Formation Kinetics of the Self-organized III-V-based Nanostructures Grown by Droplet Epitaxy

Ákos Nemcsics

Institute for Microelectronics and Technology, Óbuda University
Tavaszmező u. 17, H-1084 Budapest, Hungary; and
Research Institute for Technical Physics and Materials Science, Hungarian Academy of Sciences, P.O.Box 49, H-1525 Budapest, Hungary
e-mail: nemcsics.akos@kvk.uni-obuda.hu

Abstract: In this work, we discuss the evolution of self-assembled III-V-based nanostructures. These nanostructures were prepared from Ga droplets during a crystallization process using a droplet epitaxy technique. Different nanostructures such as quantum dots, quantum rings, double quantum rings, and nano holes were formed from Ga droplets on a (001)-oriented AlGaAs surface, but at different substrate temperatures and various arsenic background pressures. We give a qualitative description of the elemental processes for the formation of all of these nanostructures. Our description is based on the size dependence of the key material properties (solubility, tension, etc.).

Keywords: GaAs; droplet epitaxy; quantum dot; quantum ring; nano hole

1 Introduction

Low dimensional structures grown by molecular beam epitaxy (MBE) revolutionized electronic devices, both in terms of their potential and their efficiency. Nowadays, the growth of self-organized nanostructures is thoroughly investigated for basic physics and device applications. It is very important to understand their growth kinetics and to know their shape. The most widespread technique is based on the lattice mismatched Stranski-Krastanov growth mode [1, 2]. An example is the MBE growth of InAs quantum dots (QDs) driven by strain between the deposited InAs and the substrate. For a long time, this strain-induced method was the only known process for zero dimensional structure production.

A droplet epitaxial technique based on Koguchi's discoveries has evolved, giving greater opportunities for the development of the self-organizing nanostructures [3, 4]. In this method, the lattice-mismatch loses its significance and it becomes

possible to create QDs [5-9], quantum rings (QRs) [10-12], double quantum rings (DQRs) [13, 14], as well as nano holes (NHs) [15-17] (see Fig. 1). The electronic structure of these nano-objects depends very much on their shape. The droplet epitaxial process roughly consists of the following steps: first, metal (e.g. Ga) droplets are generated on the surface in the Volmer-Weber-like growth mode. After that, crystallization of the droplets occurs together with their chemical transformation, e.g. GaAs QDs under arsenic pressure. In order to control the process it is necessary to understand the kinetics of the growth processes, but no theoretical description is available yet of the underlying growth mechanism. In this paper, we study the formation and properties of GaAs nanostructures grown on AlGaAs (001) substrate as a function of the applied technology.

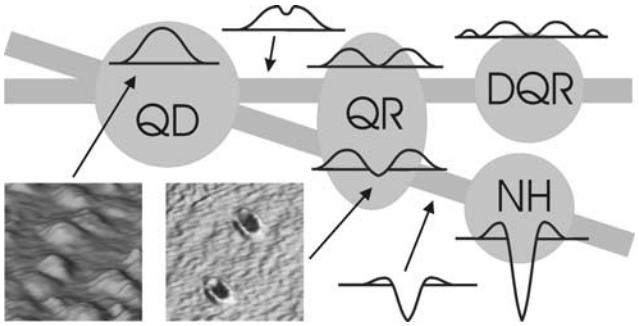


Figure 1
Overview of the III-V-based dropletepitaxially grown nanostructures such as quantum dots (QDs), quantum rings (QRs), double quantum rings (DQRs) and nano holes (NHs)

2 Experimental Preliminaries

The growth experiments were made in a solid source MBE system. The evolution of the growth front was monitored in the [1$\underline{1}$0] direction with reflection high-energy electron diffraction (RHEED). After the growth, the QDs were investigated using atomic force microscopy (AFM) in tapping mode. Droplet epitaxial GaAs QDs were formed on the AlGaAs (001) surface (with Al content of $x = 0.3$).

The main steps of the droplet epitaxial process of QDs were as follows: Following the preparation of $Al_xGa_{1-x}As$ surface, the samples were cooled to 200 and 250 °C, Ga was deposited first (the surface coverage Θ=3.75 mono layer (ML)) with the flux of 0.15 and 0.025 ML/s without the arsenic flux, respectively (sample type *a* and *b*). After a few seconds' waiting time, an annealing step was performed at 350 °C under 6.7×10^{-5} mbar arsenic pressure. A detailed experimental description is published elsewhere [18].

In the case of sample type *c* (droplet-epitaxial QRs), after the preparation of the AlGaAs layer, the sample was cooled to 300 °C, and θ=3.75 ML Ga was deposited with the flux of 0.75 ML/s without arsenic flux. After Ga deposition, 60 seconds' annealing (300 ºC), under 5.3×10^{-6} mbar arsenic pressure, was applied. A detailed description of the growth parameter is published elsewhere [19].

A further five types of nanostructures (sample types *d* - *h*)were also prepared. The technological parameters are given as follows: The solution process was investigated in two variations (sample types *d* and *e*). NHs (sample type *d*) were generated at 570 °C in AlGaAs surface applying 6.4 ML Ga [16]. In this case, the AFM measurement shows NHs and very large clusters [16]. Other types of NHs (sample type *e*) were prepared similarly, but the Ga coverage was different (3.2 ML) [15, 16]. Here, the AFM picture shows deep NHs surrounded by ring-like bulge formations and shallow NHs, with plane rims (without any bulge) [15, 16].

The crystallization process is investigated in three variations of samples (sample types *f* - *h*). The first nanostructure is a very fast-crystallized QD (sample type *f*). The crystallization of the Ga droplet occurs during the high arsenic supply (3.3×10^{-4} mbar) at 150 °C [8]. The elementary map of transmission electron microscopy (TEM) shows that the QD had Ga inclusons [8]. The second type of samples (type g) is crystalized at 2.7×10^{-6} mbar arsenic pressure at low temperature (200 °C) [13]. In this way, we received DQR [13]. The last sample (type *h*) is crystallized at about the same arsenic background (1.5×10^{-6} mbar) but at a high temperature (500 °C) [17]. The middle of the structure is surrounded by single ring-like bulge formations [17]. For the detailed technological parameters of the above samples see the given references.

3 Discussion

3.1 Formation Kinetics of Quantum Dots

The formation of the QD structures was in-situ monitored by RHEED. Firstly, Ga droplets were formed on an AlGaAs (001) surface. As the second step, the droplets were transformed into QDs under an arsenic environment. In order to investigate of droplet formation in one occasion, the preparation procedure of QDs was interrupted before the crystallization step. This particular sample was quenched and inspected by AFM immediately after the Ga droplets were formed. This AFM image was compared to another sample, on which the Ga droplets, formed under identical conditions, were reacted with As in the crystallization step. The density and the size of QDs were found to be identical to te Ga droplets. Because of the grazing incidence angle, the RHEED pattern is characteristic for the nano structure (see Fig. 2/A). The Ga clusters formed on the surface are liquid

aggregates; this status is proven by the RHEED picture as well as by the shape of the clusters. The sharp RHEED streaks characteristic of the AlGaAs (001) surface vanished after Ga deposition and the streaky RHEED picture becomes diffused because of the electron beam crossing the liquid aggregate. The RHEED temporal patterns of the QD formation are shown in middle part of Fig. 2/B-E. As mentioned before, the Ga clusters before crystallization are investigated with AFM working in tapping mode. The original shape of the droplets is conserved during the AFM measurement due to the residual adsorbate on the droplet surface, which exists also under 10^{-10} mbar [20-22]. The shape of these clusters follows the law of the surface and interface energy minimization (see later) [23-29].

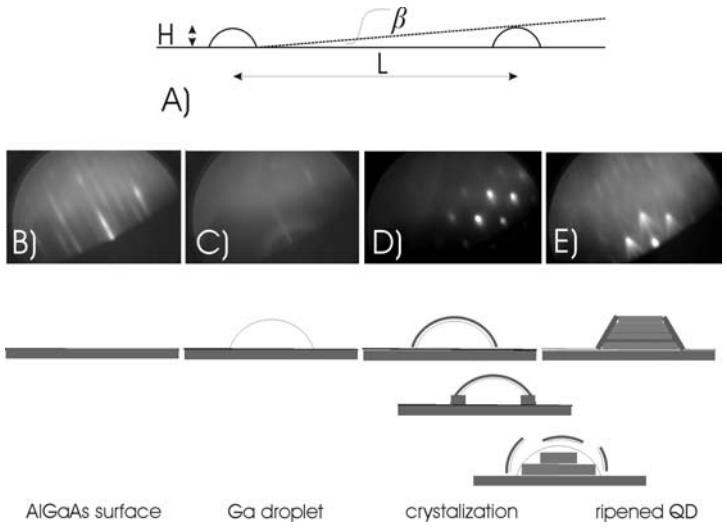


Figure 2

(A): The RHEED geometry on the surface. At N_1=1.2H10^{10} cm^{-2} and N_2=4.4H10^8 cm^{-2}, the average height (H) of the droplet and the characteristic average distance (L) between these nanostructures are H_1=7 nm, H_2=32 nm and L_1=90 nm, L_2=477 nm, respectively. The RHEED pattern is characteristic to the nano-structures, because the β_1=4º and β_2=3.8º are larger than the incidence angle (about 2º) of the electron beam; (B)-(E): The RHEED pattern and the corresponding sketched stage of the QD evolution

On the initial surface (Fig. 2/B), droplets with rounded shape are formed (Fig. 2/C). After this, a monocrystalline outer shell develops (due to the adsorbate effect, mentioned before). This is verified by the dotted RHEED pattern (Fig. 2/D). Thirdly, the crystallization process occurs, showing signs of the crystallographic planes inside the QD (Fig. 2/E). Here, the timing of the process of the RHEED sequence and the duration of each RHEED stages are also relevant. For many seconds, the dotted stage (Fig. 2/D) remains unchanged, but after that, the chevron tails develop relatively quickly (Fig. 2/E). The developing planes correspond to crystalline facets, verified by chevrons on the RHEED picture [18]. The evolution of the structure is sketched in the lower part of Fig. 2.

