was more than 66 times less in comparison to the material without nitrooxidation.

3.1.3 Microhardness Measurements

The results of microhardness testing according to Vickers, measured across the material thickness, are documented in Fig. 4. A Buehler IndentaMet 1100 Series tester was used as the measuring equipment. The force load was F = 0. 981 N (m = 100 g) and the loading time was t = 10 s. To acquire the most accurate results, the measurements were repeated three times in different places. Fig. 4 shows an increase in microhardness of more than 47% at the depth of 60 μm from the material's surface. Likewise, it can be stated that the material was affected by the nitrooxidation to the depth of 350 – 400 μm from the surface. However, the microhardness values in the area of ε-phase, where the highest values were expected, could not be obtained due to measuring equipment limitations.

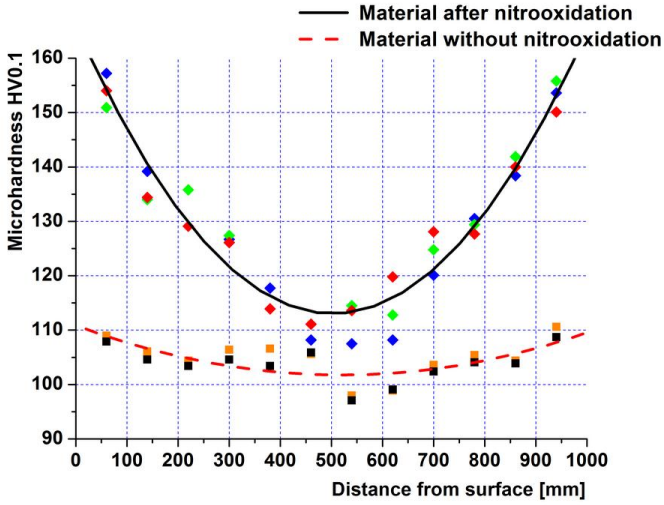


Figure 4
Microhardness trend comparison

3.1.4 Erichsen Cupping Test

An Erichsen cupping test was performed in order to evaluate the forming properties of nitrooxidized material with regard to the influence of the surface layer on deep drawability of the steel. After the process of nitrooxidation, the depth of the indent decreased by 10% in comparison to untreated material. Based on this, the material treated by nitrooxidation loses its deep-drawability, although it is still suitable for forming operations. On the other hand, the character of the fracture corresponded to the untreated material. The result of the Erichsen cupping test is documented in Fig. 5.

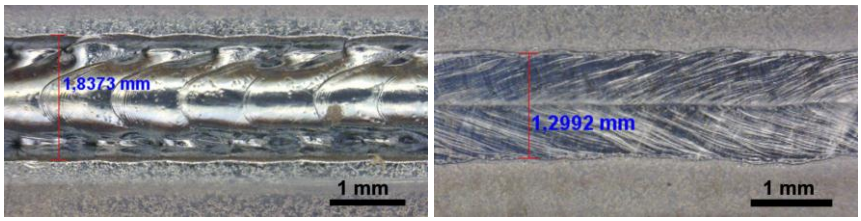
Figure 5
Sample after the Erichsen cupping test with close-up view on transverse fracture

3.2 Joints Properties

There are many factors having a direct and indirect influence on weld joint quality. In order to get the maximum available information concerning the weld joint quality, several analyses were carried out.

3.2.1 Visual Inspection

The visual inspection results of the joints welded by a LBW and PAW are shown in Fig. 6. The PAW joints (Fig. 6a) were about 50% wider than those welded by LBW (Fig. 6b).



a) b)

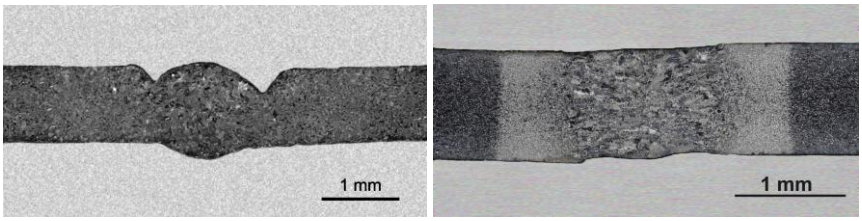
Figure 6
Visual inspection of the joints (face side of the joints)
a) plasma arc welding; b) solid-state laser beam welding

3.2.2 Macroscopic Analysis

The macroscopic analysis (Fig. 7) revealed that the joints welded by PAW (Fig. 7a) had no porosity, which was the main issue in almost every arc welding

method [8]. Nevertheless, the presence of the undercuts, situated along the joint's length, were inappropriate. Figure 7 revealed a more than three times wider Heat Affected Zone (HAZ) than in the joints made by LBW.

The joints welded by LBW welding (Fig. 7b) had a superior quality, very narrow HAZ and were defects-free. The joints had appropriate shape with no bead reinforcement.

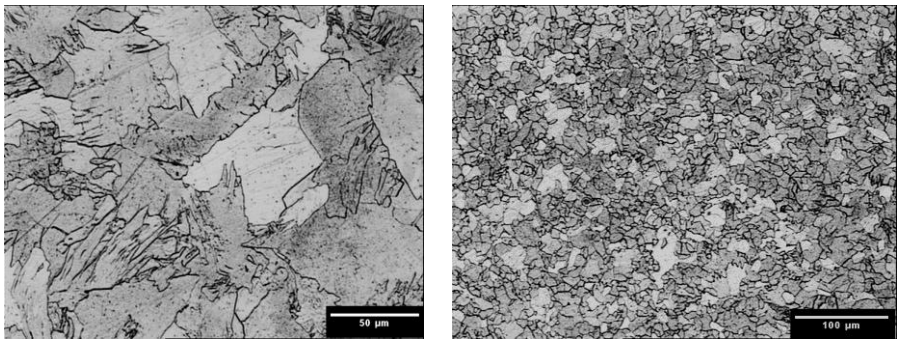


a) b)

Figure 7
Macroscopy of the joints
a) plasma arc welding; b) solid-state laser beam welding

3.2.3 Microscopic Analysis

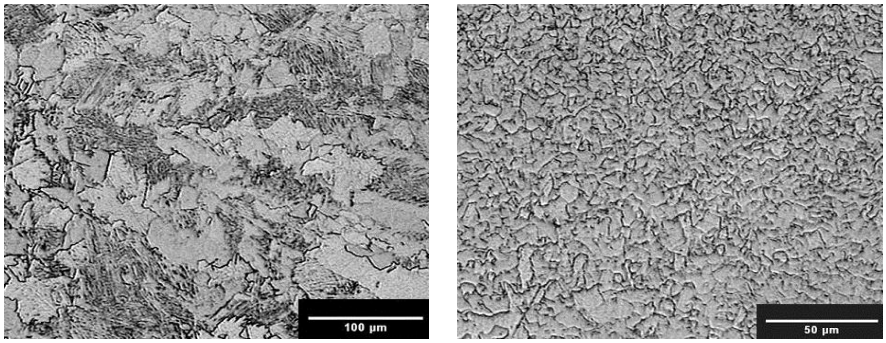
The microscopic analysis was primarily focused on the Weld Metal (WM) and HAZ area. The results of the microscopic analysis of the joints welded by PAW are shown in Fig. 8. The microstructure of WM was ferritic, mainly composed of coarse-grained acicular ferrite. The polygonal ferrite was observed in a minor amount. The HAZ consisted of polygonal ferrite with a visible fine-grained structure.



a) b)

Figure 8
Microscopic analysis of the plasma arc welding joints
a) Weld Metal; b) Heat Affected Zone

The results of microscopic analysis of the joints welded by LBW are presented in Fig. 9. The WM was created primarily by the acicular ferrite and the ferrite precipitated along to columnar grains. The composition of HAZ corresponded to the PAW joints.

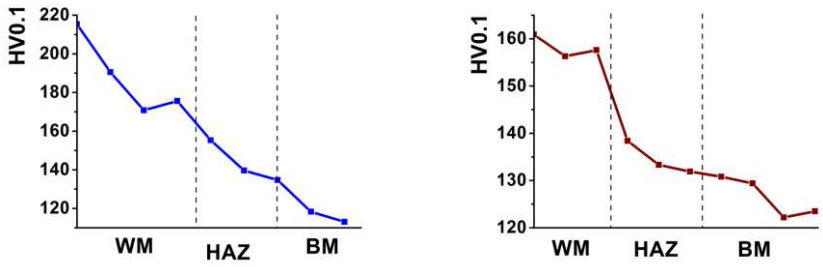


a) b)

Figure 9
Microscopic analysis of the laser beam welding joints
a) Weld Metal; b) Heat Affected Zone

3.2.4 Microhardness Measurements

Microhardness measurements of the joints were carried out in the same way as in the case of the material's examination. Typical microhardness trends are illustrated in Fig 10. The highest microhardness values were observed in WM and the lowest in the BM. The microhardness of HAZ and BM was comparable in both welding methods. The much higher values in the WM area (more than 30%) were obtained in the PAW.

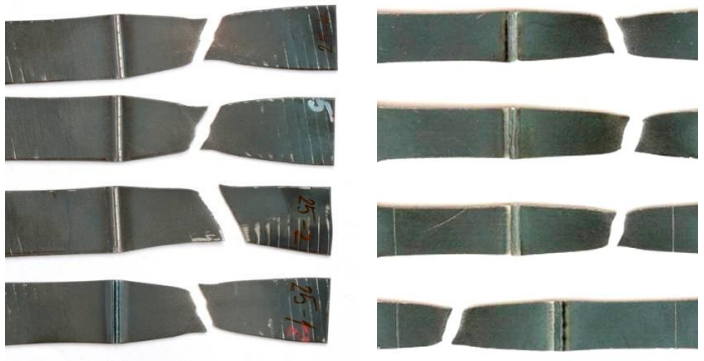


a) b)

Figure 10
Microhardness trend of the joints
a) plasma arc welded; b) solid-state laser beam welded

3.2.5 Tensile Tests

The tensile tests were accomplished on samples with the dimensions of 200×20×1 mm with the weld in the middle of the sample. As the testing device, a tensile test EU 40 machine was used. The results (Fig. 11) showed that both PAW and the LBW joints were fractured outside the joint area and thus marked as suitable.

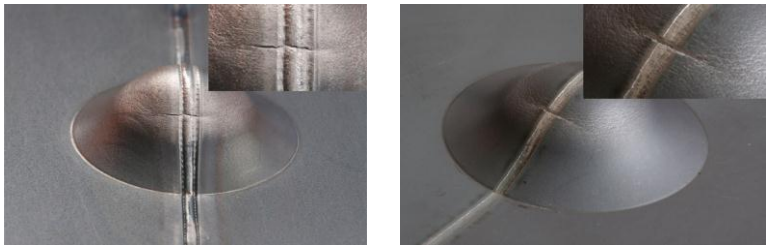


a) b)

Figure 11
Samples after the tensile tests
a) plasma arc welded; b) solid-state laser beam welded

3.2.6 Erichsen Cupping Test

The Erichsen cupping test was carried out in accordance to STN EN 1001-5. The dimension of the samples was 250×50×1 mm. The results of the samples are documented in Fig. 12. Figure 12 shows that in both PAW (Fig. 12a) and LBW joints (Fig. 12b) the transverse type of fracture was observed in every sample, which confirmed the good mechanical properties of the joints, since the fracture did not occur along the joint.



a) b)

Figure 12
Samples after the Erichsen cupping test a with close-up view on the transverse fracture
a) plasma arc welding; b) solid-state laser beam welding

The results of the deep-drawability measurements by the Erichsen cupping test are shown in Table 6. The PAW joints exhibited an increase in the depth indent by more than 11% in comparison to LBW joints.

Table 6
Depths of indent in the Erichsen cupping test

	Plasma arc welding	Solid-state laser beam welding
Specimen No.	Depth of indent	
1	**9.20**	**8.10**
2	**9.10**	**8.10**
3	**8.85**	**8.00**

4 Discussion

The comparison of the joints' widths uncovered a wider joint in PAW, by more than 50%. Likewise, the macroscopic analysis proved a three-times-wider HAZ in PAW joints. This was caused by the higher thermal density of the laser beam distributed into the narrower surface in comparison to plasma arc welding.

The LBW joints had more consistent microhardness trend along the measured length, whereas the PAW joints exhibited a continuous decrease of the microhardness towards the base material. It gave the evidence of microstructure homogeneity across the weld joint, whereas the microstructure of PAW joints was more heterogeneous.

Even though undercuts were observed in the PAW joints, the tensile tests as well as the Erichsen cupping tests proved the good mechanical properties of the joints. Nevertheless, the presence of undercuts indicates that the process parameters need optimizing.

The Erichsen cupping test proved the lower depth of indent in the case of the LBW joints, although the microhardness measurements showed lower microhardness values. These results did not correspond to the expectations of the comparable cupping index and they should be verified in further research.

Conclusion

The material treated after the process of nitrooxidation in the fluidized bed and its welding by solid-state laser beam welding as well as plasma arc welding were analysed. In the surface layer, the individual layers were identified. Likewise, the mechanical properties together with the corrosion resistance increase were documented. The material after nitrooxidation treatment loses its deep-drawability, but on the other hand, it keeps the ability to be formed.

The tests on joints welded by laser proved excellent results. The visual inspection of the joints welded by plasma arc revealed a significant presence of undercuts, whereas the macroscopic analysis confirmed the absence of porosity in the weld joint. Based on the results, there is an assumption that plasma arc welding could have a potential to become an alternative welding method to the laser beam welding of steel sheets treated by nitrooxidation.