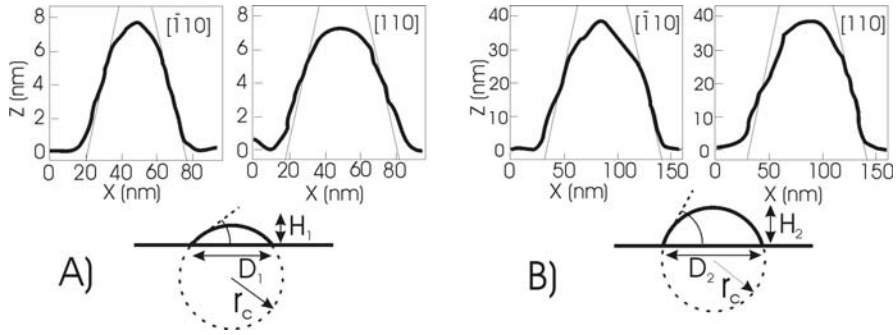


Figure 3

AFM line scans of single QDs along [1$\underline{1}$0] and [110] azimuth. (A) and (B) show typical line scans and the sketched droplets with parameters (high H, sole diameter D, critical radius r_c; the explanation see later) of the sample type *a* and *b*, respectively

In our earlier studies, the grown QDs can be classified two groups by their size and shape [18] (see Fig. 3). Here, we will show that shape of the QD depends on the shape of the droplet. Moreover, the size and the shape of the initial droplet is determined by the process of Ostwald ripening [30-36]. Furthermore, the shape of the droplets is also determined by the surface tension in the liquid aggregate and by the wetting properties at the interface.

The contact angle (the angle at which the tangent at a liquid/vapour interface meets the solid surface, called the three-phase line (TPL)) is characteristic of the surface. In practice, droplets (with mm size) are used for surface characterization [37-39]. Surface tension is caused by the intermolecular forces on surface of the liquid [40]. This tension is an effect within the surface layer of a liquid that causes that layer to behave like an elastic sheet with a thickness of approximately 10 nm [41]. Inside of this very thin sheet, the forces are independent of the length of the sheet. (Therefore, the behaviour of these forces differ fundamentally from the force in an elastic membrane.) At macroscopic scale, the interaction between forces across the three interfaces are independent of the dimension of the droplet, and therefore they can be used for surface characterization [38, 39].

These forces in the surface sheet are independent of the area and only depend on the thickness, therefore, of the number of interacting molecules in the unit area. (For example, in a free-standing membrane, the force is doubled; therefore, we must double the thickness as well in the calculations.) In reality, the dimensions of the droplet are much larger than the thickness of the surface sheet. If the dimension of the droplet decreases to the nano range, and the size becomes comparable with the dimension of the thickness of surface sheet (approx. 10 nm), then the shape of the droplet (the contact angle) changes. The tension also depends on the temperature. It decreases with increasing temperature [41], and it also decreases with increasing ambient pressure [42]. The surface wettability (and also the contact angle) depends very strongly on the properties of the substrate surface.

Molten Ga dropletes were investigated on a pure GaAs substrate in a hydrogen atmosphere [43]. At 852 °C, typical to liquid phase epitaxy, the contact angle is 28°. It was found that the temperature coefficient is 3.6°/100 °C [43]. At 250 and 200 °C the contact angles are thought to be around 50° and somewhat larger, respectively. In our case, the droplets are in a vacuum (pressure conditions are different); therefore, the contact angle can be expected to be somewhat different from the calculated one. Furthermore, it was found that on the AlGaAs substrate, the contact angle increases with the increased Al content [43]. Our substrate has x=0.3 Al content: therefore, the estimated contact angle is few degrees larger than 50°. It is evident that larger QDs originate from larger droplets (and smaller ones from smaller). The relation between density and size of QDs is governed by Ostwald ripening [30-36].

We will show later that the facet angle of the QDs depends on the contact angle of the initial droplet and, furthermore, that it is realistic to suppose that the larger droplet – similarly to the macroscopic case – has a contact angle slightly larger than 50°. It will be verified that the contact angle of the lower droplet is around 25°.

In our previous experiments, we found two distinct regimes for the QDs shapes (see Fig. 3.). When the Ga flux is low (F=0.025 ML/s, T=250 °C), the probability of the nucleation is low. With decreasing adatoms (and increasing temperature), the Ga atoms attach themselves to existing clusters rather than nucleate new ones. Therefore in this case we have larger QDs (average height n_1=32 nm, average base width d_1=110 nm) with lower density (n_1=4.4×10^8 cm^{-2}), and with a facet angle of about 55°. We observed in the experiment that the size and shape of the QDs did not differ greatly from each other. According to the experiment, it is probable that the contact angle is likely to be found at around 55° (somewhat larger than 50°). For higher Ga flux (F=0.19 ML/s, T=200 °C), the probability of nucleation is high. We observed smaller droplets (h_2=7 nm, d_2=60 nm) with higher density (n_2=1.2×10^{10} cm^{-2}). We also found a fundamental difference in the shapes of the QDs in two different growth regimes (sample types *a* and *b*, respectively). In the latter case, the side angle of the QDs was found to be 25° (where the contact angle of the droplets was also about 25°.) This proves that the quantity of the deposited Ga atoms is in correlation with the volume of the droplet, according to the results on the QD facets, shown previously [18].

Let us assume that the shape of the droplets is a segment of a sphere (see. Fig. 4/A), where R is radius of the sphere, D is a cord as shown in the figure, α is the contact angle, H is the height of the droplet, V is the volume of the droplet and n is the density of the droplets. The radius of the curvature can be calculated from the equation: $R=D/2sin\alpha$ or $R=H/(1-cos\alpha)$. For smaller droplets, the calculated radius from both formulae is the same, R = 68 nm and the contact angle is 26°. For the larger droplets, the radius – according to former calculations – is 71 nm with a contact angle 57°. The droplet volume can be calculated either from the spherical volume formula: $V=\pi h^2(R-H/3)$ or from the dot density: $V=(\Theta-1)/n$. The number of

Ga atoms in the droplets can be estimated by using both formulae. The atomic radius of Ga is 0.13 nm [44]. This corresponds closely to the half of the minimum bond length. The calculated atoms in the droplet are in order of one million. The atomic interaction in a large cluster (the atomic number $\geq 10^6$) is similar to in bulk material [45]. It is also known that the surface tension is caused by the attraction between the short radius intermolecular forces. This tension is a surface effect which behaves like a very thin (in order of the nm) elastic sheet [41].

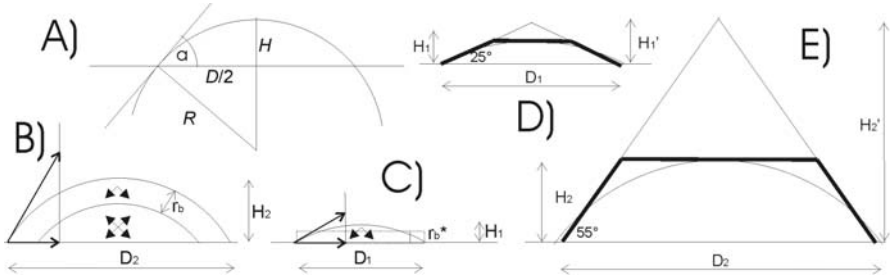


Figure 4

(A): schematic diagram of the droplet for the calculation of curvature; (B) and (C): Illustration of the effect of the contact angle and the size: for large droplet, the size is large enough; for a small droplet, the droplet size and the surface sheet dimensions are comparable. (D) and (E): Formation of truncated pyramid-like QDs (see text)

Now we can estimate the contact angle. The surface tension depends on the thickness of the surface sheet (see doubling effect). The contact angle dependence on the surface sheet can be seen in Fig. 4/B. In our calculation, we used a numerical value of r_b=10 nm [41]. For the big droplets, the volume is large enough for the surface sheet to be similar to that of the macroscopic case. The contact angle is over 50°, as shown earlier. This contact angle corresponds roughly to the side angle of the large QDs. For a small droplet, however, the situation is quite different. Here, the height of the droplet is comparable (even below) to the thickness of the surface sheet and the surface tension is proportional to the thickness of the sheet. The effective sheet thickness (r_{beff} calculated from the volume of the sheet) is approx. 6 nm, where the effective or average thickness is defined as the volume of the sphere segment normalized to the surface area of the interface between the droplet and the substrate (see Fig. 4/C). From here, the contact angle is around 30°. The contact angle corresponds roughly to the side angle of the small QD. The follow up phase in the QD production is the crystallization. The kinetics of the crystallization process is not yet known. We know that when the dotted diffraction picture of the transmission RHEED develops very quickly, it characterises the monocrystalline state. As we said before, the monocrystalline shell develops on the droplet surface due to the adsorbate effect [20]. The lateral formation of crystal surface is determined by the orientation of the new external surface element of the crystal in a defined crystal position of the melt surface adjacent to the TPL. The crystal surface coincides

approximately to the tangent to the phase boundary at a point on the TPL [46]. That is, the crystal facet of the QD corresponds to approximately the contact angle of the droplets. For a small droplet (a contact angle of less than 30°) the grown side-facets are around 25° after crystallization, corresponding approx. to (113) or (137) crystal planes (see Fig. 4/D). For a large droplet (a contact angle of more than 50°) the grown side-facet is around 55° after crystallization, corresponding to (111) crystal plane. The temporal evolution of the crystallization breaks up at one point, because the Ga atoms are all used up from the droplet; therefore, the QDs take up a truncated pyramid shape (see Fig. 4/E).

The TPL at the droplet edge initializes the crystallization. A number of crystallization seeds form at the same time in a number of equivalent places. The dotted RHEED picture indicates that all of the seeds are coherent oriented, so they have a monocrystalline character. The crystallization may occur uniformly distributed amongst these equivalent places. The liquid phase has short-range order lines and the droplet sizes are falling in this range. The liquid phase in the droplet can be described as a crystalline-like structure [45, 47]. We can conclude that the coalescence of the simultaneously appearing nucleation centers grow into a monocrystalline form and not into the usual polycrystalline form. When the surface area minimization takes place (between the droplet state and QD state), the crystal planes of the QDs must come near to the tangential planes of droplet [46]. If we assume a simultaneous crystallization process, than we come very close to the already described shapes for the QDs. The crystalized QDs are truncated, setting the sole width, the facet angle, the height and also the shape of the droplets.

The effect of the adsorbate and its segregation can strongly alter or even block the crystal growth locally and can lead to the rounding of crystal shapes [21, 22]. This effect explains the existence of the pure dotted RHEED picture during crystallization. On the RHEED picture, chevron tails form relatively quickly in the later stage of crystallization. This behaviour explains the phenomenon that the growing crystal facets explode the outer adsorbing shell, and instead of round crystal surface, in the result is a sharp crystal facet. The resulting nanocrystal corresponds to the Wulf construction [23-29].

3.2 Formation Kinetics of Quantum Rings

The QR evolution was tracked in-situ manner with the RHEED technique. In the initial stage, the perfect crystalline AlGaAs (001) surface causes sharp RHEED streaks. After the deposition of Ga, this picture becomes diffuse, due to the amorphous nature of the phase present on the surface. The annealing phase begins after the offering of an arsenic component with the pressure of 4×10^{-6} Torr, while the substrate temperature is 300 °C (sample type *c*). After the introduction of arsenic, some time is needed for the formation of the characteristic sharp pattern, representative of the crystalline structure (see Fig. 5/A). This shows that the liquid state on the surface stays for longer time and the material transport processes in it

propagate the formation of the QRs. This is different to the QD formations, where the RHEED pattern characteristic of the crystalline phase coincides with the opening of the arsenic cell [48]. The QR structures were ex-situ investigated with AFM. The perspective AFM image and the top view with line scans are shown in Figs. 5/B and C, respectively. The dimensions of the QRs were determined from individual line scans. Two typical adjacent AFM line scans are shown in the right side of Fig. 5/C. The density of the QRs was determined from the AFM pictures as 1.5×10^{9} cm^{-2}. It is observable that the middle of the QRs is located deeper than the original surface level. It can be shown from the AFM measurement that the shape and size of the QRs are near uniform, but we can observe small deviations from the average size and form. It has been observed frequently that the smaller diameter QRs have deeper cavities in the centre and opposite; that is, the larger diameter QRs have shallower cavities in the middle. (In the figures, the smaller and larger objects are labelled with "S" and "L", respectively.) Fig 5/C shows that the QRs are a little elongated (see top view), the phenomenon of which is based on the different binding properties in [110] and [1$\underline{1}$0] directions. Here, the phenomenon is neglected, because it does not influence our order of ideas, since the larger QR is larger in both lateral directions, and the smaller one is the opposite.

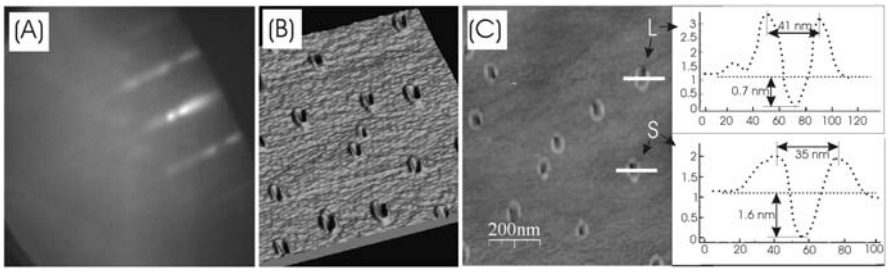


Figure 5

(A): The RHEED pattern of the QR structures; (B) Perspective AFM picture of sample surface with QRs; (C):Top view of the AFM image ith line scans (see text)

This interesting contradictory property of the QR aspect ratio can be explained as follows. During the deposition of the Ga, droplets form on the AlGaAs surface. It can be presumed that the larger droplets lead to the development of rings of larger diameter, and the smaller droplets to smaller rings. The intersection of the crystal surface with the droplet edge is the three-phase-line, which serves as initial place of crystallization [49, 50]. The three-phase-line of larger diameter attaches to the larger droplet and the one of smaller diameter attaches to the smaller droplet. It is known from the liquid phase epitaxy that thermal etching takes place at the Ga melt and AlGaAs surface [51]. This phenomenon was confirmed via analytical transmission electron microscopy in the case of droplet epitaxy [52]. The Ga melting can solve the arsenide molecules (e.g. GaAs). These arsenide molecules originate partly from the thermal etching of the crystal surface and partly from the

reaction of the external arsenic atoms. (When a Ga atom of the droplet meets an arsenic atom from the environment, they form a GaAs molecule.) These molecules causing thermal movement in the droplet can reach the three-phase-line, where the crystallization takes place. During the process of this solidification, a material transport takes place from the middle to the edge of the nanostructure. Thus, a circular crystalline phase is formed at the droplet edge. During the solidification process, the amount of Ga atoms in the droplets decreases, so the droplet size decreases, as well. The proposed kinetics of the formation is sketched in Fig. 6. During thermal etching, not only arsenic and gallium but also aluminium leaves from the AlGaAs surface, and is built into the lattice matched crystalline pair, which was shown earlier for QD [52]. We can suppose that the situation for QR is similar.

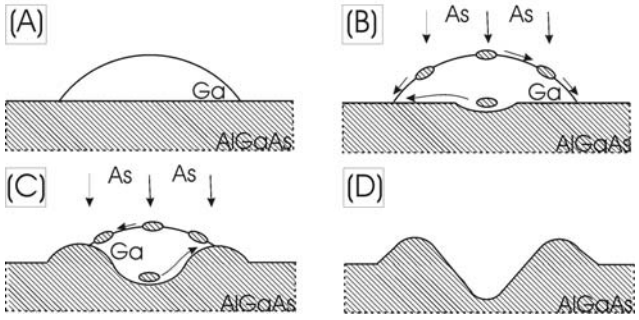


Figure 6
(A)-(D): Different states of the evolution of the droplet-epitaxial QR (see text)

It is known that the melting point decreases with the reduction in the particle size [53, 54]. Fig. 7/A shows the normalized melting curve for a metal vs. diameter of the particle. The figure shows that if the size is less than 50 nm, then the melting point depends very strongly on the size. In the nano range, the melting point is determined more strongly by the size than by bulk properties. The experimental melting curves for near spherical metal nano particles exhibit similarly shaped curves. For our case, this curve is gives direction for qualitative consideration. It means that the melting point in the larger and smaller Ga droplet can differ strongly. Fig. 7/B shows the solubility curves for different particle sizes. The figure shows that under the same temperature, the larger droplet has a lower saturating concentration and the smaller droplet has a higher one [55]. This means that the crystallization in the larger droplet will take place earlier that at a smaller arsenide concentration. The smaller droplet will crystallize later only at higher arsenide concentrations.

The temporal evaluation of the smaller and larger QR is shown in Fig. 7/C. In other words, in the larger droplets, the probability of the formation of crystallization seeds is higher; therefore, the crystallization takes place earlier, so less time is spent on material transportation, causing the development of the hole

in the middle. In the smaller droplets this probability is smaller; crystallization starts later, so more time is spent on deepening the middle of the ring. This process is influenced by other factors as well. As the melting temperature of the nanostructure drops with its diminishing size, a smaller droplet will stay longer in liquid state at the same temperature; therefore, more time can be spent on the central cavity formation. This diameter versus central depth relationship can only be considered on statistical grounds. The likely reason is that the forming process depends not only on the size of the droplets, but also on the arsenic molecules, whose presence is independent of the size of the droplet. Homogeneity of the distribution and the size of the nanostructures are quite uniform. We also monitored some deviations from the rule.

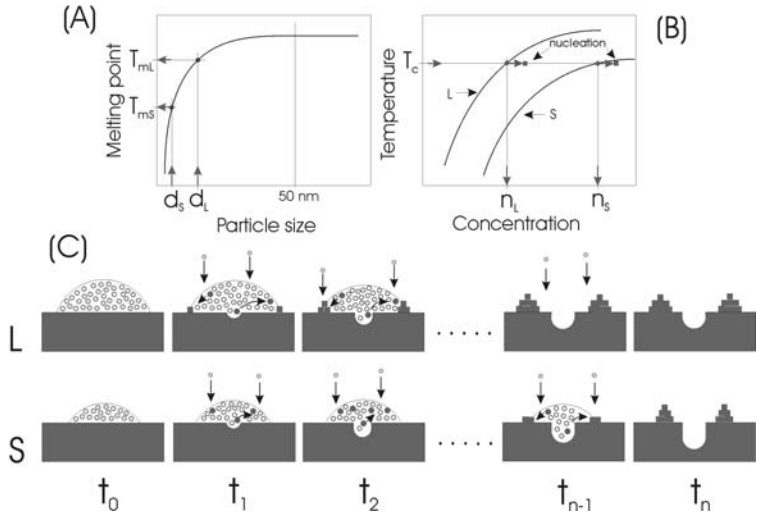


Figure 7

(A) The melting point dependence vs. particle diameter. If the size less than 50 nm then the dependence is very strong. (B): The solubility for different droplet sizes; (C): Temporal evolution of the QR for larg and for small droplet size

3.3 Formation Kinetics of Special Shaped Quantum Structures

Here we deal with the formation kinetics of particular shaped droplet epitaxially grown nanostructures. As is widely known, the properties in the nano region differ from the bulk properties. Generally, the nano-properties are unknown and only the tendency of the change is established, relative to the bulk properties. Usually the starting and the end states are known. The explanation of the final process comes from the coherent explanations of the component processes. These process explanations are coherent and consistent only when they describe the growth of a large number of different kinds of nano-structures. Only then do they become consistent.