Further research activity will be focused on the optimization of the plasma arc welding parameters. The plasma and shielding gas flow rate, the torch angle and the electrode stickout will be taken into account.

Acknowledgements

This paper was prepared within the support of Slovak Research and Development Agency, grant No. 0057-07 and Scientific Grant Agency, grant No. 1/0203/11.

References

[1] Michalec, I.: CMT Technology Exploitation for Welding of Steel Sheets Treated by Nitrooxidation. Diploma thesis, Trnava 2010

[2] Michalec, I., et. al.: Metallurgical Joining of Steel Sheets Treated by Nitrooxidation by a Hybrid CMT - Laser Process, 20th Anniversary International Conference on Metallurgy and Materials, May 18-20, 2011, Brno, Czech Republic

[3] Bárta, J.: Welding of Special Treated Thin Steel Sheets: Dissertation thesis, Trnava, 2010

[4] Marônek, M. et. al.: Laser Beam Welding of Steel Sheets Treated by Nitrooxidation, 61st Annual Assembly and International Conference of the International Institute of Welding, Graz, Austria, 6-11 July 2008

[5] Lazar, R., Marônek, M., Dománková, M.: Low Carbon Steel Sheets Treated by Nitrooxidation Process, Engineering Extra, 2007, No. 4, p. 86

[6] Marônek, M. et. al.: Comparison of Laser and Electron Beam Welding of Steel Sheets Treated by Nitrooxidation, Congresso da ABM (CD-ROM). - ISSN 1516-392X. - 65th ABM international Congress. 18th IFHTSE Congress. 1st TMS/ABM : Brazil, Rio de Janeiro, July 26-30, 2010

[7] Bárta, J. et al.: Joining of Thin Steel Sheets Treated by Nitrooxidation, Proceeding of Lectures of 15th Seminary of ESAB + MTF-STU in the scope of seminars about welding and weldability. Trnava, AlumniPress, 2011, pp. 57-67

[8] Marônek, M. et al.: Welding of Steel Sheets Treated by Nitrooxidation, JOM-16: 16th International Conference on the Joining of Materials & 7th International Conference on Education in Welding ICEW-7, May 10-13, Tisvildeleje, Denmark, ISBN 87-89582-19-5

[9] Viňáš, J.: Quality Evaluation of Laser Welded Sheets for Cars Body. In: Mat/tech automobilového priemyslu: Zborník prác vt-seminára s medzinárodnou účasťou : Košice, 25.11.2005. Košice: TU, 2005. pp. 119-124. ISBN 80-8073-400-3

Non-Linear Behavior of Sands under Longitudinal Resonance Testing

Merouane Mekkakia Maaza, Ahmed Arab, Mostefa Belkhatir, Saaed Hammoudi

Laboratory of Materials Sciences & Environment, University of Chlef (Algeria)
e-mail: mek_mer@yahoo.fr, Ah_arab@yahoo.fr, abelkhatir@yahoo.com, hamoudisaaed@yahoo.fr

Minh Phong Luong

Laboratory Solid Mechanics, CNRS, Polytechnic School of Palaiseau (France)
e-mail: luong@lms.polytechnique.fr

Abdelatif Benaissa

Civil Engineering Dept, University of Science and Technology of Oran (Algeria)
e-mail: dzbenaissa @yahoo.fr

Abstract: One of the fundamental features needed to evaluate soil response during earthquakes, is the study of controllable external variables that may affect the instability phenomena of granular materials under vibration, such as acceleration, frequency, the interaction of grains and their arrangements. Despite previous researches in this field, an understanding of these phenomena is still incomplete. A more accurate description of one of the phenomena that we will see, is how the resonance curve changies and how the jump occurs with the frequency change. For this purpose, a series of longitudinal resonance excitation laboratory tests were carried out on dry sandy soils with different grain size distributions (spread and tight) and different densities to identify the instability zone.This type of test may be assimilated to a system subjected to a forced excitation with damping. The test results confirm the existence of a non-linearity zone represented by a "jump" just after the resonance for tight-grained sand. Moreover, this study shows that the grains interact with the contact forces. Indeed, a slight local density increase induces more collisions and friction, and therefore more dissipation, creating a pressure drop that attracts the neighboring particles and finally a low damping.

Keywords: sand; resonance; dynamic; vibration; non-linearity; frequency; velocity

1 Introduction

The topic we develop in this laboratory investigation concerns the problem of earthquake hazards. Indeed, the earthquake appears on the ground surface in terms of soil vibrations, which may induce phenomena whose consequences can be devastating for both human and socio-economic field. The soil is an assembly of grains and particles much more complex than the regular assembly of the spheres used in the linear elastic theory of Hertz. However, the dynamic study of such soil assembly provides us with very basic information on these phenomena. In addition, direct contact between the grains plays an important role when the soil deposit starts moving.

Granular materials have been the subject over decades of a significant number of previous research works. According to Jae (1996) and Mue (1998), granular assemblies are random arrangements of rubbing grains with a geometrical disorder. Granular soil deposits are characterized by a non-linear response, making their overall behavior surprising and complex (Evesq 2002). The importance of this non-linear soil behavior is commonly accepted in the earthquake engineering field (Lopez Caballero 2003)**.** And according to (Roscoe and Burland 1968, Hicher 1985 and 1996, Biarez and Hicher 1994, Maalej et al 2007), this has already been demonstrated earlier by different studies; the mechanical behaviour of sands in the range of small strains ($\varepsilon \approx 10^{-5}$) revealed a non-linear elastic behaviour, which depends on the evolution of the modulus of elasticity. Non- linear behaviour can be characterized by rigidity and the degree of non-linearity (Atkinson 2000). Miksic (2008) showed that the complexity of granular soil is strongly linked to the inherent disorder of these deposits, due to the heterogeneity of the contact forces between the grains.

For this purpose, we propose to study the nonlinear phenomenon whose effects are often considered disturbing, leading to spectacular effects. This analysis leads logically to the examination and description of the dynamic properties of the material in terms of transfer curves. For this, many researchers have undertaken studies on the linear and non-linear response of a vibration on solids. Dublin (1959) found that the shape of the curve acceleration response-frequency (resonance curve) depends on the amplitude and shape of excitement. Harris and Crede (1961) studied the phenomenon of non-linear resonance curves. It was shown that two types of jumps appear: one on the right of the peak and the other to the left of the peak. These jumps represent the region of instability of the system, and the position of the jumps depends on the direction of frequency (decreasing or increasing the frequency).

The studies of Anand (1966) on non-linearity show areas of instability on the resonance curve, and show that the excitation forces play an important role in the presence or absence of jumps. The instability zone is primarily due to the frictional forces, which results in a jump which can behave like a linear system by introducing dissipative elements (an increase in the damping) (Mathey R and

Rocard Y 1963). Valette and Cuesta (1993) carried out a vibration study on a vibrating cord; they found that the resonance frequency increases with the excitation force and the non-linearity, but decreases when the damping coefficient increases. According to Lalanne (1999), the jump of the resonance curve is unstable and therefore cannot represent the transfer function of a physical system. The sources of non-linearity particularly are: large displacements, dry friction and the non-linearity of the material (Girard and Roy 2003).

Despite all of these works, the problem of non-linearity of materials is still poorly understood; fundamental aspects remain to be solved, including not only the underlying causes of the non-linear phenomenon, but also the effects that occur**.** The tests with the sinusoidal forced vibrations with damping make it possible to clarify certain points, in particular the curve shape, the amplitude versus the frequency, and resonance frequencies, as well as the instability phenomena that take place when the excitation frequency is varied (Perez J. P 1995). It is therefore necessary to supplement existing research with specific investigations to identify clearly the phenomenon. For this purpose, we carried out an experimental program on two sands with different grain size distributions and with different initial densities to highlight another fundamental aspect of the non-linear behaviour of the soil. This aspect is studied through the influence of some factors affecting the shape of the resonance curve and the resonance frequency. We subject the sand specimen to vibrations with an acceleration range of excitation ($0.25g \leq \Gamma \leq 1g$) and a variable frequency sweeping velocity (V = 0.2 Hz/s and V = 0.1 Hz/s). We consider the existence of a confining pressure within the sample created by the vacuum (σ_c) equal to 100 kPa, using an electromagnetic vibrator called "vibrant pot".

2 Tested Materials

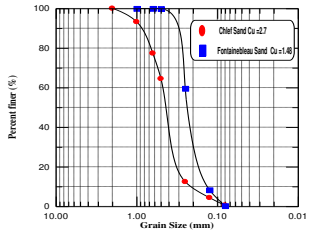
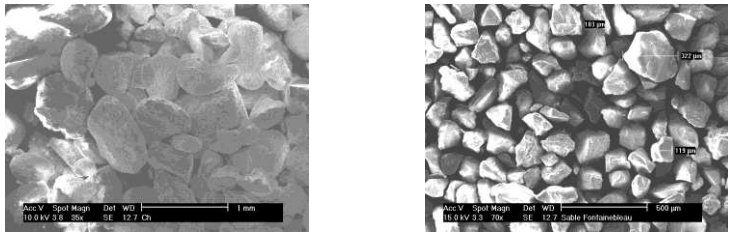


Figure 1

Grain size distribution curves of tested materials

The tested materials are granular materials. We used two different sands (Chlef and Fontainebleau). Figure 1 shows the grain size distributions of those two sands and Figure 2 illustrates the images taken by the scanning electronic microscope

(SEM), identifying clearly the texture of the studied sands. Table 1 presents their physical properties.



(a) (b)

Figure 2

Image SEM: (a) Chlef sand and , (b) Fontainebleau sand

Table 1

Physical properties of tested sands

Material	γ_s (g/cm^3)	e_{max}	e_{min}	Grain shape
Chlef Sand	2.68	0.85	0.53	Rounded
Fontainebleau Sand	2.63	0.94	0.54	Mostly angular

3 Experiment and Materials Studied

3.1 Equipment

The apparatus used is an electromagnetic vibrator (vibrant pot) controlled by an electronic control unit of power. The capability of the vibrator (TW DERRITRON 3000) is 5 kN dynamic, with a range of sinusoidal frequencies between 20 and 10 kHz and a rack steering frequency, acceleration or speed imposed. A computer records and processes the data with suitable software developed at the Laboratory of Mechanics of Solids (Figure 3). Four sensors are connected to amplifiers and signal conditioners: one to measure the acceleration at the top of the sample (attached to the mass), a second for the acceleration at the bottom of the sample, a third to measure the force that is applied to the sample (under the brass plate), and a fourth connected to the pot to control the imposed acceleration. The mass is placed on the top of the sample that weighs 2850 g (Figure 4). The hammer is used to compact the samples to vary the density.

Figure 3
The electromagnetic vibrator connected to the electronic control unit

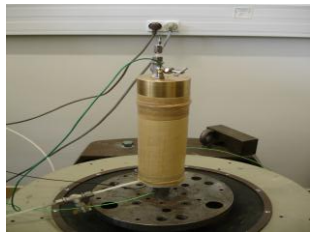


Figure 4
The Specimen placed on the vibrant pot

3.2 Preparation of the Samples and Procedure

The samples of dry sand that we used for a series of tests were prepared in a metal mold of 70 mm in diameter and 160 mm in height, on which was placed a rubber membrane. Pressed against the inner surface of the mold was a vacuum. We applied an air depression created by a vacuum pump (-100 kPa) through the opening, then we poured the sand into the mold in five layers of 200 g, and each layer was compacted. The various densities of the samples varied according to the number of compaction blows. Once the sample was placed at the initial density, a mass was placed at the top of the sample, applying a vacuum within the sample to allow its manipulation and to put it in an upright position (in equilibrium) on the pot. This was a very important point, because the pressure inside the sample was considered as a given pressure confined by compressed air σ_C (Figure 5). After reaching the desired vacuum, the tap was closed, and the mold and supports were removed. Finally, we had our sample perpendicular and in equilibrium (see Figure 4).

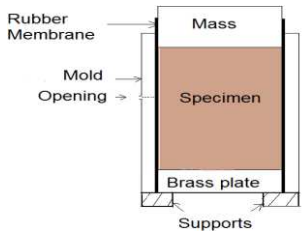


Figure 5
Sample preparation mold

3.3 Testing Procedure under Longitudinal Resonant Excitation

This test has been used by many authors including Hardin and Richart (1963) and Saada and Al (1978) together with the torsional resonance instrument to measure the velocity of longitudinal waves at the resonance of a cylindrical sample. Boelle (1983) developed this test to measure the Young modulus and Poisson's ratio at small strains ($\varepsilon \approx 10^{-5}$) (El Hosri 1984).

Our soil specimen to be tested was encased by a rubber membrane. It was then placed on a base attached to the oscillating diaphragm by a brass plate. A mass was placed on the top of the specimen, which was then placed under a vacuum, considered as a confining pressure. At the beginning of the test we applied a vibration with a frequency scanning varying from 300 Hz to 30 Hz (in decreasing the frequency), while the velocity and acceleration were imposed (see Figure 5). We could assimilate our sample to an oscillatory system by a single mass supported by a spring and damper (viscoelastic model). The support received an excitation, which is defined by an acceleration known (Γ); the excitement spread towards the mass through the elements K and C. The vibration that supports the mass translates into a response movement (Figure 6).