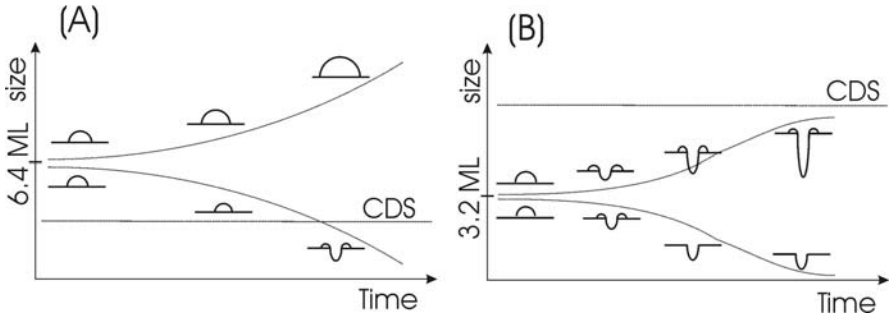


Figure 8

(A): Explanation of the temporal diferentiation and solution process in case of large amount of deposited Ga (6.4 ML), (B) Explanation of this process at small Ga amount (3.2 ML) (see text)

After the deposition of Ga, part of the deposited Ga will combine with the surface arsenic atoms and the rest will form droplets. In order to form droplets at this temperature, the Ga atoms must migrate (process *i*) on the GaAs and the Ga surfaces. It is well known that with diminishing size the melting temperature drops and saturation concentration increases. This describes our present Ga droplets well (process *ii*) (this is the dominant process in cases of a complete lack of or only small quantities of arsenic). It is also well known that the crystallization starts at the three-phase-line when the conditions became favourable. In our case, this starts in the line of the rim of the droplet (process *iii*). It is noted that the excess arsenic incorporates into the GaAs epitaxy when the temperature is low (about 300 °C or lower), creating stress in the lattice. (In presence of a larger quantity of arsenic, these are the dominant processes).

Process (*i*) allows for the phenomenon called Ostwald ripening [30-36]. At the same time, process (*ii*) causes the differentiation between QD-s, QR-s and NH-s. Due to this process, at the same temperature and during the same time, the hole which originates under the smaller droplets is deeper than that which is under the larger droplets (Fig. 8). (That is supposing the time duration is not so long asto run out of the material from the smaller droplets.) These findings are proved by the two experiments, where the deposited Ga quantity was 6.4 ML (Fig. 8/A) and 3.2 ML (Fig. 8/B), respectively. At a given temperature there is a critical droplet size (CDS) under which the solution begins. After the Ga deposition, droplets form, followed by the growth of the larger droplets at the expense of the smaller ones. When the critical size is reached, the substrate solution by the droplet begins. We start by investigating the case of Ga 6.4 ML [16]. During the experiment, the formation of small NH-s and large QD-s can be observed (sample type *d*) [16]. We can follow the process on Fig. 8/A, where a large quantity of Ga is deposited. The sizes of the droplets formed are above the CDS. After the deposition, the differentiation of the droplets begins. The smaller droplets reach critical size and start solving the substrate. This state is frozen via opening the arsenic cell.

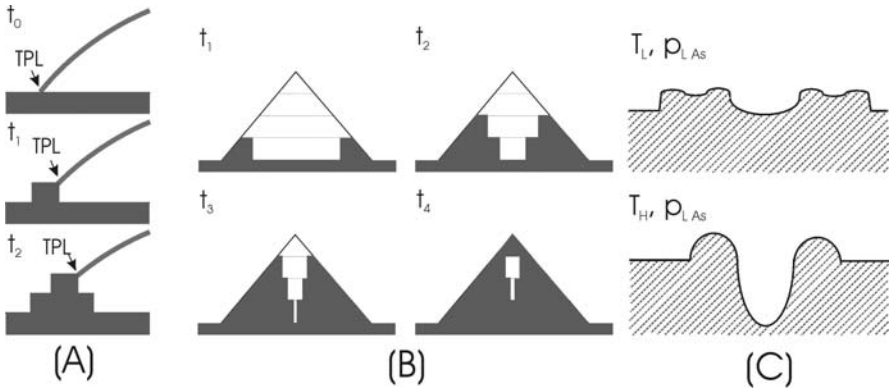


Figure 9

(A): Explanation of the crystal seed formation at the droplet rim; (B): Explanation of the formation of Ga inclusion inside of the QD; (C): upper and lower part show cross section of a DQR and NH, respectively

The second case is when the deposited Ga is 3.2 ML [15, 16]. Here, we can observe shallow NHs, with plane rims (without any bulge) and deep NHs surrounded by ring-like bulge formations (sample type *e*) [15, 17]. The explanation is given in Fig. 8/B. In this case, the quantity of the deposited Ga is small. The formed droplets are under the CDS; therefore the solution starts under the droplets. Under the small droplets the solution is faster, but the material is used up in a short time. The reduction in material is due first to the solution and second to the material's migration towards the larger droplets. After a short time, at the smaller droplets the solution stops, whilst it carries on further under the larger ones. The larger droplets will not be spent, and therefore the surrounding ring will freeze after the opening of the arsenic cell.

Due to the third process, the crystallization originates at the rim of the droplets. At a lower temperature the excess arsenic is incorporated into the lattice. The interstitial arsenic migrates towards the inside of the structure, reducing the lattice stress. A new interface is created in the inner and upper regions of the crystal centres. The growth of these centres propagate in those directions (Fig. 9/A). The arsenic is fed from the outer surface of the structure. It was observed that when the arsenic quantity was large, the QD had Ga inclusions (sample type *f*) [8]. Their origin is explained above. The growth in height is similar to that of the basic circle; so the speed of growth in height is a linear function. The speed of growth towards the inner regions accelerates because, assuming constant arsenic absorbing surface, the concentric area diminishes with the reduced radius going towards the centre (Fig. 9/B). The presence of the excess arsenic at a lower temperature is the explanation for the formation of the DQRs. The migration and the crystallization of the Ga would result in the creation of a larger ring (upper part of Fig. 9/C, sample (type *g*)) [13]. Instead, in addition to the original ring we end

up with a second ring, and this limiting effect on the size of the area is the result of the elimination of the lattice stress. The creation of DQRs can only be observed at a lower temperature. At a higher temperature DQRs cannot be created because the arsenic is not incorporated into the lattice, which causes the stress in the lattice (lower part of Fig. 9/C, sample (type *h*)) [17].

Conclusions

Here, we described the formation of several droplet-epitaxial nano-objects. In the first part, two separate regimes are investigated for the strain-free GaAs QD shape, using Ga droplet epitaxy for the growth. On one hand, the accurate facet angles are determined by the crystalline planes; on the other, the coarse angles of inclination of the QDs are determined by the contact angles of the droplets. A spherical approach was used for the describing shape of droplets. The volume and the shape were controlled with the data from AFM. The resulting contact angle was correlated with the surface tension dependence on sheet thickness (size effect). The shape of the QDs depends on the droplet volume. Larger QDs have a strongly truncated pyramid-like shape with side-facets 55°. Smaller QDs have a slightly truncated pyramid-like form with side-facets 25°, verified by AFM results. The correlation between the size and shape of the droplets and the QDs was also investigated. We showed that the shape of the droplet in the nano range depends on its size. We also showed that the nano structure during crystallization has a monocrystalline character and a rounded outer surface due to the effect of adsorbate. This is in an agreement with the dotted RHEED picture. The chevron tails which form in the late stage of crystallization process can be explained by a break in the adsorbate shell of the structure. The results correspond with the Ostwald ripening and with the Wulf construction.

In the second part of this paper we attempt to explain the kinetics of the epitaxial growth of QRs. What triggered the investigation was the realization of one peculiar characteristic feature of QRs during experimentation. That is, often the middle of the larger rings is shallower than that of the smaller ones. The arsenic molecules arrive to the surface in random distribution and therefore the above phenomena is superimposed on random noise. The cause of the phenomenon is the dependence of the melting point and the saturation concentration on the size of the nanostructure in this size range. We are confident that with this interpretation, we approach the correct explanation of the kinetics of the droplet epitaxy.

In the last part, we explain the formation kinetics of a few special-shaped nano-structures. One of them is the QD with Ga inclusion. The explanation is based on the excess arsenic content in the grown crystal at low temperature. Temporal differentiation was observed during the growth of NHs. We introduced the CDS to explain this phenomenon.

Acknowledgement

This work was supported by Hungarian Scientific Research Foundation (OTKA K75735 and K77331) and by research grant of Óbuda University.

References

[1] D. Leonard, M. Krisnamorthy, C. M. Reaves, S. P. Denbaas, P. M. Petroff; Appl. Phys. Lett. 63 (1993) 3203

[2] V. Bessler-Hill, S. Varma, A. Lorke, B. Z. Nosho, P. M. Petroff; Phys. Rev. Lett. 74 (1995) 3209

[3] N. Koguchi, S. Takahashi, T. J. Chikyow; J. Cryst. Growth 111 (1991)

[4] N. Koguchi, K. Ishige; Jpn. J. Appl. Phys. 32 (1993) 2052

[5] T. Mano, K. Watanabe, S. Tsukamoto, H. Fujikoa, M. Oshima, N. Koguchi; Jpn. J. Appl. Phys. 38 (1999) L1009

[6] J. M. Lee, D. H. Kim, H. Hong, J. C. Woo, S. J. Park; J. Cryst. Growth 212 (2000) 67

[7] Ch. Heyn, A. Stemmann, A. Schramm, H. Welsch, W. Hansen, Á. Nemcsics; Appl. Phys. Lett. 90 (2007) 203105

[8] T. Mano, K. Mitsuishi, Y. Nakayama, T. Noda, K. Sakoda; Appl. Surf. Sci. 254 (2008) 7770

[9] Á. Nemcsics, L. Tóth, L. Dobos, Ch Heyn, A. Stemmann, A. Schramm, H. Welsch, W. Hansen; Superlatt. Microstr. 48 (2010) 351

[10] T. Mano, N. Koguchi; J. Cryst. Growth 278 (2005) 108

[11] T. Kuroda, T. Mano, T. Ochiai, S. Sanguinetti, K. Sakoda, G. Kido, N. Koguchi; Phys. Rev. B 72 (2005) 205301