The program “pot" recorded time, scanning, acceleration at the top, acceleration at the bottom, and the dynamic force. It also displayed the amplitude and the signal phase of the acceleration at the top, taking into account the acceleration at the bottom, and the transfer curve.

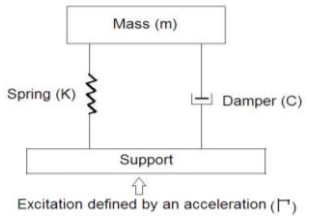


Figure 6
Schematic of a system-mass-spring-damper

4 Analysis of Experimental Results

4.1 The Influence of Particle Size on the Resonance Curve

Many authors have studied the influence of grain shape, grain size and mineralogy of the materials on mechanical properties at the resonant longitudinal column in the range of small deformations (Constantino 1988). Skoglund, Macurson and Cunny (1976) have studied the influence of soil structure on the modulus of the resonant column using two types of soil, sand and silt clay (Homsi, 1986). According to Luong (1986), we can see that by its morphology, a sandy soil acts as a filter frequency Low-pass.

The experimental program conducted in the laboratory involved a series of tests of longitudinal resonance on the material described above for various conditions of initial density and acceleration for a range of excitation ($0.25 \leq \Gamma \leq 1$g), with a speed and a confining pressure imposed. This allowed us to investigate on the shape and evolution of the resonance curve, and how the jump occurs according to the excitation for a given size. Figures 7 and 8 illustrate the qualitative aspect of the jump phenomenon that appears after the resonance in the form of a straight line, which explains: for a frequency value there is an infinity of responses.

Figure 7 shows the evolution of acceleration versus frequency. We note that different curves show a linear shape around the resonance for the sand well graded (diversified grain sizes) and that the frequencies of its resonance change gradually as the acceleration of excitation increases. We can then say that the Chlef sand (C_u >2) can be characterized by its linearity during the resonance; we can say that it is easy to identify the dynamic parameters at each point of response curves. Figure 8 shows curves with a nonlinear form around the resonance represented by a jump, which varies according to the acceleration of excitement. We deduce that the uniformly-graded, tight-grained sand subjected to vibratory loading presents a nonlinear resonance curve resulting in a jump, and the calculations are very difficult to perform, especially for bandwidth.

However, the analysis of the behavior of granular soils at resonance (Figures 7 and 8) shows that the sand grains is a very complex assembly, and provides information that can be summarized in the following:

This linear behaviour for the graded sand can be explained by the fact that the agitated grains of sand settle down and form regular geometrical figures, although they are highly agitated under the effect of vibrations with maximum amplitude (resonance). The rearrangement of grains (slips and rotation) is fast, since the various sizes of grains and their round shape (see paragraph 2) facilitates the movement of these and come together during the vibration motion corresponding to low amplitude. The test results show that the graded sand follows a linear variation before and after resonance. This is shown in Figure 7.

However, there is a non-linearity around the resonance for the uniform graded sand (C_u< 2). It is more significant when the acceleration of the excitation is important (Γ> 1g). The results of the tests carried out on the Fontainebleau sand show unusual shapes of the resonance curves with the appearance of a decreasing jump just after the resonance (Figure 8) This phenomenon is explained by the variation of the acceleration inside the sample (large amplitudes of Γ) during the resonance, in addition to the direct contact between the particles, which plays an important role when the soil specimen is put in motion. When the amplitude becomes maximum (resonance) and the time is very short, the rearrangement of the grains occurs with difficulty due to their size (uniformity) and shape (angularity). The jump that appears represents the nonlinear behavior, and for a single frequency value corresponds to several values of acceleration (undefined number).

Thus, we can say that the factors leading to the non-linearity are: grain size distribution, their arrangements, and dry friction around the resonance.

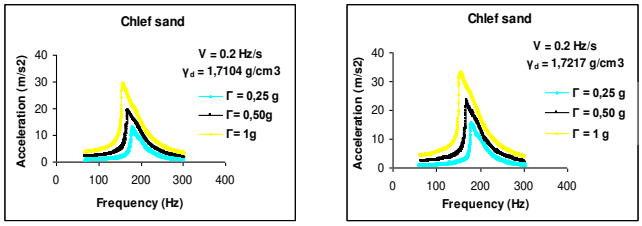


Figure 7

Resonance curves (σ_c =100 kPa)

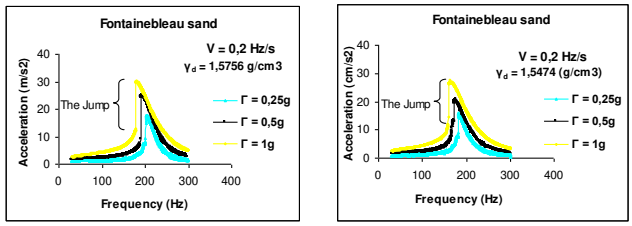


Figure 8

Resonance curves with an instability zone (σ_c =100 kPa)

4.2 Effect of the Velocity on the Nonlinearity

The appropriate choice of scanning mode requires a scan rate slow enough so that the response reaches a large percentage of the steady-state response. Indeed, an extremely slow scanning allows for measuring and plotting the transfer function of

the system with a degree of freedom without distortion and for obtaining values of the resonance frequency and voltage (Lalanne, 1999).

To show the effect of velocity on the nonlinearity, we carried out tests on longitudinal resonance excitation on two different dry sands, Chlef sand (well graded) and Fontainebleau sand (poorly graded), under different densities and excitation speeds equal to 0.1 Hz/s and 0.2 Hz/s.

Curves 9 and 10 show that the sand of Fontainebleau presents an instability zone despite a decrease in the velocity from 0.2 to 0.1 Hz/s, but the jump varies in size according to the velocity: as agitation decreases, the jump decreases. Meanwhile, the Chlef sand maintains its linearity; then we can say that in this case the nonlinearity is independent of the velocity.

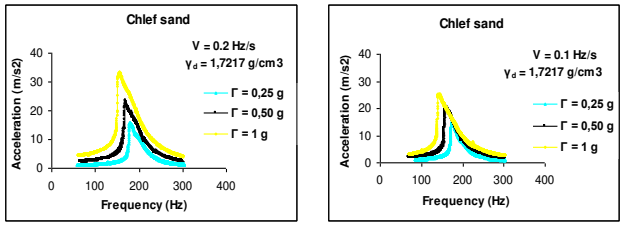


Figure 9

Resonance curves (σ_c =100 kPa)

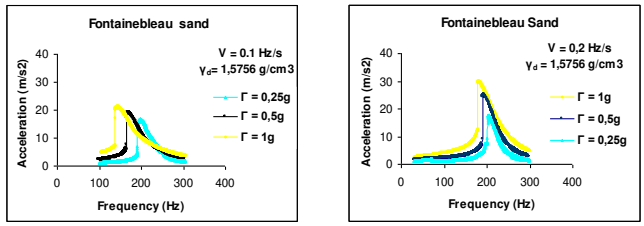


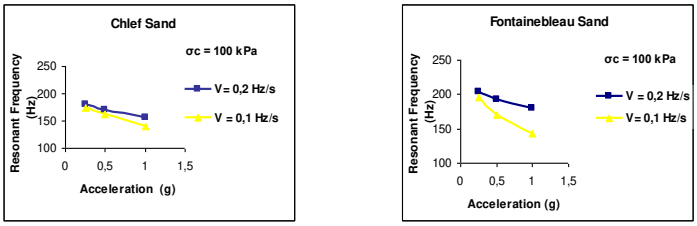
Figure 10

Resonance curves (σ_c =100kPa)

4.3 Effect of the Velocity on the Frequency of the Resonance

Our results (Figures 9 and 10) show that when the scanning speed is slow, the response curves of the acceleration-frequency (resonance curves) show a reduction in the maximum peak acceleration, a shift of the abscissa of the maximum, and a displacement of the center line of the curve (which loses its symmetry); as we remark that an increase of the bandwidth.

Finally, we can say according to the Figure 11 that the resonance frequency increases with the scan rate for all types of sand, and that the acceleration measured at the top of the sample depends on the scan velocity rate.



(a) (b)

Figure 11

Effect of propagation velocity on the resonant frequency

Conclusions

Our study has highlighted the nonlinear behavior of uniform sand (Cu <2) under vibration loading. This nonlinearity appears around the resonance, wich means that this step in the rearrangement of the grains is difficult because of the sudden agitation, uniform particle size and the angular shape of the grains. This is in contrast to the well graded (Cu >2) sand, where the diversity and rounded shape of the grains facilitate their quick rearrangement during the resonance phenomenon, which guaranteed the non-appearance of the instability zone (jump). This type of nonlinearity appears on the resonance curves in a particular manner that is represented by a jump (a zone of instability). Our work has also led us to define the state of resonance frequency linear variation depending on the speed excitement and the resonance curve asymmetry. However, the variation of the scanning velocity has no effect on the nonlinearity of the material.

Acknowledgement

The authors thank the entire team of the Laboratory of Solid Mechanics of Polytechnic School of Palaiseau and in particular the Director of the Lab, Professor B. HALPHEN, for their contribution to this work.

Notation

g (m/s^2): Gravitation

γ_s (g/cm^3): Solid unit weight

γ_d (g/cm^3): Dry unit weight

e_{max} : Maximum void ratio

e_{min}: Minimum void ratio

C_u : uniformity coefficient (Hazen coefficient)

C: Damper coefficient

K: spring constant

Γ : Acceleration

V: Velocity

σ_c : Confining pressure

References

[1] Anand G. V: Nonlinear Resonance in Streched Strings with Viscous Damping, Journal of the Acoustical Society of America 40, 1966, pp. 1517-1528

[2] Atkinson J. H: Nonlinear Soil Stiffness in Routine Design, Géotechnique 50, N°. 5, 2000, pp. 487-508

[3] Biarez J., Hicher P. Y: Elementary Mechanics of Soil Behaviour, Belkema, Rotterdam, The Netherlands, 1994

[4] Boelle J. L: Mesure en Regime Dynamique des Propriétés Mécaniques des Sols aux Faibles Deformations, Thèse de Doctorat, Ecole Centrale de Paris, 1983

[5] Caballero F: Influence du Comportement Non Linéaire du Sol sur les Mouvements Sismiques Induits dans des Géo-Structures, Thèse de Doctorat, Ecole Centrale, Paris, 2003

[6] Constantino R. R: Détermination des Propriétés Mécaniques et des Argiles en Régime Dynamique et Cycliques aux Faibles Déformations, Thèse de Doctorat, Ecole Centrale de Paris, 1988

[7] Dublin M: The Nature of the Vibration Testing Problem, Bulletin n°27, Shock, Vibration and Associated Environments, 1959, pp. 1-6

[8] El Hosri M. S: Contribution à L'étude des Propriétés Mécaniques des Matériaux, Thèse de Doctorat, Université Pierre et Marie Curie, Paris 6, 1984

[9] Evesq P:. Quelques Aspects de la Dynamique des Milieux Granulaires , Poudres & Grains 13 (4), 2002, pp. 40-73

[10] Girard A, Roy N: Dynamiques des Structures Industrielles, Hermes Science, Paris, 2003, pp. 325- 347

[11] Hardin B. O, Richart F. E:. Elastic Wave Velocities in Granular Soils, J. SMFS, ASCE, Vol. 89, SM 1, 1963, pp. 33-65

[12] Harris C. M, Crede C. E: Shock and Vibration handbook, Mc Graw-Hill Book Campany, New York, 1961

[13] Hicher P. Y: Comportement des Argiles Saturées sur Divers Chemins de Sollicitations Monotones et Cycliques, Application à une Modélisation Elastoplastique et Viscoplastique, Thèse d'état, Université Paris 6, France, 1985

[14] Hicher P. Y: Elastic Properties of Soils, Journal of Geotechnical Engineering, Vol. 122, N°. 8, 1996, pp. 641-648

[15] Homsi M: Contribution à L'étude des Propriétés Mécaniques des Sols en Petites Déformations à L'essai Triaxial, Thèse de Doctorat, Ecole Centrale de Paris, 1986

[16] Jae et al: Granular Solids, Liquids and Gases, Reviews of Modern Physics, Vol. 68, N° 4, 1996, pp. 1259-1273

[17] Lalanne C: Vibrations Sinusoïdale. Hermes Sience Publications, TOME 1, Paris, 1999, pp. 277-279

[18] Luong M. P: Mesure des Propriétés Dynamiques des Sols. Revue Française de Géotechnique, 34, Paris, 1986, pp. 18-28

[19] Maalej Y, Dormieux L, Sanahuja J : Comportement Elastique Non - Linéaire d'un Milieu Granulaire: Approche Micromécanique, Comptes Rendus Mécanique, Vol. N° 335, Issue 8, August 2007, pp. 461-466

[20] Mathey R, Rocard Y : Physique des Vibrations Mécaniques. Dunod, Paris. France, 1963, pp. 146-154

[21] Miksis A: Etude des Propriétés Mécaniques et Acoustiques d'un Milieu Granulaire sous Chargements Cycliques, Thèse de Doctorat, Université – Paris-Est, France, 2008

[22] Mue et al: Force Distribution in a Granular Medium, Physical Review E-Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Ttopics, Vol. 57, Issue 3, Suppl. B, 1998, pp. 3164-3169

[23] Perez J. P: Mécanique Points Matériels, Solides Fluides, Masson, 4ème Edition, 1995, pp. 142-153

[24] Roscoe K. H, Burland I. B: On the Generalised Stress-Strain Behaviour of Wet Clay, Engineering Plasticity, Cambridge, U.K., 1968, pp. 535-609

[25] Saada A. S, Bianchini G. F, Shook, L. P:The Dynamic Response of Normally Consolidated Anisotropic Clay. Proceedings of the Specialty Conference on Earthquake Engineering and Soil Dynamics, ASCE, June, Pasadena, California, Vol. II, 1978, pp. 777-801

[26] Skoglund G. R, Marcuson W. F, Cunny R. W: Evaluation of Resonant Column Test Devices. J. GED., ASCE, Vol. 102, GT 11, 1976, pp. 1147-1158

[27] Valette C, Cuesta C: Mécanique de la Corde Vvibrante, Maison Hermes, Paris, 1993, pp.135-158

On the Calculus of Centrifugal Moments for Plane Plates and Plane Bars

Mihail Boiangiu

Department of Mechanics, "Politehnica" University of Bucharest, Splaiul Independentei 313, sector 6, cod 060042, Bucharest, Romania, E-mail: mboiangiu@gmail.com

Georgeta Boiangiu

Math. Prof., master student, "Politehnica" University of Bucharest, Romania

Abstract: This paper is focused on the calculus of centrifugal moments for plane plates and bars, starting from the definition. General cases of plane plates and bars are studied. General formulae of calculus for centrifugal moments are established. These formulae are based on the positions of the mass centers of the rotation surfaces and rotation bodies generated by the bars and plates in rotation, respectively.

Keywords: centrifugal moments; plane plate; plane bar

1 Introduction

A number of problems on the dynamics of rigid bodies [1] are solved by the application of the theorem of the angular momentum, or d'Alembert's principle [2], [3], [4], [5].

In order to solve the problems on the dynamics of plates and bars by using this theorem, it is necessary to find the centrifugal moments by a calculus, which can sometimes be difficult. In the technical literature [4], [5] this is done by integration, starting from the definition.

In this paper, the authors propose two general formulae for the calculus of the centrifugal moments for plane plates and bars. The formulae proposed are original and are based on the positions of the mass centers of the rotation bodies and rotation surfaces generated by plates and bars in rotation, respectively.

2 Centrifugal Moments for Plates

Let us consider a homogeneous plane plate with the mass *m*, area *A* and surface density ρ. We relate the plate to a Cartesian reference system (Figure 1a) so that the plate will be situated in the *xOz* plane. The Ox and Oz axes do not cut the plate. The center of mass, C, has the coordinates $C(\xi,0,\zeta)$.

We isolate an element with infinite little area $dA = dxdz$, with the mass dm. Starting from the definition of the centrifugal moment, we can write:

$$J_{xz} = \int_{(D)} xzdm = \rho \iint_{(S)} xzdA = \rho \iint_{(S)} xzdxdz = \frac{\rho}{2\pi} \iint_{(S)} z2\pi xdxdz = \frac{\rho}{2\pi} \iiint_{(V)} zdV , \qquad (1)$$

where $dV = 2\pi xdA = 2\pi xdxdz$ is the volume of an infinite little element generated by the plate in rotation around the Oz axis (Figure 1b).

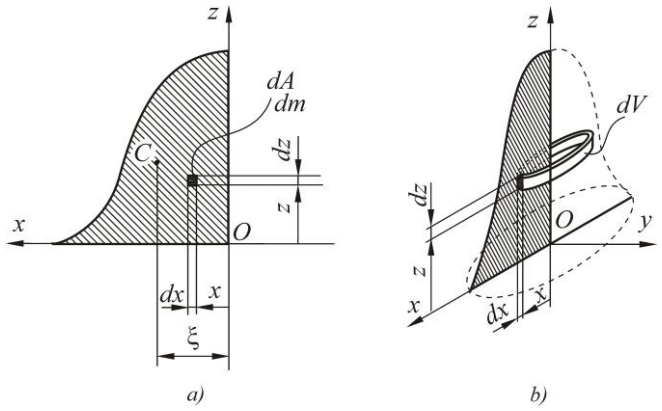


Figure 1
Plane plate completely situated on the same part of the rotation axis (Oz): a) calculus of centrifugal moment, starting from definition; b) rotation body generated by plate in rotation

If we consider that the mass center coordinate z_{rc} of the rotation body generated by the plate rotating around the Oz axis is given by the relation:

$$z_{rc} = \frac{\iiint_{(V)} z\,dV}{\iiint_{(V)} dV}, \qquad (2)$$

the relation (1) becomes:

$$J_{xz} = \frac{\rho}{2\pi} \iiint_{(V)} zdV = \frac{\rho}{2\pi} z_{rc} \iiint_{(V)} dV = \frac{\rho}{2\pi} z_{rc} \iint_{(S)} 2\pi xdA . \qquad (3)$$

Taking into account that the mass center coordinate ξ of the plane plate on the axis which is perpendicular on the rotation axis Oz (used to generate the rotation body) is given by the relation:

$$\xi = \frac{\iint\limits_{(S)} x dA}{\iint\limits_{(S)} dA}, \tag{4}$$

the relation (3) becomes:

$$J_{xz} = \frac{\rho}{2\pi} z_{rc} \iint\limits_{(S)} 2\pi x dA = \rho z_{rc} \iint\limits_{(S)} x dA = \rho z_{rc} \xi \iint\limits_{(S)} dA = \rho z_{rc} \xi A = m z_{rc} \xi, \tag{5}$$

where $m = \rho A$ represents the mass of the plate.

So, we obtain the following formula for the centrifugal moment:

$$J_{xz} = m \xi z_{rc}, \tag{6}$$

or

$$J_{xz} = \rho A \xi z_{rc}. \tag{7}$$

In conclusion, the centrifugal moment is equal to the product of the mass of the plate, the mass center coordinate of the rotation body generated by the plate, and the mass center coordinate of the plate on the axis which is perpendicular on the rotation axis.

If we consider the second Guldin's law, $V_{Oz} = 2\pi \xi A$ (where V_{Oz} is the volume of the rotation body generated by the plate in rotation around the Oz axis), the relation (7) becomes:

$$J_{xz} = \frac{\rho z_{rc} V_{Oz}}{2\pi}. \tag{8}$$

The geometric centrifugal moment will be:

$$I_{xz} = A \xi z_{rc}, \tag{9}$$

or

$$I_{xz} = \frac{z_{rc} V_{Oz}}{2\pi}. \tag{10}$$

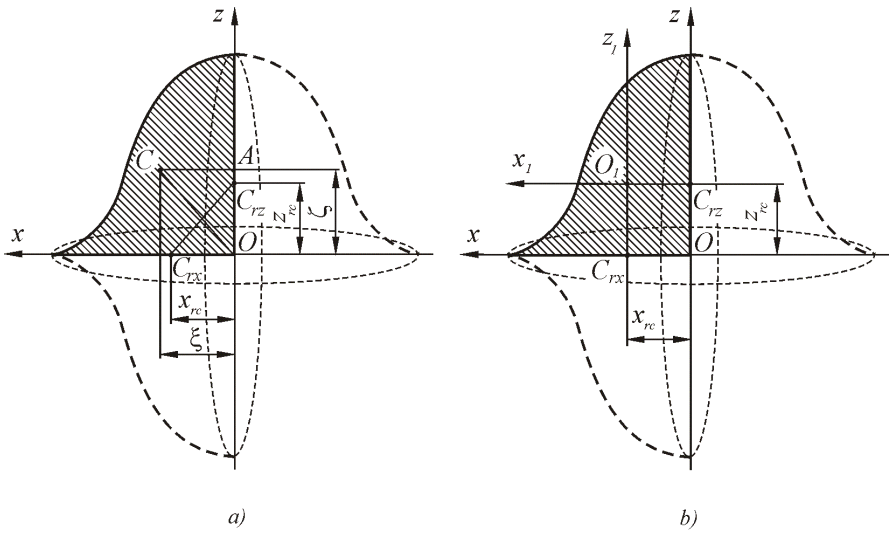


Figure 2

Plane plate rotated around two orthogonal axes: a) rotation bodies generated by the plate in rotation around two orthogonal axes; b) principal axes of inertia

Let us consider a plate like in Figure 2a. With the axes as in the figure, when we rotate the plate around the Oz axis, we obtain $J_{xz} = \rho A \xi z_{rc}$, and when we rotate the plate around the Ox axis, we obtain $J_{xz} = \rho A \zeta x_{rc}$. It follows that:

$$\xi z_{rc} = \zeta x_{rc}, \tag{11}$$

or

$$\frac{\xi}{\zeta} = \frac{x_{rc}}{z_{rc}}. \tag{12}$$

From the relation $\frac{\xi}{\zeta} = \frac{x_{rc}}{z_{rc}}$ and Figure 2a it results that the triangles OAC and $C_{rx}OC_{rz}$ are similar.

From the relation $J_{xz} = m\xi z_{rc}$ it results that, if $z_{rc} = 0$, then $J_{xz} = 0$. It follows that axis $C_{rz}x_1$, which crosses the mass center of the rotation body generated by the plate, is a principal axis of inertia (Figure 2b). Also, from the relation $J_{xz} = \rho A \zeta x_{rc}$ it results that axis $C_{rx}z_1$ is a principal axis of inertia.

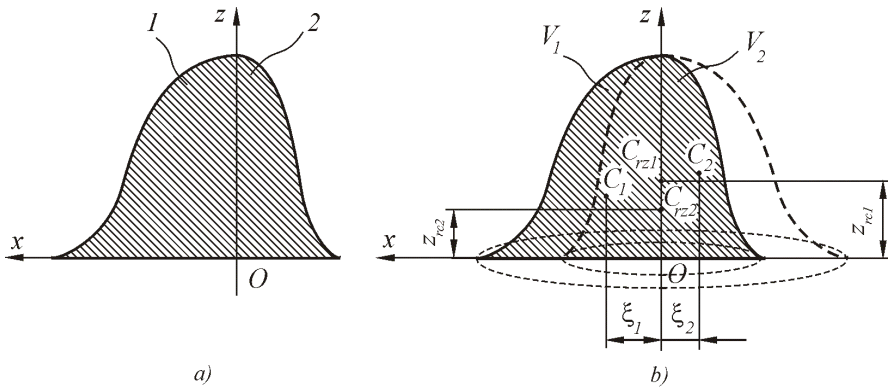


Figure 3
The case when the rotation axis cuts the plane plate: a) plane plate cut by an axis; b) rotation bodies generated by plate in rotation around the axis that cuts the plate

The relations (6), (7), (8), (9), (10) are available if the plate is fully situated on the same part of the rotation axis. In the case when the rotation axis cuts the plate (Figure 3a), the latter is split into two parts by areas A_1, A_2 and mass center coordinates ξ_1 and ξ_2, respectively (Figure 3b). By the rotation of these two parts, two rotation bodies are generated, with volumes V_{1Oz} and V_{2Oz}, respectively. In this case the centrifugal moment is:

$$J_{xz} = J_{1xz} + J_{2xz} = \rho A_1 \xi_1 z_{rc1} + \rho A_2 \xi_2 z_{rc2} = \frac{\rho\left(2\pi A_1 \xi_1 z_{rc1} + 2\pi A_2 \xi_2 z_{rc2}\right)}{2\pi} =$$

$$= \frac{\rho}{2\pi}\left[2\pi A_1 \xi_1 z_{rc1} - 2\pi A_2\left(-\xi_2\right) z_{rc2}\right] = \frac{\rho\left(z_{rc1} V_{1Oz} - z_{rc2} V_{2Oz}\right)}{2\pi}. \tag{13}$$

The geometric centrifugal moment will be:

$$I_{xz} = A_1 \xi_1 z_{rc1} + A_2 \xi_2 z_{rc2} = \frac{\left(z_{rc1} V_{1Oz} - z_{rc2} V_{2Oz}\right)}{2\pi}. \tag{14}$$

Let us consider the example of a homogeneous plane plate OAB (Figure 4a), quarter of disk of radius R. The surface density of the material is ρ (kg/m^2). With the axes as in the figure, we want to determinate the centrifugal moment J_{xz}.

First, the "classic way" will be used, the calculus by integration. Let dA be the area of an element of the plate, with the mass dm, which corresponds to an angle $d\theta$ and a radius r (Figure 4a). For this element we can write:

$dm = \rho dA = \rho r dr d\theta$; $x = r\cos\theta$; $y = r\sin\theta$.