[12] Á. Nemcsics, Ch. Heyn, A. Stemmann, A. Schramm, H. Welsch, W. Hansen; Mat. Sci. Eng. B 165 (2009) 118

[13] S. Sanguinetti, M. Abbarchi, A. Vinattieri, M. Zamfirescu, M. Gurioli, T. Mano, T. Kuroda, N. Koguchi; Phys. Rev. B 77 (2008) 125404

[14] T. Kuroda, T. Mano, T. Ochiai, S. Sanguinetti, T. Noda, K. Kuroda, K. Sakoda, G. Kido, N. Koguchi; Physica E 32 (2006) 46

[15] Ch. Heyn, A. Stemmann, W. Hansen; J. Cryst. Growth 311 (2009) 1839

[16] Ch. Heyn; Phys. Rev. B 83 (2011) 165302

[17] Zh. M. Wang, B. L. Liang, K. A. Sablon, G. J. Salamo; Appl. Phys. Lett. 90 (2007) 113120

[18] Ch. Heyn, A. Stemmann, A. Schramm, H. Welsch, W. Hansen, Á. Nemcsics, Appl. Phys. Lett. **90**, 203105 (2007)

[19] Heyn Ch, Stemmann A, Schramm A, Welsch H, Hansen W, Nemcsics Á. Phys. Rev. B 2007; 76: 075317

[20] M. Kalft, G. Comsa, T. Michely; Phys. Rev. B 81, 1255 (1998)

[21] P. B. Barna; Proc. of Int. Summer School on Diagnostics and Application of Thin Films, Chulum u Trebone, Czechoslovakia (1991) pp. 295-310

[22] P. B. Barna, F. M. Reicha; Proc. of 8th Int. Vac. Congr., Cannes, France (1980) Vol. 1, pp. 165-168

[23] C. Herring; Phys. Rev. **82**, 87 (1951)

[24] P. Müller, R. Kern; Surf. Sci. **457**, 229 (2000)

[25] P. Müller, R. Kern; Appl. Surf. Sci. **162-163**, 133 (2000)

[26] P. Müller, R. Kern; Appl. Surf. Sci. **164**, 68 (2000)

[27] J. J. Métois, P. Müller; Surf. Sci. **548**, 13 (2004)

[28] M. Degawa, E. D. Williams; Surf. Sci. **595**, 87 (2005)

[29] M. Degawa, F. Szalma, E. D. Williams; Surf. Sci. **583**, 126 (2005)

[30] W. Ostwald; Z. Phys. Chem. 34, 495 (1900)

[31] M. Zinke-Allmang, L. C. Feldman, S. Nakahara; Appl. Phys. Lett. 51, 975 (1987)

[32] M. Zinke-Allmang, L. C. Feldman, W. van Saaloos; Phys. Rev. Lett. 68, 2358 (1992)

[33] G. Z. Pan, K. N. Tu; Appl. Phys. Lett. 68, 1654 (1996)

[34] G. R. Carlow, M. Zinke-Allmang; Phys. Rev. Lett. 78, 4601 (1997)

[35] K. Shorlin, S. Krylov, M. Zinke-Allmang; Physica A 261, 248 (1998)

[36] A. Raab, G. Springholz; Appl. Phys. Lett. 77, 2991 (2000)

[37] E. Z. Luo, Q. Cai, W. F. Chung, M. S. Altman, Appl. Surf. Sci. 92, 331 (1996)

[38] N. Kaiser, A. Cröll, F. R. Szofran, S. D. Cobb, K. W. Benz, J. Cryst. Growth 231, 448 (2001)

[39] A. Cröll, N. Salk, F. R. Szofran, S. D. Cobb, M. Volz, J. Cryst. Growth 242, 45 (2002)

[40] F. R. de Boer, R. Boom, W. C. M. Mattens, A. R. Midema, A. K. Niessen, Cohesion in Metals (North-Holland, New York, 1988) p. 127

[41] Á. Budó, Experimental Physics, Vol. I. (NTK-Publ., Budapest, 1997), p. 235

[42] X. Huang, S. Togawa, S.-I. Chung, K. Terashima, S. Kimura, J. Cryst. Growth **156**, 52 (1995)

[43] U. König, W. Keck, J. El.chem. Soc. **130**, 685 (1983)

[44] B. Jones, J. Chem. Phys. **41**, 3199 (1964)

[45] R. Popescu, E. Müller, M. Wanner, D. Gerthasen, Phys. Rev B **76**, 235411 (2007)

[46] G. A. Satukin, J. Cryst. Growth **255**, 170 (2003)

[47] N. Eustathopoulos, B. Drevet, E. Ricci, J. Crys. Growth **191**, 268 (1998)

[48] Nemcsics Á, Heyn Ch, Stemmann A, Schramm A, Welsch H, Hansen W. Mat. Sci. Eng. B 1999; 165: 118

[49] SatukinGA. J. Cryst. Growth 2003; 255: 170

[50] Nemcsics Á, Tóth L, Dobos L, Stemmann A. Microel. Reliab. 2011; 51: 927

[51] E. Lendvay, T. Görög, V. Rakovics. J. Cryst. Growth 1985; 72: 616

[52] Nemcsics Á, Tóth L, Dobos L, Heyn Ch, Stemmann A, Schramm A, Welsch H, Hansen W. Superlatt. Microstr. 2010; 48: 351

[53] Jiang A, Awasthi N, Kolmogorov AN, Setyawan W, Borjesson A, Bolton K, Harutyunyan AR, Curtarolo S. Phys. Rev. B 2007; 75: 205426

[54] Sun J, Simon SL. Therochimica Acta 2007; 463: 32

[55] Wautelet M, Beljonne D, Brédas JL, Cornil J, Lazzaroni R, Lecère P, Alexanre M, Gillis P, Gossuin Y, Muller R, Ouakssim A, Roch A, Duviver D, Robert J, Gouttebaron R, Hecq M, Monteverde F; Nanotechnologie, Oldenbourg Verlag München, 2008

Factors Limiting Controlling of an Inverted Pendulum

Tobiáš Lazar, Peter Pástor

Department of Avionics
Faculty of Aeronautics
Technical University of Košice
Rampová 7, 041 21 Košice, Slovakia
E-mail: tobias.lazar@tuke.sk, pastor_peto@yahoo.com

Abstract: The aim of this paper is to show the limitation during an inverted pendulum control process. Assume the control signal and its derivate are limited. The goal is to find the maximum permissible value of the θ angle and state if this value can be determined only by symbolical calculation by using Maple software. This maximum value must guarantee the stability of whole system and the quality of the transient process. The nonlinear mathematical model of the inverted pendulum implemented in Simulink is utilized for result verification. A detailed description of these limitations is important for the application of advanced control methods based on expert knowledge to aircraft equipped with a thrust vectoring nozzles system.

Keywords: inverted pendulum; transfer function; nonlinear analyses; maple

1 Introduction

An inverted pendulum is an inherently unstable system. This system approximates the dynamics of a rocket immediately after lift-off, or dynamics of a thrust vectored aircraft in unstable flight regimes in negligible small dynamic pressure conditions [2]. Assume the force for the inverted pendulum control represents the force generated by a thrust vectoring nozzles system. The nozzle deflection is limited up to ± 20 deg, the rate of deflection is limited up to ±60 deg/sec and the nozzle dynamics is described by 2nd order transfer function, similarly as in the publication [1]:

$$\frac{400}{s^2 + 40s + 400} \tag{1}$$

The dynamics of the pendulum is given by following nonlinear differential equations system [7]:

$$\left(M+m\right)\frac{d^2x}{dt^2}+ml\frac{d^2\theta}{dt^2}\cos\theta-ml\left(\frac{d\theta}{dt}\right)^2=u \tag{2}$$

$$\left(J+ml^2\right)\frac{d^2\theta}{dt^2}=-ml\frac{d^2x}{dt^2}\cos\theta+mgl\sin\theta \tag{3}$$

where M – cart mass (in this case it can be neglected), m – pendulum mass ($m=15180\ kg$), l – length to the pendulum centre of gravity ($l=5{,}4\ m$), J – moment of inertia of the pendulum ($J=4.2138\cdot10^5\ kg\cdot m^2$), g – gravity ($g=9.81 m\cdot s^{-2}$), θ – the angle between pendulum and vertical axes [3].

The θ angle transfer function can be obtained after linearization of the system described by equations (2), (3):

$$\frac{\theta(s)}{U(s)}=\frac{K}{s^2+\omega_0^2}=\frac{-\frac{l}{J}}{s^2-g\frac{ml}{J}}=\frac{-1,2815\cdot10^{-5}}{s^2-1,90836} \tag{4}$$

The algorithm for pendulum control is given by following control law:

$$F(s)=sD\theta(s)+P\theta(s)+\frac{I}{s}\left[\theta(s)-\theta_z(s)\right] \tag{5}$$

where $F(s)=U(s)$ – inverted pendulum control signal, P – proportional coefficient, I – integral coefficient, D – derivative coefficient. The control system structure is depicted in Figure 1.

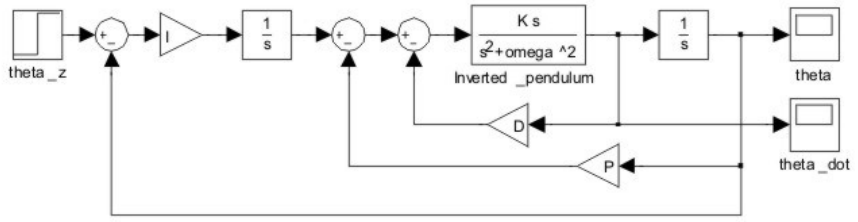


Figure 1
Control system structure with inverted pendulum transfer function

The final transfer function of the system shown in Figure 1 is:

$$\frac{KI}{s^3+KDs^2+\left(KP+\omega_0^2\right)s+KI} \tag{6}$$

2 Control Signal Limitation in Steady State

Utilize the equation (3) for maximal θ angle computation. Condition θ=*const* is valid for steady state. If θ=*const*, its derivate is zero and its second derivate is also zero. The following equation can be obtained by solving equation (3):

$$ml\frac{d^2x}{dt^2}\cos\theta = mgl\sin\theta \tag{7}$$

Expression $m\frac{d^2x}{dt^2}$ represents the control signal, the maximum value of which is given by: $F_{max}=Tsin\varphi$ [6], where φ – the angle of deflection of vectored nozzle. Assume the thrust and aircraft's weight are equal ($T=G=mg$). It is possible to transform equation (7) to get the following equation:

$$mg\sin\varphi\cos\theta = mg\sin\theta \tag{8}$$

Divide equation (8) by expression $cos\theta$:

$$\sin\varphi = \frac{\sin\theta}{\cos\theta} = tg\theta \tag{9}$$

The condition (10) for maximum value of θ angle in steady state has been obtained by solving equation (9):

$$\theta_{\max} = arctg\left(\sin\varphi_{\max}\right) \tag{10}$$

3 Limitation during Transient Process

Inverted pendulum control signal is denoted as *Z(s)* and is depicted in the structure shown in Figure 2. This structure can be utilized for Z(s) transfer function calculation.

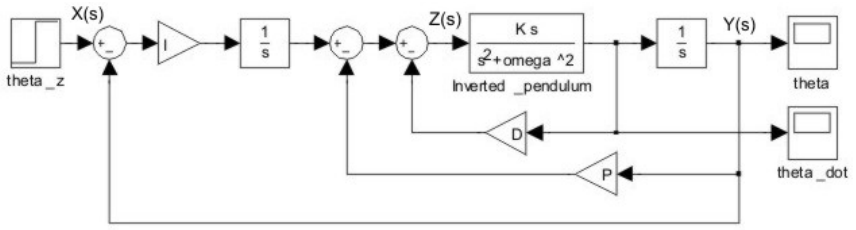


Figure 2
Control system structure with signal's description

The following equation is valid for *Z(s)*:

$$\frac{I}{s}\left[X(s)-\frac{K}{s^2+\omega_0^2}Z(s)\right]-\frac{KP}{s^2+\omega_0^2}Z(s)-\frac{KDs}{s^2+\omega_0^2}Z(s)=Z(s) \tag{11}$$

Solve the equation (11) and place the expression involving *Z(s)* on the right side:

$$\frac{I}{s}X(s)=Z(s)+\frac{KI}{s\left(s^2+\omega_0^2\right)}Z(s)+\frac{KP}{s^2+\omega_0^2}Z(s)+\frac{KDs}{s^2+\omega_0^2}Z(s) \tag{12}$$

Z(s) transfer function can be calculated from the previous equation:

$$\frac{Z(s)}{X(s)}=\frac{I\left(s^2+\omega_0^2\right)}{s^3+KDs^2+\left(KP+\omega_0^2\right)s+KI} \tag{13}$$

Denominators of transfer functions (6) and (13) are equal and represent the poles of the transfer function and their values guarantee whole system stability and transient process quality. Because the proportional, derivative and integral coefficients influence the poles' placement, it is necessary to select optimal values. The 3rd order polynomials in denominator of transfer functions (6) and (13) are the same. Assume that the 3rd order polynomial has one real root and two complex conjugate roots:

$$\left(s^2+2\xi\omega_Z s+\omega_Z^2\right)\left(s+\omega_Z\right) \tag{14}$$

where ω_Z is the desired natural frequency of the system and ξ is the desired system damping. Apply convolution operations to compute the product of polynomial in equation (14) to obtain the generalized 3rd order polynomial form:

$$s^3+\left(2\xi+1\right)\omega_Z s^2+\left(2\xi+1\right)\omega_Z^2 s+\omega_Z^3 \tag{15}$$

Substitute the denominator of transfer function (13) by the generalized 3rd order polynomial given by equation (15):

$$\frac{Z(s)}{X(s)}=\frac{I\left(s^2+\omega_0^2\right)}{s^3+\left(2\xi+1\right)\omega_Z s^2+\left(2\xi+1\right)\omega_Z^2 s+\omega_Z^3} \tag{16}$$

Transfer function (16) must be transformed into time domain by applying the inverse Laplace transformation for maximum positive and negative values determination. Maple software is used to provide this transformation [4]. It is possible to find a time function of equation (16), but this function is complicated for further symbolical analyses. State the damping value of the system as $\xi=1$ and substitute this value into equation (16):

$$\frac{Z(s)}{X(s)} = \frac{I\left(s^2 + \omega_0^2\right)}{s^3 + 3\omega_Z s^2 + 3\omega_Z^2 s + \omega_Z^3} \tag{17}$$

Polynomial of equation (17) has a triple root and is relatively simple for further symbolical analyses and represents the ideal transient process with acceptable quality. The optimal coefficient of the PID regulator can be found by comparing denominators of transfer functions (6), (17):

$$P = -\left(3\omega_Z^2 \frac{J}{l} + mg\right) \quad \left[kgms^{-2}\right] \tag{18}$$

$$I = -\frac{J}{l}\omega_Z^3 \quad \left[kgms^{-3}\right] \tag{19}$$

$$D = -3\omega_Z \frac{J}{l} \quad \left[kgms^{-1}\right] \tag{20}$$

The derivative of the control signal in time domain can be obtained by applying inverse Laplace transform to equation (17):

$$z'(t) = \frac{1}{2} I e^{-\omega_Z t}\left[t^2\left(\omega_Z^2 + \omega_0^2\right) - 4\omega_Z t + 2\right] \tag{21}$$

Control signal step response in 's' domain is given by following equation:

$$\frac{Z(s)}{X(s)} = \frac{I\left(s^2 + \omega_0^2\right)}{s\left(s^3 + 3\omega_Z s^2 + 3\omega_Z^2 s + \omega_Z^3\right)} \tag{22}$$

The control signal in time domain can be obtained again by applying the inverse Laplace transform to equation (22) and is described by the following equation:

$$z(t) = \frac{I}{\omega_Z^3}\left\{\omega_0^2 - \left[\frac{t^2}{2}\left(\omega_Z^4 + \omega_Z^2\omega_0^2\right) + t\left(\omega_Z\omega_0^2 - \omega_Z^2\right) + \omega_0^2\right]\right\} \tag{23}$$

In Figure 3 is shown the inverted pendulum's control signal step response given by equation (23). Value ω_0^2 is given in transfer function (4) and desired natural frequency value has been selected ($\omega_Z=2$).

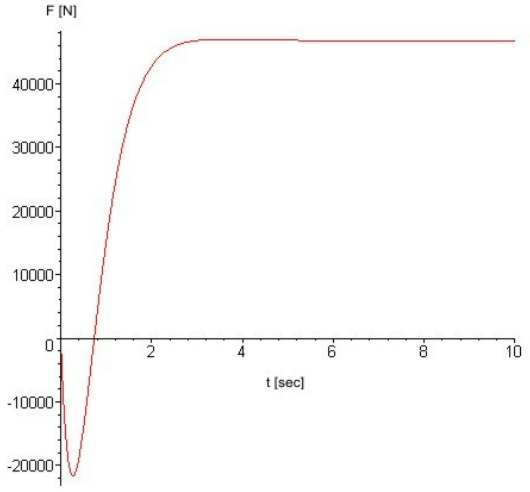


Figure 3
Control signal time response

The extreme value theorem states that if a function f is defined on a closed interval [a,b] (or any closed and bounded set) and is continuous, then the function attains its maximum, i.e. there exists c Є [a,b] with f(c) ≥ f(x) for all x Є [a,b]. The same is true for the minimum of f. The derivative of function f in c is zero. Equation (21) represents the derivative of the control signal. The equation has two roots:

$$t_{1,2} = \frac{2\omega_Z \pm \sqrt{2\left(\omega_Z^2 - \omega_0^2\right)}}{\omega_Z^2 + \omega_0^2} \tag{24}$$

The function described by equation (23) reaches its maximum positive value in time t_1 (t_1=*3.56 s*) and its maximum negative value in time t_2 (t_2=*0.27s*). It can be observed in Figure 3. Equations (25) and (26) represent maximum positive and negative values of control signal and are gained by substituting (24) into (23):

$$F_{\max,t_1} = \frac{-I}{\omega_Z^3}\left[\left(\omega_0^2 + \omega_Z^2 + \omega_Z\sqrt{2\omega_Z^2 - 2\omega_0^2}\right)e^{\Omega_1} - \omega_0^2\right] \tag{25}$$

$$F_{\max,t_2} = \frac{I}{\omega_Z^3}\left[\omega_0^2 - \left(\omega_0^2 + \omega_Z^2 - \omega_Z\sqrt{2\omega_Z^2 - 2\omega_0^2}\right)e^{\Omega_2}\right] \tag{26},$$

where Ω_1 and Ω_2 are given by equation (27), (28) respectively:

$$\Omega_1 = -\frac{\omega_Z\left(2\omega_Z + \sqrt{2\omega_Z^2 + 2\omega_0^2}\right)}{\omega_Z^2 + \omega_0^2} \tag{27}$$

$$\Omega_2 = \frac{\omega_Z \left(-2\omega_Z + \sqrt{2\omega_Z^2 - 2\omega_0^2} \right)}{\omega_Z^2 + \omega_0^2} \tag{28}$$

Maximum force is transformed into maximum angle by using following assumption:

$$T_{\max} \sin \varphi_{\max} = \theta_{\max} F_{\max} \Rightarrow \theta_{\max} = \frac{T_{\max} \sin \varphi_{\max}}{F_{\max}} \tag{29}$$

The maximum θ angle value for the desired frequency ω_Z can be calculated by utilizing equation (29). θ_{max} values are depicted in Figure 4.