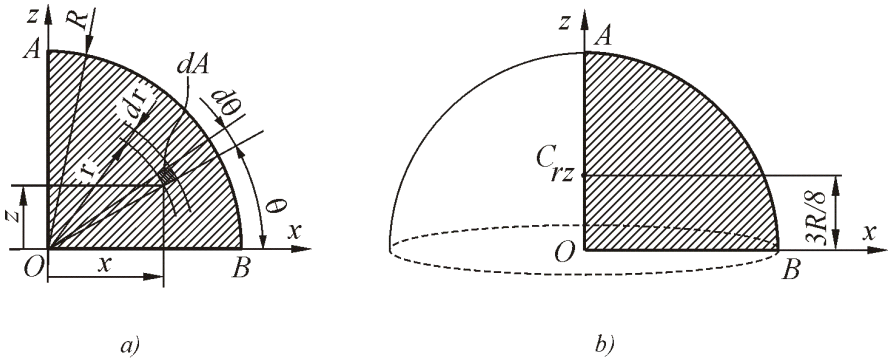


Figure 4

Determination of centrifugal moment for a plane plate quarter of disk: a) calculus of centrifugal moment for a plane plate quarter of disk, starting from definition; b) rotation body generated by plane plate in rotation

$$J_{xz} = \int_{(D)} xzdm = \iint_{(S)} r^2 \sin\theta\cos\theta\,\rho r dr d\theta = \rho\int_0^R \int_0^{\pi/2} r^3 \sin\theta\cos\theta dr d\theta =$$

$$= \rho\int_0^R r^3 dr \int_0^{\pi/2} \sin\theta\cos\theta d\theta = \rho\frac{R^4}{8}\int_0^{\pi/2} \sin\theta\cos\theta d\theta .$$

With the change of variable $v = \sin\theta$, $dv = \cos\theta d\theta$ it follows that:

$$J_{xz} = \rho\frac{R^4}{8}\int_0^1 v dv = \rho\frac{R^4}{8}\cdot\frac{1}{2} = \rho\frac{R^4}{8}.$$

The same result can be obtained quickly using the relation (8). During the rotation, the plate describes a semi sphere whose volume is $\frac{2}{3}\pi R^3$ (Figure 4b). The mass center of the semi sphere is on the axis of symmetry (Oz), $z_{rc} = \frac{3}{8}R$. So, for the centrifugal moment it results that:

$$J_{xz} = \rho\frac{\frac{3}{8}R\cdot\frac{2}{3}\pi R^3}{2\pi} = \rho\frac{R^4}{8}.$$

3 Centrifugal Moments for Bars

This study is similar to the one for plates presented above. A homogeneous plane curve bar is considered, with the mass *m*, length *l* and density ρ. We relate the bar to a Cartesian reference system (Figure 5a) so that the bar should be situated in the *xOz* plane. The Ox and Oz axes do not cut the bar. The center of mass C has the coordinates $C(\xi,0,\zeta)$.

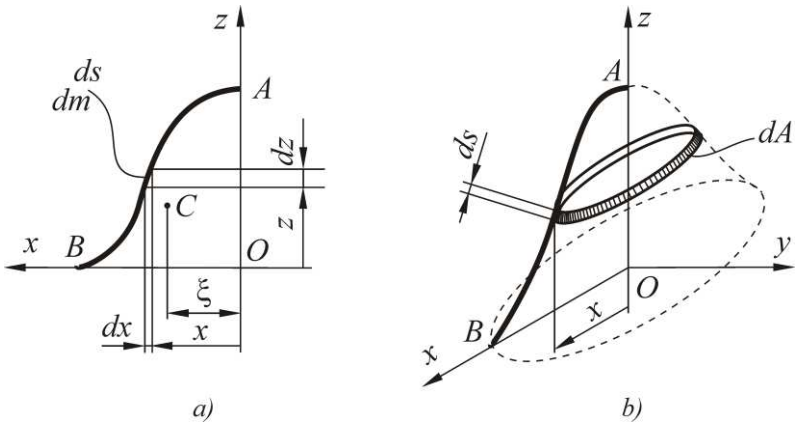


Figure 5
Determination of centrifugal moment for a plane bar: a) calculus of centrifugal moment, starting from definition; b) rotation surface generated by bar in rotation

We isolate an element with infinite little length $ds = \sqrt{1+(z')^2}\,dx$, with the mass dm. Starting from the definition of the centrifugal moment, we can write:

$$J_{xz} = \int_{(D)} xz\,dm = \rho \int_{(L)} xz\,ds = \frac{\rho}{2\pi} \int_{(L)} z 2\pi x\,ds = \frac{\rho}{2\pi} \iint_{(S)} z\,dA, \tag{15}$$

where $dA = 2\pi x\,ds = 2\pi x\sqrt{1+(z')^2}\,dx$ is the area of an infinite little element generated by the bar in rotation around the Oz axis (Figure 5b).

Taking into account the fact that the mass center coordinate z_{rc} of the rotation body (surface) generated by the bar in rotation around the Oz axis is given by the relation:

$$z_{rc} = \frac{\iint_{(S)} z\,dA}{\iint_{(S)} dA}, \tag{16}$$

the relation (15) becomes:

$$J_{xz} = \frac{\rho}{2\pi} \iint_{(S)} z dA = \frac{\rho}{2\pi} z_{rc} \iint_{(S)} dA = \frac{\rho}{2\pi} z_{rc} \int_{(L)} 2\pi x ds \,. \tag{17}$$

Taking into consideration that the mass center coordinate ξ of the plane bar on the axis which is perpendicular on the rotation axis Oz (used to generate the rotation surface) is given by the relation:

$$\xi = \frac{\int_{(L)} x ds}{\int_{(L)} ds}, \tag{18}$$

the relation (17) becomes:

$$J_{xz} = \frac{\rho}{2\pi} z_{rc} \int_{(L)} 2\pi x ds = \rho z_{rc} \int_{(L)} x ds = \rho z_{rc} \xi \int_{(L)} ds = \rho z_{rc} \xi l = m z_{rc} \xi \,, \tag{19}$$

where $m = \rho l$ represents the mass of the bar.

So we obtain the following formula for the centrifugal moment:

$$J_{xz} = m \xi z_{rc} \,, \tag{20}$$

or

$$J_{xz} = \rho l \xi z_{rc} \,. \tag{21}$$

In conclusion, the centrifugal moment is equal to the product of the mass of the bar, the mass center coordinate of the rotation surface generated by the bar, and the mass center coordinate of the bar on the axis which is perpendicular on the rotation axis.

Considering the first Guldin's law, $A_{Oz} = 2\pi \xi l$ (where A_{Oz} is the area of the rotation surface generated by the bar in rotation around the Oz axis), the relation (21) becomes:

$$J_{xz} = \frac{\rho z_{rc} A_{Oz}}{2\pi} \,. \tag{22}$$

Let us consider the example of a homogeneous plane straight bar AB (Figure 6a). We know the angle α between the bar and the Oz axis. The linear density of the material is ρ (kg/m). With the axes as in the figure, we want to determinate the centrifugal moment J_{xz}.

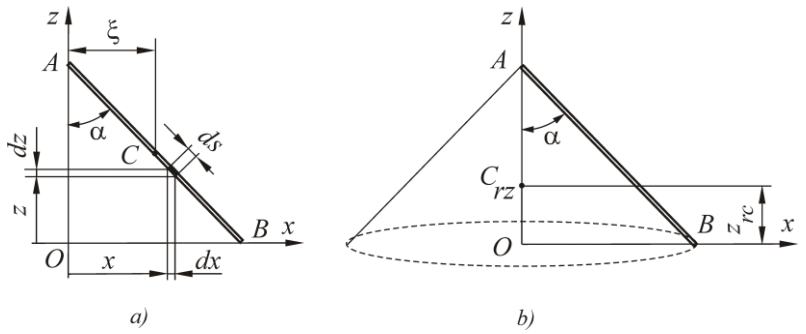


Figure 6

Determination of centrifugal moment for a plane straight bar: a) calculus of centrifugal moment for straight bar, starting from definition; b) rotation surface generated by straight bar in rotation

First the "classic way" will be used, the calculus by integration. Let ds be the length of an infinite little element of the bar, with the mass dm, which corresponds, on the axes, to the infinite little distances dx and dz, respectively (Figure 6a). For this element we can write:

$$z = -xctg\alpha + l\cos\alpha\ ;\ z' = \frac{dz}{dx} = -ctg\alpha\ ;\ dm = \rho ds = \rho\sqrt{1+(z')^2}\,dx = \frac{\rho}{\sin\alpha}dx\,.$$

$$J_{xz} = \int\limits_{(D)} xzdm = \frac{\rho}{\sin\alpha}\int\limits_{(L)} xzdx = \frac{\rho}{\sin\alpha}\int\limits_{0}^{l\sin\alpha} x\left(l\cos\alpha - xctg\alpha\right)dx =$$

$$= \frac{\rho}{\sin\alpha}\left(l\cos\alpha\int\limits_{0}^{l\sin\alpha} xdx - ctg\alpha\int\limits_{0}^{l\sin\alpha} x^2dx\right) = \frac{ml^2}{6}\sin\alpha\cos\alpha$$

This result can be obtained faster by using the relation (20). During the rotation, the bar describes a cone surface whose centre of mass is situated on the (Oz) axis (Figure 6b), $z_{rc} = \frac{1}{3}l\cos\alpha$. The mass center coordinate of the bar is $\xi = \frac{l}{2}\sin\alpha$. So, for the centrifugal moment it results that:

$$J_{xz} = m\frac{l}{2}\sin\alpha\cdot\frac{1}{3}l\cos\alpha = \frac{ml^2}{6}\sin\alpha\cos\alpha\,.$$

Conclusions

In this paper the authors have proposed formulae for the calculus of the centrifugal moments for plane plates and bars, based on the positions of the centers of mass.

To sum up, for a plane plate the centrifugal moment is equal to the product of the mass of the plate, the mass center coordinate of the rotation body generated by the

plate, and the mass center coordinate of the plate on the axis which is perpendicular on the rotation axis. Also, for a plane bar the centrifugal moment is equal to the product of the mass of the bar, the mass center coordinate of the rotation surface generated by the bar, and the mass center coordinate of the bar on the axis which is perpendicular on the rotation axis.

Taking into consideration the fact that in the technical literature it is easy to find the positions of the mass centers for a lot of bodies (necessary in statical calculus), the formulae proposed here are accessible because they replace the integral calculus with arithmetical calculus.

References

[1] M., Boiangiu, A., Boiangiu, On the Support of the Resultant Vector of d'Alembert's Fictitious Forces for Bars and Plates in Rotation Motion, Acta Polytechnica Hungarica, Vol. 8, No. 2, 2011, pp. 81-89

[2] V., I., Arnold, Mathematical Method of Classical Mechanics, second edition, Springer science, 1989

[3] G., R., Fowles, G., L., Cassiday, Analytical Mechanics, Harcourt College Publishers, 1998

[4] M., Radoi, E., Deciu, Mecanica, Editura Didactica si Pedagogica, Bucharest, 1993, p. 542

[5] R., Voinea, D., Voiculescu, V., Ceausu, Mecanica, Editura Didactica si Pedagogica, Bucharest, 1975, p. 85

Real-time Traffic Sign Recognition with Map Fusion on Multicore/Many-core Architectures

Kerem Par, Oğuz Tosun

Computer Engineering Department, Bogaziçi University, 34342 Bebek, Istanbul, Turkey, e-mail: k.par@ieee.org, tosuno@boun.edu.tr

Abstract: This paper presents a parallel implementation and performance analysis of a system for traffic sign recognition with digital map fusion on emerging multicore processors and graphics processing units (GPU). The system employs a particle filter based localization and map matching and template-based matching for sign recognition. In the proposed system, a GPS, odometer and camera are fused with digital map information. The system utilizes the depth sensor of a Kinect camera for the detection of signs and achieves high recognition rates for both day and night conditions. Tests were performed on real data captured in the vehicle environment comprising various road and lighting conditions. Test results show that speed increases of up to 75 times for localization and 35 times for sign recognition can be achieved on parallel GPU implementation over sequential counterparts. As those speedups comply with real-time performance requirements, high computational cost of using map topology information with large number of particles in localization implementation and template based matching for sign recognition is proven to be handled by emerging technologies. The system is unique since it is not limited to certain sign types; it can be used in both day and night conditions and utilizes a Kinect sensor to achieve a good price/performance.

Keywords: traffic sign recognition; particle filter; Kinect; multicore; gpu computing

1 Introduction

With the rise of multicore and many-core processors, the way of computing has been evolving into a new era. The high computational power, energy efficiency and programmability of these emerging general purpose processors make them a good candidate for a unified vehicle computing platform to host advanced driving assistance systems (ADAS) and autonomous vehicle applications by replacing specialized hardware and/or software platforms for each application. On the other hand, meeting the real-time performance requirements of those applications on such a platform is a challenge. Parallelization and using parallel programming techniques is one of the key methods to speed up applications on multicore architectures.

In this work, we present a parallel implementation and performance analysis of a complete system for sign recognition with map fusion including localization and map matching, both on a multicore processor using Open Multi-Processing (OpenMP) and on a graphics processing unit (GPU) using Compute Unified Device Architecture (CUDA).

The proposed system is unique, with many features, since it is not limited to speed signs, uses topological features of digital maps, and shows good performance in ambient lighting conditions. As a side contribution, the system utilizes a Kinect sensor, which simplifies sign detection radically and lowers overall system cost.

In the area of sign recognition with map fusion, the localization and map matching step is generally ignored and assumed as perfect. We provide a complete system, including the map matching and localization. We use a particle filter-based matching and localization algorithm proposed in [1] by the authors where GPS (Global Positioning System) and odometer data is fused with the topology of the digital map data as an additional sensor. The algorithm also generates a probabilistic measure for the correctness of the map matching. This measure is taken into account while using the digital map for sign recognition.

The system utilizes the depth sensor of a Kinect camera for the detection of signs. Classification is carried out by a template matching based algorithm, where the digital map information and vehicle position provided by the localization and map matching module are fused.

The target architecture is a combination of a multicore CPU and a many-core GPU, which is very likely to take place in a production vehicle environment as a unified computing platform in the near future. Both modules run on the same platform. The system overview can be seen in Fig. 1.

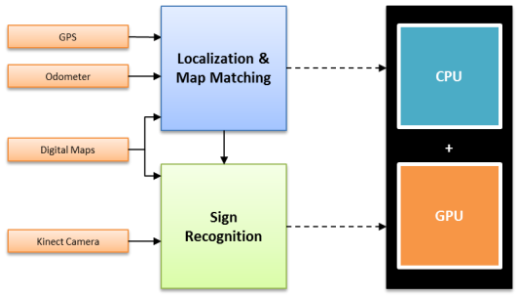


Figure 1
System overview

Both particle filter-based localization and map matching and template matching based sign recognition are computationally intensive applications where high success rates and real-time performance cannot both be achieved simultaneously using sequential implementations. The proposed system achieves both targets by employing parallelization on a hybrid multicore/many-core architecture.

We consider a multi-hypothesis localization and map matching algorithm where map topology information is used in terms of route-ability as the likelihood calculation in the particle filter to increase map matching performance, at the same time further increasing the computational cost of the algorithm.

We first characterized the execution profile of the particle filter algorithm for the different number of particles using a sequential implementation. Critical function blocks in terms of execution time were identified. We also investigated the effect of the number of particles employed by the algorithm on the error rate of localization and map matching. We then mapped the algorithm to the multicore CPU and the GPU platforms to accelerate bottlenecks and to see if the required speedups are realizable.

For our template-matching based sign recognition algorithm, which can be applied to a wide range of traffic signs, we employed a similar approach; we first observed detection rates using the Kinect sensor and recognition rates with the map fusion, and we also characterized the execution profile using a sequential implementation of the algorithm. We then tested the parallel implementations on our test system having two 6-core CPUs and two 512-core GPUs with real video and positioning data captured in the vehicle environment under various road and lighting conditions.

In the rest of this paper, we first give information about related work in this area. We then describe the particle filter-based localization and map matching algorithm and the sign recognition approach with map fusion. We give details about our parallel implementations for both modules. We continue with tests and experimental results. We finish with the interpretation of those results.

1.1 Related Work

Traffic sign recognition is one of the key components of advanced driver assistance systems and has been worked on for a long time in the intelligent vehicles domain. Although the appearance of the traffic signs was originally designed to be easily distinguishable from natural objects, the reliable, automated recognition of traffic signs, especially under adverse environmental conditions, remains a complex task.

Recent approaches tend to use a scheme of three stages: detection of sign candidates, classification of the candidates and tracking of the sign candidates over time. Many algorithms available in the literature generally differentiate in using different methods in the detection and classification stages [2]-[5].

There are also some attempts to enhance the performance of visible light cameras for sign recognition using infrared cameras [6]-[8]. They are limited to a subset of traffic signs and use expensive hardware, and the recognition rate is lower than the proposed system.

Most of the work done in this field so far has been strictly bounded by available computing capacity. However, recent developments in multicore and many-core architectures present a research challenge, also in this area, to meet real-time performance requirements with a parallel processing model. There are a very limited number of studies in the literature for parallel implementations of traffic sign recognition. [9] and [10] describe the detection and classification of traffic signs on an application-specific multicore processor. A real-time template-based approach for the recognition of speed limit signs using GPU computing is described in [11]. A feature-based speed limit sign detection system using a GPU is described in [12]. The studies cover only speed limit signs and do not include map integration.

We propose a generic template-based approach which can be applied to a wide range of traffic signs and the parallel implementation on a multicore CPU and GPU platform. Our approach uses a new sensor (Kinect) which provides both color and infrared images of the traffic scene, which enhances the detection stage, and we also propose using digital map information to augment template matching in the classification stage in order to increase the robustness of the recognition and to contribute to real-time performance.

Kinect has become very popular in a very short time since its launch in November, 2010, not only for playing games, but also, with its relatively low price, in robotics research for depth sensing and 3D vision. However, we have not yet encountered an application of Kinect in intelligent vehicles research.

Our approach employs a particle filter-based localization and map matching. Particle filters are among the principal tools for the on-line estimation of the state of a non-linear dynamic system [13]. Particle filtering has been applied widely in applications in tracking, navigation, detection and video-based object recognition [14]. Although, in general, particle filtering methods yield improved results compared to other Bayesian filters, it is difficult to achieve real time performance as the algorithm is computationally intensive [15]. This has been a prohibitive factor for real-time implementations for many applications of particle filtering.

A number of methods for software and hardware implementations of particle filtering have been proposed in the literature. Special architectures [16], field-programmable gate arrays (FPGAs) [17], and SIMD processor arrays [18] have been utilized for various types of problems. Many of the GPU implementations are focused on low-level stream processing or OpenGL [19].

Although emerging multicore processors and GPUs are good candidates for parallel particle filter implementations, multicore implementations, especially using the GPU computing concept and the platforms and tools such as NVIDIA's CUDA architecture, are still very recent and few. Some of the recent studies [20]-[21] utilize the general particle filter algorithm, but they differ significantly in their calculation of the likelihood phase. This variety also influence the approach used in parallelization.

Our work is a part of a research project addressing the challenge of meeting the real-time performance requirements of ADAS and autonomous vehicle applications by efficiently mapping them on multicore and/or many-core architectures, and, to our knowledge, this is the first parallelization effort of traffic sign recognition with map fusion using Kinect sensor and a particle filter targeted to localization and map matching using map topology.

2 Localization and Map Matching

2.1 Particle Filter

Particle Filters, also known as Sequential Monte Carlo (SMC) methods, are iterative methods that track a number of possible state estimates, so-called particles, across time and gauge their probability by comparing them to measurements.

We are considering a dynamic system with state x_t at a given time *t*. The *system model* is a Markov process of the first order. We assume that the system state can only be tracked by measurements y_t, which may be influenced by noise. The relation between measurements and system states is described by the *measurement model*.

The *sampling importance resampling* (SIR) algorithm is one of the most widely used sequential Monte Carlo methods. The SIR algorithm has following stages iterated over discrete time steps:

Sampling (Prediction): To follow the state during subsequent iterations, the system model is used to obtain a possible new state for every particle x_t^i based on its last state x_{t-1}^i where u_{t-1} is measured inputs and v_{t-1} unmeasured forces or faults:

$$x_t^i = Ax_{t-1}^i + Bu_{t-1} + v_{t-1}, \quad i = 1,\ldots,N \tag{1}$$

Importance (Update): The measurement model is evaluated for every particle and the current measurements to determine the *likelihood* that the current measurement y_t matches the predicted state x_t^i of the particle. The resulting likelihood is assigned as a weight w_t^i to the particle and indicates the relative quality of the state estimation:

$$w_t^i = w_{t-1}^i p(y_t \mid x_t^i), \quad i = 1,\ldots,N \tag{2}$$

At this point, when the particles are weighted, a state estimation can easily be obtained via various techniques, such as using the highest-weighted (highest-probability) sample, or using the weighted sum of the particles to get a mean-equivalent, or using the average of particles within some distance from the best particle.

$$\hat{x}_t \approx \sum_{i=1}^{N} w_t^i x_t^i \tag{3}$$

Resampling: If the number of effective samples fall below a certain value, resampling is required. Particles with comparatively high weights are duplicated and particles with low weights are eliminated. This can be done by calculating the number of effective particles N_{eff} as follows:

$$N_{eff} = \frac{1}{\sum_i (w_t^i)^2} \tag{4}$$

Effective sample size (ESS) is another metric to decide if resampling is required.

2.2 Particle Filter for Localization and Map Matching

For the vehicle localization problem, state is represented as a four-dimensional vector $x = [Lon, Lat, \Theta, L]$ where *Lon*, *Lat*, Θ and *L* stand for position, orientation and link or road segment on the map database, respectively.

Basically, the new location of the vehicle is predicted using the odometer data in the prediction stage and corrected by the GPS measurements and a map based likelihood function in the weight update stage. The operations performed in the main stages of the particle filter can be summarized as the following:

Prediction: The data coming from the odometer is used to measure vehicle displacement. The new location (*Lon*, *Lat*) of the vehicle is randomly calculated for each particle in the range of this displacement. This stage requires a high number of random number generations for the calculation of the new values of each state variable.

Weight Update: Weights are updated using the GPS readings first. The likelihood function is designed so that the particles that are within the error range of the GPS reading get higher weights. Then the weights are augmented with the map data by multiplying them with the probabilities derived from the map:

$$w_t^i = w_{t-1}^i \times p(\text{zone}) \times p(\text{topology}) \tag{5}$$

Two types of map attributes are used in the likelihood calculation. The first feature is the type of area where the particle resides on the map (road segment, building, parking area, etc.). The probability of being in a certain type of zone or road class (e.g. motorway, major road, local road, residential road, etc.) is calculated based on the speed of the vehicle (e.g., for a vehicle at a speed of 120 km/h, the relative probability of being on a motorway is chosen to be higher than being on a residential road).

The second feature of the map is the topology. Given the previous location of a particle, the probability of travelling to a new location on a certain road segment is calculated using the map topology. Possible reachable roads are searched in the road network in forward and backward directions for the distance travelled measured from the odometer. If the predicted location of the particle is found to be reachable, a high probability is assigned, otherwise a low probability is assigned (e.g., due to the connectivity, direction of traffic flow, turn restrictions, etc.).

Estimation: The location component of the system state is calculated as the weighted mean of each particle's location information. Map matching is achieved by selecting the road segment with highest weight as the matched link on the map. The flow of our particle filter algorithm for localization and map matching is shown in Fig. 2.

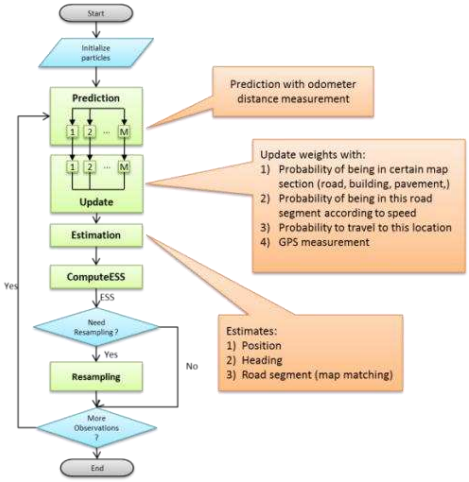


Figure 2
Particle filter localization and map matching

3 Sign Recognition

The proposed traffic sign recognition algorithm is implemented based on a template matching pipeline. The Kinect camera's depth image output is used to determine candidate regions on the RGB image. A special color segmentation scheme is applied to candidate regions. Template matching is employed for classification. A distance is calculated between the candidate region in the source image and different sizes of template images in the template database based on a difference function. The template having the minimum distance is denoted as the matched or recognized sign. Fig. 3 summarizes the design of the algorithm and the following sections describe the stages of the sign recognition flow in detail.

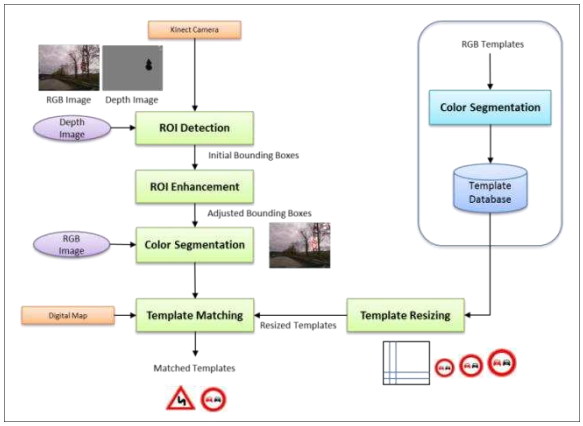


Figure 3
Sign recognition algorithm

3.1 Template Database and Map-based Probabilities

A template database is created from the sign images. Each sign template has two versions, one with a white background, the other with a black background. When sign recognition is carried out under night conditions, templates with black background are needed. Templates are also converted to four colors by use of color segmentation. Only black, white, red and blue are preserved in the image. We used an automatic resizing function according to the size of the region of interest found in the scene.

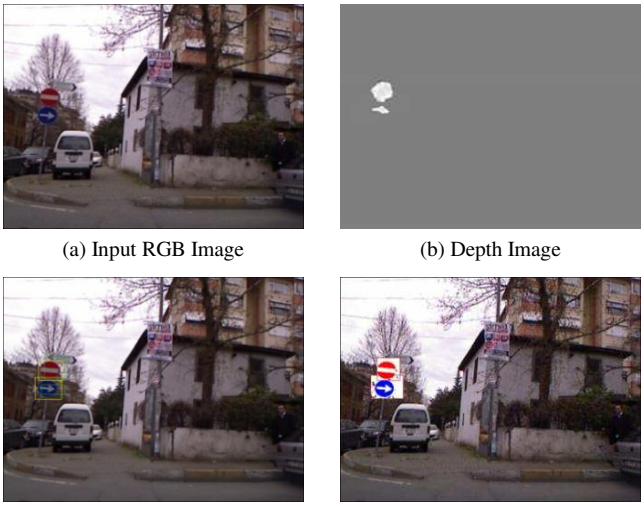
The localization and map matching algorithm determines the vehicle location and the map segment. By use of the matched segment, we can calculate map based probabilities for each sign in the database, considering different map contexts for various sign classes. Table 1 summarizes the sign classes and their respective map based context.