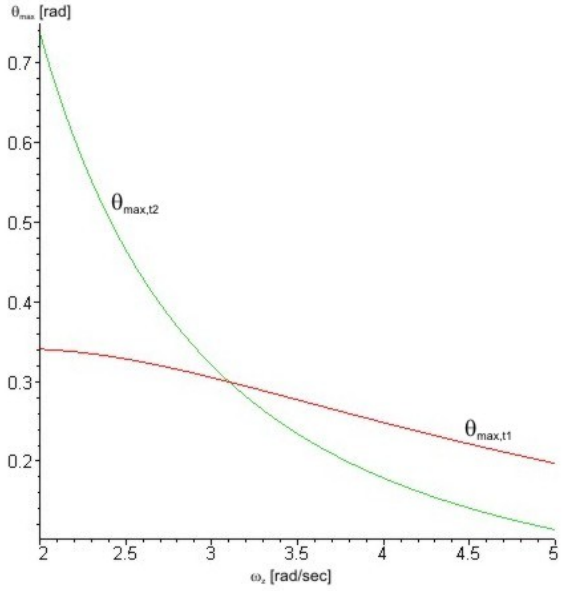


Figure 4
Maximum θ angle values depicted as a function of the desired natural frequency value

4 Limitation Given by Vectoring Nozzle Deflection Rate

The derivative of the control signal given by equation (21) is depicted in Figure 5.

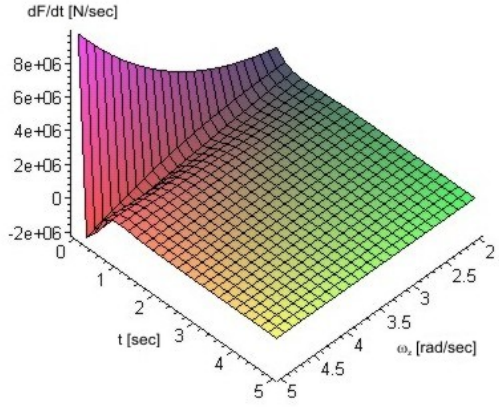


Figure 5
Control signal derivation depicted in 3-dimensional graph for ω_Z values in region from 2 to 5 rad/sec

It can be observed in Figure 5, that the control signal derivative reaches its maximum value in time *t=0*. The following expression can be utilized for maximum θ angle computation:

$$\theta_{\max} = \frac{\frac{dF_{\max}}{dt}}{z'(0)} \tag{30}$$

The maximum derivative of the control signal can be determined by applying the derivative operation to the right side of equation (29):

$$\frac{dF_{\max}}{dt} = \frac{d}{dt}\left(T_{\max}\sin\varphi\right) = T_{\max}\frac{d\varphi}{dt}\cos\varphi \tag{31}$$

Assume in time t=0 the nozzle deflection is zero so $\cos\varphi=1$. The maximum nozzle deflection rate is 60 deg/sec (approximately $\pi/3$ *rad/sec*) and the maximum thrust is supposed to be constant ($T_{max}=148916N$):

$$T_{\max}\frac{d\varphi}{dt}\cos\varphi \approx T_{\max}\frac{d\varphi}{dt} = 155945\ N/s \tag{32}$$

5 Nonlinear Analyses

The structure of the system used for nonlinear analysis is shown in Figure 6 and consists of two main blocks. Block 'vectored_nozzles' describes the vectored nozzles together with their dynamics and limitations mentioned in the introduction of this paper. The nonlinear mathematical model of inverted pendulum given by equations (33) and (34) is implemented in block 'Inverted pendulum'.

$$\left[(M+m)-\frac{(ml)^2\cos^2\theta}{J+ml^2}\right]\frac{d^2x}{dt^2}=ml\left(\frac{d\theta}{dt}\right)^2-\frac{(ml)^2 g\sin\theta\cos\theta}{J+ml^2}+u \tag{33}$$

$$\left[\left(J+ml^2\right)-\frac{(ml)^2\cos^2\theta}{M+m}\right]\frac{d^2\theta}{dt^2}=mlg\sin\theta-\frac{ml\cos\theta}{M+m}\left[ml\left(\frac{d\theta}{dt}\right)^2+u\right] \tag{34}$$

These equations are based on equations (2) and (3) that have to be rewritten for algebraic loop elimination [2]:

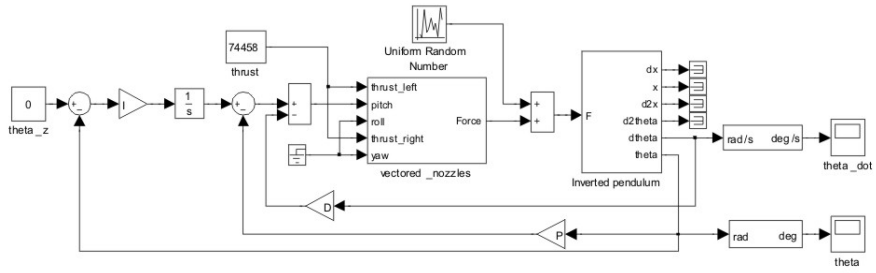


Figure 6
Structure used for nonlinear analyses

It can be seen in Figure 6 that the force generated by the vectoring system is controlled by the pitch command. The coefficients of the PID regulator can be calculated by dividing equations (18), (19) and (20) by the maximum thrust value ($T_{max}=148916N$). It is possible to consider this simplification only for small angle (up to *20 deg*). The following m-file was utilized for coefficients' calculation:

```
m=15180;%[m] mass of the aircraft
J=4.2138e5;%[kg*m^2] moment of inertia
l=5.4;%[m] CG position
T=148916;%[N] thrust
g=9.81;%[m*s^2] gravity
omega=2;%desired natural frequency value
P=-(3*omega^2*(J/l)+m*g)/T;%proportional coefficient
I=-(J*omega^3)/(l*T);%integral coefficient
D=-(3*omega*J)/(l*T);%derivate coefficient
```

The natural frequency desired value is shown in the first column of Table 1. The values in 2nd and 3rd column are depicted in Figure 4. The values in the 4th column are the minimum of the values calculated according to equations (10), (25) and (26). The values obtained from nonlinear analyses when the rate limitation and nozzle dynamics has not been assumed are in the 5th column. In the 6th column are values computed calculated according to equation (30) in Chapter 4. The values obtained from nonlinear analyses with rate limitation and nozzle dynamics are in the last column of Table 1.

Table 1

ω_Z	θ_{maxt1}	θ_{maxt2}	$\theta(2,3)$	θ_{max}	$\theta(4)$	θ_{max}
2	0.341	0.737	0.3295	0.324	0.2498	0.327
2.5	0.329	0.465	0.328	0.313(0.316)	0.128	0.279 (0.297)
3	0.305	0.321	0.305	0.298(0.303)	0.074	0.144 (0.178)
3.5	0.277	0.234	0.234	0.276(0.287)	0.047	0.081(0.093)
4	0.248	0.179	0.179	0.245(0.261)	0.031	0.042 (0.058)
4.5	0.221	0.141	0.141	0.211(0.231)	0.022	0.037
5	0.197	0.114	0.114	0.181(0.203)	0.016	0.023

The values obtained from nonlinear simulation are in the 5th and 7th column of Table 1. Both values delineate maximum value of given θ angle of the system in stable conditions but transient process for the first values is with acceptable quality (Figure 7) and transient process for the values in brackets is with poor quality (Figure 8). The desired frequency for both responses was $\omega_Z=4$.

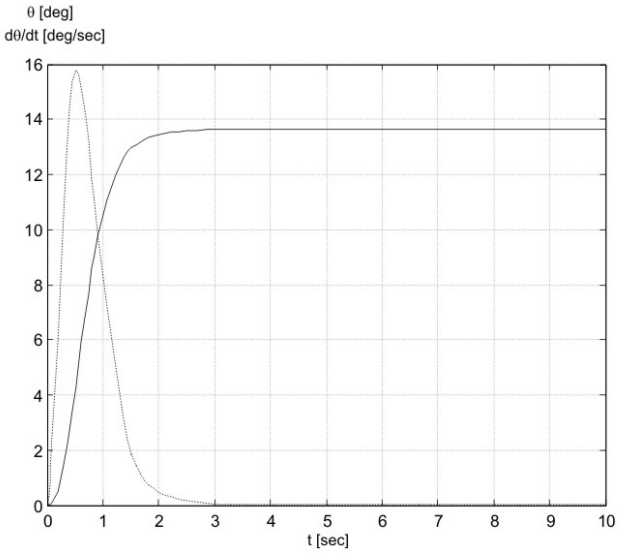


Figure 7

θ angle (solid line) and rate (dotted line) time response with acceptable quality of the transient process

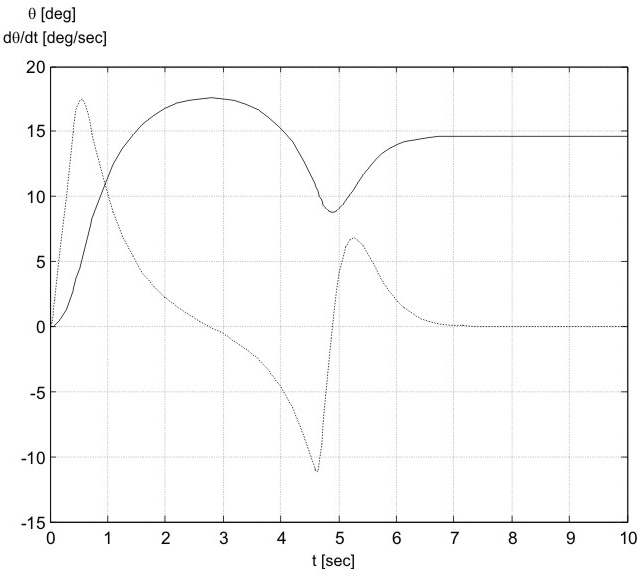


Figure 8

θ angle (solid line) and rate (dotted line) time response with poor quality of the transient process

Time response in Figure 8 converges, but the observable oscillations increase the settling time.

From Table 1, it is visible that the calculated values approximately describe the limiting conditions. This behaviour of the system is explained in Figure 9b where the time response of the control signal is depicted. It can be observed that the control signal reaches its maximum value. The oscillations are observed exactly in time when control signal reaches its maximum value.