Table 1
Traffic signs and their respective map context

Sign Class	Signs	Map Context
Speed Signs		Road Class, Speed Limit
Manoeuvres		Manoeuvres (restrictions), One way information and map topology
Bends		Map topology, existence of a bend in the driving direction is checked.
Junctions		Map topology, existence of a junction and type of junction is checked
School		POI, existence of a school is checked.
Parking		Road Class

3.2 ROI Detection with Kinect Sensor

The Kinect camera's depth sensor consists of an infrared laser projector combined with a monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions. When used in outdoor environment, we end up with a very effective function of Kinect; it detects reflective surfaces (signs in traffic) and thus makes region of interest (ROI) detection very easy and robust. Fig. 4 (a) and (b) show the RGB image and the depth image coming from Kinect camera for the same scene. The depth image shows the region of interest in the RGB image.



(a) Input RGB Image (b) Depth Image

(c) Initial and Enhanced ROIs (d) Color Segmented ROIs

Figure 4
ROI detection with Kinect sensor

The initial bounding boxes are created by following the neighboring pixels within a given pixel tolerance. As seen in the Fig. 4 (c), the initial bounding boxes (green) are not perfect. There are several reasons for this. The IR camera and the RGB camera of the Kinect sensor have different fields of view and focal lengths. The RGB camera has a wider field of view. This is why, when objects get closer to the image edge, the difference in pixel locations increases. The two cameras are separated from each other by 2.5 cm. Sometimes the pixel image coming from the IR camera does not cover the whole sign. Sometimes the two signs are so close (their distance is smaller than the pixel tolerance when calculating the bounding boxes) that only one bounding box is found for two signs. There may be some small bounding boxes caused by reflections coming from other surfaces. An algorithm has been developed to overcome these errors. The initial and enhanced ROIs are shown in Fig. 4 (c) in green and yellow rectangles, respectively.

3.3 Template Matching

For a better matching, color segmentation is applied to the regions of interest first. All the colors in the image are segmented in four colors: red, blue, white and black. An example of color segmentation can be seen in Fig. 4 (d). After color segmentation, the templates are matched against the region of interests by computing the sum of the differences between pixel color values. For each region of interest, templates are resized based on the size of the bounding box before matching. Several template sizes with different aspect ratios are tried. Starting from corner of the region of interest, the difference between the template and the region of interest is calculated. The difference value is normalized according to the template size. The template with the lowest difference value is selected as the match.

3.4 Map Fusion

Template matching generates a likelihood measure for each sign. This measure is the distance between the template image and the camera image. Since we successfully detect the location of the sign on the camera image, the sign with the lowest distance value can be selected as the matched sign most of the time. But some of the signs are very similar to each other. Also, even if we find the location of the sign successfully, the camera image may not be clear. As a result of this, the algorithm returns very close likelihood results. When we fuse this information with the probabilities coming from the map, the correct sign can be selected. The recognition performance of our algorithm increases radically. Fig. 5 shows two examples of template matching results, with and without map fusion.

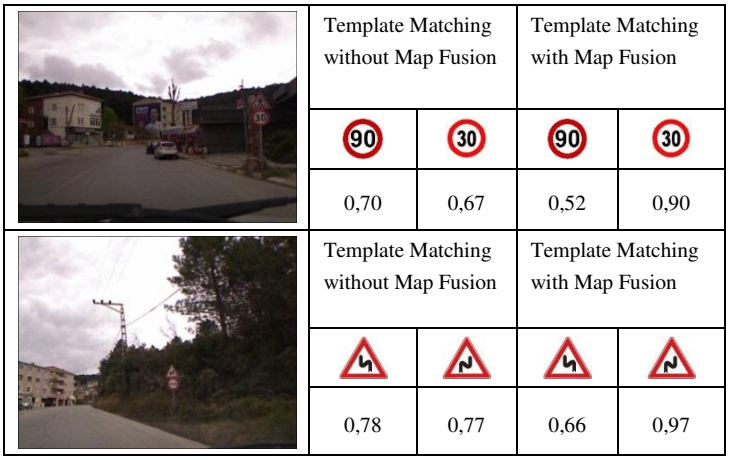


Figure 5
Sign recognition with map fusion

4 Parallel Implementations

4.1 Localization and Map Matching

Before attempting parallel implementations, we first characterized the execution profile of the particle filter algorithm for different number of particles using a sequential implementation. We see that the prediction and update sections dominate the execution time by a large margin. Therefore, those sections were selected as the first targets of parallelization in both platforms.

Particle filters heavily use random number generation. Our implementation uses the Mersenne-Twister random number generation algorithm. An existing implementation has been adopted for both the CPU and GPU platforms.

4.1.1 Multicore (OpenMP) Implementation

We used OpenMP programming model [22] for the parallelization of the predict and update sections of the particle filter on a multicore CPU. Since the same operations are repeated for all the particles in a loop for both the predict and update sections and the particles can be processed independently of each other, the iterations (effectively the particles) have been distributed among the cores. Each core therefore performs the prediction and update steps on a subset of particles. The static scheduling mechanism of OpenMP is used for the predict part and dynamic scheduling has been employed for the update part, in order to have a better workload distribution among the cores since the complexity of map based operations for each particle in the update step can be different.

4.1.2 GPU (CUDA) Implementation

In our GPU implementation, we used the CUDA programming model [23]-[24]. This actually represents a hybrid (CPU+GPU) implementation of particle filter. We implemented most of the main steps of the filter in C using CUDA Toolkit 3.2. The *Prediction*, *Update*, *Estimation*, and *ComputeESS* parts were implemented as kernels to run on GPU (device), where resampling part is run on CPU (host). The CUDA implementation flow is illustrated in Fig. 6.

Since the prediction and update parts of a particle filter work on particles independently, a separate thread is created for each particle on the GPU for the *predict* and *update* kernels. This is accomplished by using the appropriate execution configuration parameters, when the kernels are launched. Each thread determines which particle it should process via built-in variables, the thread block index, the thread index within its block, and the block size.

The states of particles are stored in the global memory, and during initialization both host and device memory are allocated for particles, and the initial particle data are copied to the device. The global memory is used to pass on data from one

kernel to the next. Map data is also transferred to the device memory during initialization. Each thread is enabled to use its own random number generator instance. Initial Twister states for the maximum possible number of threads are created on the host and transferred to the device memory at the initialization.

For the *update* kernel, measurement values are passed as parameters at the kernel launch for each iteration. The *estimation* part consists of the summation and normalization of the weights and the calculating weighted mean of the state variables. This part is divided into three separate kernels: *summation*, *normalizeWeights* and *mean* kernels. The division of the workload into separate kernels was necessary due to the fact that the only way to enforce synchronization between all concurrent CUDA threads in a grid is to wait for all kernels running on that grid to exit.

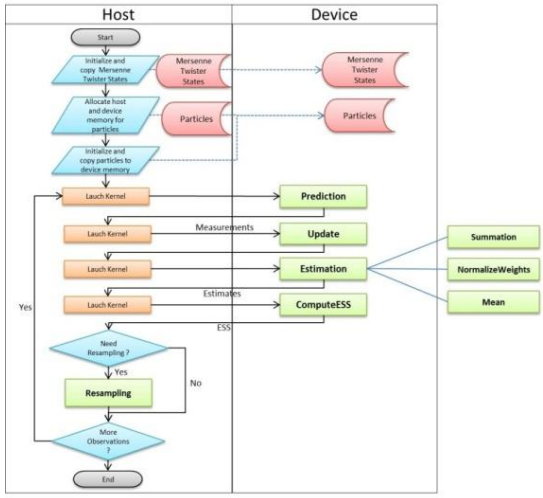


Figure 6

CUDA implementation of particle filter localization and map matching

For the *summation kernel*, the parallel prefix sum technique is used to calculate the partial sums within each block, and these partial sums are added to the global sum by using global atomicAdd operations. The *normalizeWeight* kernel is implemented similar to the *predict* and *update* kernels. Each thread adjusts its weight independently by using the sum value which is passed to it as a parameter at the kernel launch. The *mean* kernel and the *computeESS* kernel also use the parallel reduction technique similar to the summation kernel. After the estimation is completed, the estimated state variables are transferred to the host.

The amount of data transfers between the host and device has been kept very small for the iterations where resampling is not required. If resampling is required, the current weights of the particles are transferred to the host, and the surviving particles are calculated on the host.

4.2 Sign Recognition

The execution time profile of the sequential implementation shows that the template matching process has the highest computational cost, more than 98 percent of the total execution time. This has been chosen as the target for parallelization. The matching process for each video frame involves the following parameters:

r	number of ROIs detected in the frame
n	number of templates in the template database
m	number of different sizes for each template to be used for matching
s	number of different starting positions for matching in each ROI
w	width of template in pixels
h	height of template in pixels

Assuming (x,y) denotes the starting search image coordinates and (i,j) denotes the template image coordinates, the time required for the matching process for each frame can be defined as the following:

$$t = r \times n \times m \times s \times \sum_{i=0}^{h} \sum_{j=0}^{w} Diff(x+i, y+j, i, j) \tag{6}$$

Three parallel implementations have been developed for multicore CPU, single GPU and multi GPU architectures. For all cases, the detection stage is performed on the host sequentially, which is performed very fast with the help of the Kinect camera.

4.2.1 Multicore (OpenMP) Implementation

The multicore CPU implementation is performed using the OpenMP programming model. The matching operations for each template are distributed among the multiple CPU threads. The number of threads is determined by the maximum number of cores in the system. For each region of interest, the work is distributed on a templates basis.

4.2.2 GPU (CUDA) Implementation

The GPU implementation is performed using CUDA. The pixel level matching operations are designed to run on GPU in parallel. A kernel (*matching* kernel) has been implemented to perform the matching of a ROI to a resized template and produce the sum of differences (SAD) values. A separate thread is created for each pixel operation when the kernel is launched. Initially, all memory allocations are done for RGB images, resized templates and SAD values on both host and device. For each video frame, detection is performed on the host and ROIs are found. If at least one ROI is found in the depth image, the RGB image is copied to the device memory. Each ROI found in the frame is matched against different sizes and starting positions of all the templates by calling the *matching* kernel.

Resizing is done on the host, each template is resized based on the size of ROI, and resized templates are copied to the device memory before launching the *matching* kernel. Since the RGB image and the resized templates are already in the device, the kernel is then called with only the corner positions of the region of interest, the template number and the size of the template.

Since the number of pixels in region of interests are relatively small (e.g. 49x48) compared to whole images (640x480), to be able to achieve maximum occupancy of GPU cores, the *matching* kernel is designed to compute SAD values for all different starting positions (4x5) of a resized template each time it is launched. So each launch of the *matching* kernel performs 20 matching operations in parallel at the template level in addition to the pixel level parallelism (e.g., for a 44x40 pixel region of interest, 35200 threads are created instead of 1760, corresponding to 138 blocks instead of 7 blocks, respectively).

Kernels for each different size of the same template are launched concurrently using different streams. Concurrent kernels is a scheduling convenience allowing different streams of the same context to run simultaneously. It enables to increase the efficiency if there are inefficient low block count kernels, mostly by reducing idle streaming multiprocessor count while kernels are finishing up. The maximum number of concurrent kernels that can be executed on a Fermi GPU is 16. The number of different sizes (4x4) to be matched for each template is also 16 in our implementation. This enables the matching of all different sizes of a template to be launched concurrently. SAD values are accumulated in the global memory by using AtomicAdd operations. For each template, after calling the kernels for all variations, the SAD values are copied back from the device to host, and for each region of interest, the SAD values are processed to determine the result of recognition. The flow of the implementation is shown in Fig. 7.

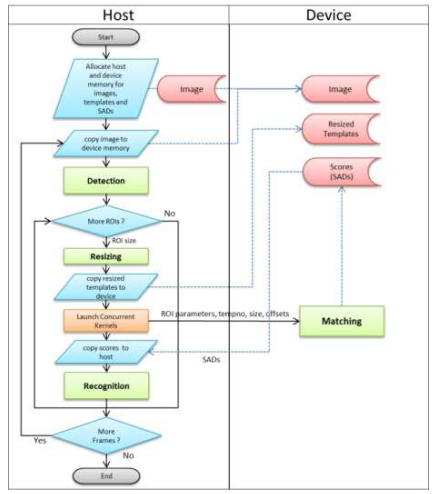


Figure 7
CUDA implementation of sign recognition algorithm

4.2.2 Multi GPU Implementation

The multi GPU solution can be used with any number of GPUs. This is also a hybrid implementation. Five CPU threads are used. The *detection* thread gets the depth, and RGB image frames, perform the detection phase, for each ROI found in the depth image, resizes the templates based on the size of ROI and puts the related data into a queue to be passed to a GPU to perform the matching. The *Dispatcher* thread keeps track of the availability of GPUs and determines the target GPU that will process the next ROI data and assigns the device number to the data slot in the queue. Each GPU has to be controlled by one CPU thread in multi GPU programming with CUDA Toolkit 3.2. Two *matching* threads are responsible for controlling the GPUs, including sending the required data (i.e. RGB image, ROI boundary, resized templates) to the device, launching the *matching* kernels concurrently, receiving the SAD values from the device and storing them into the results queue. The *recognition* thread processes the SAD values and determines the sign recognized for each ROI. The implementation details are depicted in Fig. 8.

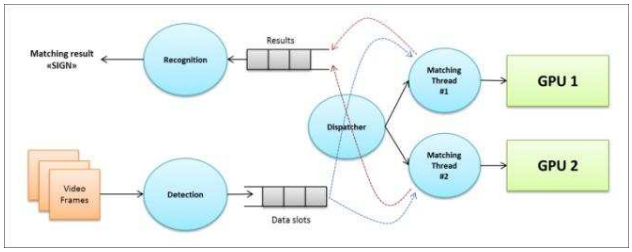


Figure 8
Multi GPU sign recognition implementation

5 Experiments

We have tested performance of our parallel implementations using real video, GPS and odometer data captured in six test routes comprising various road (highways, urban traffic, etc.) and lighting conditions (night/day, sunny/cloudy). Parallelization tests were performed on our test platform, a dual processor HP® Z800 workstation having two Intel® Xeon® 5660 6-core processors running at 2.80 GHz and two NVIDIA® GeForce GTX580 graphics processing units.

The GTX580 GPU has NVIDIA's new generation CUDA architecture called Fermi and has 16 streaming multiprocessors, each having 32 streaming processors, and thus in total has 512 processing cores. Hence, it is capable of running 512 threads simultaneously. Each core runs at 1.544GHz. Each streaming multiprocessor has 64KB configurable L1 cache. All cores shares a 768MB L2 unified cache and a 1512MB global memory.

5.1 Localization and Map Matching

We used one 6-core CPU and one 512-core GPU in our tests. Tests were repeated on each platform for different number of particles ranging from 256 to 128K. For the multicore CPU tests, we ran the OpenMP implementation with 6 threads. For the CPU+GPU tests, the block size was chosen as 256.

The OpenMP implementation provided approximately a 4.7x speedup with a theoretical maximum increase of 5.4x on a 6-core CPU. We observed similar speedups after the number of particles exceeds 4096.

With the CUDA implementation, we achieved increasing speedups of up to 75x when the number of particles reached 128K. We see that the performance of GPU is better exploited when the number of particles or threads is increased. The relatively low speedups for the smaller number of particles are mainly due to the low occupancy of streaming multiprocessors. Fig. 9 shows the execution times of sequential, multicore CPU and GPU implementations for different number of particles and the relative speedups.

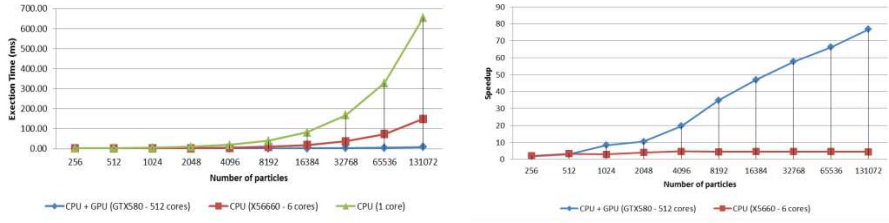


Figure 9

Execution time and speedup comparisons of sequential and parallel implementations

When we examine performance of kernels separately, we see that speedups can be as high as 150x for the *predict* kernel, where there are no data dependencies among threads and operations performed are almost identical for all threads. We see 100x speedups for the *update* kernel, where we observe the negative effect of branching and divergence on the performance since road network is traversed to a new location for some particles which causes different execution paths for threads. We see speedups around 10x for the *estimation* and *computeESS* kernels, where synchronization requirements within blocks and global atomic operations reduce speedups. However, overall speedups achieved are sufficient for real-time localization and map matching using a high number of particles.

We examined the sensitivity of the localization and map matching performance to the number of particles to determine the optimum number. The error rate is calculated as the ratio of the number of wrong map matches to the total number of positions on the test routes. We see that the error rate decreases significantly until the number of particles exceeds 32K. Fig. 10 shows the error rates for two different routes.

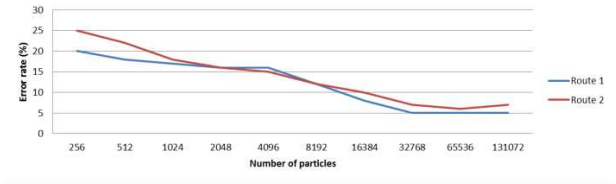
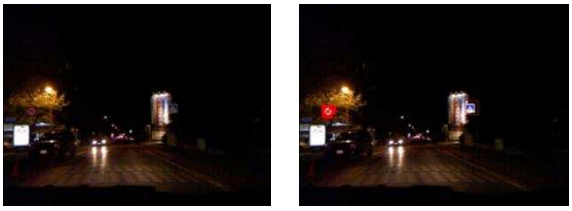


Figure 10

Effect of number of particles on the error rate of map matching algorithm

5.2 Sign Recognition

Since detection of ROIs are handled by the the Kinect camera, detection is successful even in very bad lighting conditions. We have observed that the system can detect signs that can hardly be seen by human eye. Fig. 11 shows an example of a successful recognition at night. Table 2 summarizes success rates of detection and recognition for different route types. We see that map fusion improves recognition performance dramatically especially under night conditions.



(a) Input RGB Image (b) Recognized signs

Figure 11

Successful recognition at night conditions

Table 2

Detection and recognition rates for traffic signs using Kinect camera

Route Type	Detection Rate	Without map fusion	With map fusion	Improvement
Cloudy, Residential Roads - Urban	93%	84%	92%	9%
Sunny, Residential Roads - Urban	89%	71%	85%	20%
Cloudy, Main Roads	91%	71%	86%	20%
Cloudy, Connecting Roads- Rural	95%	50%	83%	66%
Night, Main Roads	94%	55%	88%	60%
Night, Residential Roads	92%	40%	80%	100%

Multicore CPU and GPU implementations were tested on the same platform. The average processing time for frames was measured and the execution time of sequential implementation was taken as a reference in the speedup calculations. For each ROI, 16 (4x4) different starting positions, and for each template, 20 (4x5) different sizes, were used. Tests were performed with a template database having 52 templates. The recognition of each ROI involved 16,640 matchings.

Multicore CPU implementation were tested with different numbers of threads, ranging from 1 to 24. Speedups of up to 10.6x were achieved. We observed linearly increasing speedups until the number of threads reached the number of cores in the system. After that point, we observed that the speedups were not improved with the increasing number of threads, but rather stayed in the range between 8.7 and 9.7. The execution time at the maximum speedup was around 250ms corresponding to 4 frames per second. The linear speedups show that we can further increase frame rates when we have a higher number of cores in the system.

Speedups up to 18.1x and 35.2x were achieved on a single GPU and multi GPU tests, respectively. The execution times at the maximum speedups approximately correspond to 7 and 13 frames per second. For GPU tests, we used 256 threads as the block size. We observed that 100% occupancy was achieved. Speedups and execution times for all implementations are shown in Fig. 12.

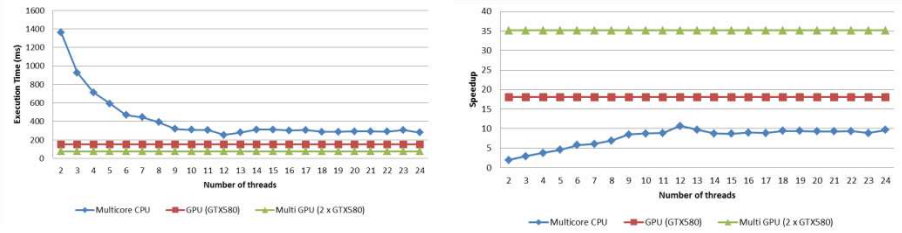


Figure 12

Execution time and speedup comparisons of sequential and parallel implementations

Conclusions

We introduced a real-time traffic sign recognition system with digital map fusion, and we examined parallel implementations and performance analysis on emerging multicore CPUs and GPUs. Test results show that up to 75 times speedups can be achieved for particle filter based localization and map matching on GPU over sequential implementation, and real-time performance is possible in the case of high computational cost of using map topology information. We showed that success of localization and map matching can be increased by employing a high number of particles where real-time performance can be achieved only by parallelization.

The speedups achieved for our sign recognition system show that the template matching based recognition approach with map augmentation, which is a simple but computationally intensive technique, can be used with real-time performance in the vehicle environment. We observed detection rates over 90% using the Kinect sensor and recognition rates over 80% for various road and lighting conditions. Test results show that the system performs very well even in night conditions. The proposed system is unique since it is not limited to certain sign types, can be used for recognition of wide range of traffic signs, can be used in

any lighting conditions, utilizes the Kinect sensor to achieve a good price/performance, and runs on commercially available parallel hardware. Our future work will include investigating the co-scheduling of other tasks that can run simultaneously on the same platform with sign recognition and localization while delivering required throughput and minimal affordable latency.

Acknowledgement

This research was supported by Bogazici University Research Fund, contract No. 5522.

References

[1] K. Par, and O. Tosun, "Parallelization of Particle Filter Based Localization and Map Matching Algorithms on Multicore/Manycore Architectures", Intelligent Vehicles Symposium (IV), 2011 IEEE, pp. 820-826, June 2011

[2] B. Hoferlin and K. Zimmermann, "Towards Reliable Traffic Sign Recognition", IEEE Intelligent Vehicles Symposium, pp. 324-329, June 2009

[3] A. de la Escalera, J. Ma Armingol, M. Mata, "Traffic sign recognition and analysis for intelligent vehicles", Image and Vision Computing, pp. 247-258, Vol. 21, 2003

[4] C. Bahlmann et al., "A System for Traffic Sign-Detection, Tracking and Recognition Using Color, Shape and Motion Information", IEEE Intelligent Vehicles Symposium (IV), pp. 255-260, June 2005

[5] J. Ban, M. Feder, M. Oravec, and J. Pavlovicova, "Non-Conventional Approaches to Feature Extraction for Face Recognition", Acta Polytechnica Hungarica, Vol. 8, No. 4, pp. 75-90, 2011

[6] Weijie Liu, and Maruya, K., "Detection and Recognition of Traffic Signs in Adverse Conditions", IEEE Intelligent Vehicles Symposium, pp. 335-340, June 2009

[7] Tsz-Ho Yu, Yiu-Sang Moon, Jiansheng Chen, Hung-Kwan Fung, Hoi-Fung Ko, Ran Wang, "An Intelligent Night Vision System for Automobiles", IAPR Conference on Machine Vision Applications, May 2009

[8] B. Kuljic, J. Simon, and T. Szakall, "Pathfinding Based on Edge Detection and Infrared Distance Measuring Sensor", Acta Polytechnica Hungarica, Vol. 6, No. 1, pp. 103-116, 2009

[9] R. Ach, N. Luth, A. Techmer, "Real-Time Detection of Traffic Signs on a Multi-Core Processor", IEEE Intelligent Vehicles Symposium, pp. 307-312, June 2008

[10] R. Ach, N. Luth, A. Techmer, A. Walther, "Classification of Traffic Signs in Real-Time on a Multi-Core Processor", IEEE Intelligent Vehicles Symposium (IV), pp. 313-318, June 2008

[11] V. Glavtchev, P. Muyan-Ozcelik, J. M. Ota, and J. D. Owens, "Feature-based Speed Limit Sign Detection Using a Graphics Processing Unit", Intelligent Vehicles Symposium (IV), 2011 IEEE, pp. 195-200, June 2011

[12] P. Muyan-Ozcelik, V. Glavtchev, J. M. Ota, and J. D. Owens, "A Template-based Approach for Real-Time Speed-Limit Sign Recognition on an Embedded System Using GPU Computing", Proceedings of the 32nd DAGM conference on Pattern recognition, pp. 162-171, 2010

[13] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "A Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation," IEE Proceedings on Radar and Signal Processing, Vol. 140, No. 2, pp. 107-113, 1993

[14] F. Gustafsson, F. Gunnarson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P. Nordlund, "Particle Filters for Positioning, Navigation and Tracking", IEEE Transactions on Signal Processing, Vol. 50, No. 2, pp. 425-437, February 2002

[15] D. Kocur, J. Gamec, M. Svecova, M. Gamcova and J. Rovnakova, "Imaging Method: An Efficient Algorithm for Moving Target Tracking by UWB Radar", Acta Polytechnica Hungarica, Vol. 7, No. 3, pp. 6-24, 2010

[16] Bolic, M., "Architectures for Efficient Implementation of Particle Filters", PhD Dissertation, Stony Brook University, August 2004

[17] M. Happe, E. Lübbers, and M. Platzner, "A Multithreaded Framework for Sequential Monte Carlo Methods on CPU/FPGA Platforms", Proceedings of the 5th Inte Workshop on Reconfigurable Computing: Architectures, Tools and Applications, pp. 380-385, 2009

[18] H. Mederios, J. Park, and A. Kak, "A Parallel Implementation of the Color-based Particle Filter for Object Tracking", IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-8, June 2008

[19] G. Hendeby, R. Karlsson, and F. Gustafsson, "Particle Filtering: The Need for Speed", EURASIP Journal on Advances in Signal Processing, 2010

[20] J. F. Ferreira, J. Lobo, and J. Dias, "Bayesian Real-time Perception Algorithms on GPU", J. Real-Time Image Proc., Springer, 2010

[21] M. A. Goodrum, M. J. Trotter, A. Aksel, S. T. Acton, and K. Skadron, "Parallelization of Particle Filter Algorithms", Proc. of 3rd Workshop on Emerging Applications and Many-core Architecture (EAMA), 2010

[22] B. Chapman, G. Jost, and R. van der Pas, "Using OpenMP, Portable Shared Memory Parallel Programming", The MIT Press, 2007

[23] J. Nickolls, W. J. Dally, "The GPU Computing Era", IEEE Micro, pp. 56-69, March-April 2010

[24] NVIDIA CUDA Programming Guide 3.2, 2010