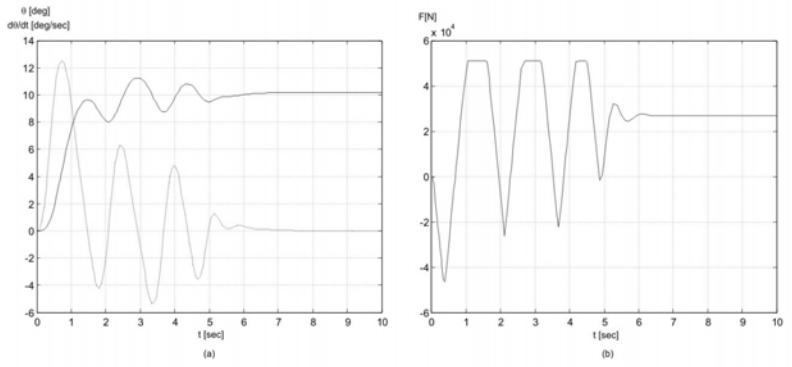


Figure 9

θ angle (solid line), rate (dotted line) and control signal time response

If the given θ angle value does not exceed the limiting conditions, calculated according to the procedure shown in Chapters 2 and 3, the system's time response will certainly be stable with required transient process quality. It is necessary to perform an experiment (e.g. nonlinear simulation) for accurate marginal θ angle value determination. The dynamic properties of the nozzle are much more limiting for the higher desired natural frequency (approximately from $\omega_Z=4$). It can be seen by comparing limiting θ angle values in the 5th and in the last column of Table 1.

Conclusions

The possibility to symbolically calculate limitation by using linear analyses and Maple software was shown in this paper. This procedure is appropriate for relatively simple transfer functions and the calculated results only approximately describe the limiting conditions, but do guarantee the system stability. It was shown that it is necessary to perform an experiment for marginal value determination. The obtained values provide a better idea about inverted pendulum dynamics and all factors considered for successful realization of the control system. These facts are also expected to be utilized during application of advanced control methods based on expert knowledge into the inherently unstable systems.

Acknowledgement

This work was supported by projects: KEGA 001-010TUKE-4/2010– Use of intelligent methods in modelling and control of aircraft engines in education.

References

[1] Adams, Richard J. – Buffington, James M. – Sparks, Andrew G. – Banda, Siva S.: Robust Multivariable Flight Control, London: Springer-Verlag, 1994. ISBN 3-540-19906-3

[2] Dabney, J. B – Harman, T. L: Mastering Simulink, Pearson Prentice Hall, 2004, ISBN 0-13-142477-7

[3] Lazar, T. – Adamčík, F. – Labun, J.: Modelovanie vlastností a riadenia lietadiel, Technická Univerzita v Košiciach – Letecká Fakulta, 2007. ISBN 978 80 8073 8396 [in Slovak]

[4] Maple help – a part of Maple software

[5] Škrášek, Josef – Tichý, Zdeněk: Základy aplikované matematiky II, Praha: SNTL, 1986 [in Czech]

[6] Taranenko, V. T.: Dinamika samoleta s vertikaľnym vzletom i posadkoj, Moskva: Mašinostrojenije, 1978 [in Russian]

[7] http://www.profjrwhite.com/system_dynamics/sdyn/s7/s7invp1 /s7invp1.html#equations

Model Reference Fuzzy Control of Nonlinear Dynamical Systems Using an Optimal Observer

Mojtaba Ahmadieh Khanesar, Mohammad Teshnehlab

Department of Control
Faculty of Electrical and Computer Engineering
K. N. Toosi University of Tech.
Seyyed Khandan, Tehran, Iran
ahmadieh@ee.kntu.ac.ir

Okyay Kaynak

Department of Electrical and Electronic Engineering
Bogazici University
Bebek, 80815 Istanbul, Turkey
okyay.kaynak@boun.edu.tr

Abstract: This paper proposes a novel indirect model reference fuzzy control approach for nonlinear systems, expressed in the form of a Takagi Sugeno (TS) fuzzy model based on an optimal observer. In contrast to what is seen in the literature on adaptive observer-based TS fuzzy control systems, the proposed method is capable of tracking a reference signal rather than just regulation. Additionally the proposed algorithm benefits from an adaptation algorithm which estimates the parameters of observer optimally. The stability analysis of the adaptation law and the controller is done using an appropriate Lyapunov function. The proposed method is then simulated on the control of Chua's circuit and it is shown that it is capable of controlling this chaotic system with high performance.

Keywords: fuzzy control; model reference fuzzy control; observer design

1 Introduction

Fuzzy controllers can be viable alternatives to classical controllers when there are experienced human operators who can provide qualitative control rules in terms of vague sentences. Although fuzzy controllers have been successfully used in many industrial applications, in cases when some adaptation is required, there may not be enough expert knowledge to tune the parameters of the controller. This has motivated the design of adaptive fuzzy controllers which can learn from data and

one of the early suggestions in this respect is the approach described in [1], named the linguistic self-organizing controller (SOC).Such early fuzzy adaptive systems have suffered from the lack of stability analysis, i.e. the stability of closed-loop system is not guaranteed and the learning process does not lead to well-defined dynamics [2]. In order to overcome this problem, the use of classical controllers have been suggested to complement adaptive fuzzy controllers. The fusion of fuzzy systems with classical control approaches makes it possible to benefit from the general function approximation property of fuzzy systems, as well as its power to use expert knowledge and the well established stability proof of classical control systems. For example in [3], [4] and [5], sliding mode fuzzy controllers are proposed, and in [6], [7] and [8], a H_∞-based fuzzy controller, a fuzzy-identification-based back-stepping controller and a model reference controller with an adaptive parameter estimator based on Takagi Sugeno (TS) fuzzy models are proposed, respectively.

Using a model system to generate the desired response is one of the most important adaptive control schemes [9] studied in such hybrid approaches, and to date, different fuzzy model reference approaches have been proposed. The indirect model reference fuzzy controllers described in [8], [10]-[13] and the direct model reference fuzzy controllers for TS fuzzy models described in [14], [15], and [16] can be cited as some examples. Most of the model reference fuzzy controller schemes existing in the literature assume that the full state measurement of the plant is available [8], [10]-[16]. However, in some practical applications state variables are not accessible for sensing devices or the sensor is expensive, and the state variables are just partially measurable. In such cases it is very essential to design an observer to estimate the states of the system. There has been a tremendous amount of activity on the design of nonlinear observers using fuzzy models, based on approaches like LMI [17]-[19], SPR Lyapunov function [20], [21] and adaptive methods [22].

The TS fuzzy system is one of the most popular fuzzy systems in model-based fuzzy control. A dynamical TS fuzzy system describes a highly nonlinear dynamical system in terms of locally linear TS fuzzy systems. The overall fuzzy system is achieved as a fuzzy blending of these locally linear systems [24]. Using this approach, it is possible to deal with locally linear dynamical systems rather than the original nonlinear dynamical system. When there are difficulties in the measurement of the states, the design of a fuzzy observer using the TS fuzzy model is considered in a number of different papers. In [22] an adaptive approach is proposed to design observer and controller for TS fuzzy system. However the proposed method considers only the regulation problem, tracking control is not addressed.

In this paper, a novel indirect model reference fuzzy controller is described that uses a novel optimal fuzzy observer. The optimality of the fuzzy observer is achieved by finding the optimal solution of an appropriate cost function. The optimal adaptation law used in the design and its stability analysis are quite

similar to the optimal adaptation law proposed in [23] and its stability analysis given therein. The superiority of the proposed controller over the one described in [23] is that the current work uses an observer so that full state measurement is not necessary. The stability analysis of the proposed observer and the controller is done using a Lyapunov function. Not only can the proposed indirect model reference fuzzy controller regulate the states of the system under control but also it can make the system track a desired trajectory. This is another benefit of the current work over the observer based model fuzzy controllers available in the literature, such as that described in [22]. To demonstrate the efficacy of the proposed method, it is then used to control a chaotic system. It is shown that by the use of the proposed approach, it is possible to make the chaotic system follow the reference model.

This paper is organized as follows. In Section 2 a brief study of zero-th order TS fuzzy systems is given. The structure of the proposed observer and its optimal adaptation law are introduced in Section 3, continuing by the stability analysis of the observer using a proper Lyapunov function. In Section 4 the proposed optimal observer is used in the design of the model reference fuzzy controller. Simulation results are presented in Section 5. Finally concluding marks are discussed in the next section.

2 Takagi-Sugeno Fuzzy Systems

In 1985 Takagi and Sugeno [25] proposed a new type of fuzzy systemin which the i-th rule of the fuzzy system is as follows:

$$R^i : If\ x_1\ is\ A_{i1}\ and\ x_2\ is\ A_{i2} ... and\ x_n\ is\ A_{in}\ Then\ y = F_i(x_1, x_2, ..., x_n) \tag{1}$$

In this fuzzy system $x_1, x_2, ..., x_n$ are the inputs of the fuzzy system and $F_i(.)$ is a function of inputs. This system can be seen as an extension of singleton fuzzy systems. It is to be noted that only the premise part of a TS fuzzy system is linguistically interpretable and that $F_i(.)$ are not fuzzy sets. The functions $F_i(.)$ can be chosen in different ways. If the functions are chosen as constants (θ_i), a singleton fuzzy system is recovered. This case is generally called a zero-th order TS fuzzy system, since a constant can be seen as a zero-th order Taylor expansion of a function. Another well-known possible selection of $F_i(.)$ is to select the rule consequent as a linear function of the inputs. The resulting fuzzy system is called first order TS fuzzy system.

The output of a TS fuzzy system can be calculated by

$$y = \sum_{i=1}^{m} h_i F_i \tag{2}$$

In which (h_i) is the normalized firing of the i-th rule and (m) is the number of the rules. In this paper we use zero-th order TS fuzzy system. The output of the zero-th order fuzzy system is calculated as:

$$y = \sum_{i=1}^{m} h_i \theta_i \tag{3}$$

3 The Structure of the Optimal Observer and its Stability Analysis

In this section, an optimal indirect adaptive fuzzy observer for a nonlinear system is proposed. The adaptation law for the estimation of the parameters of the nonlinear system is derived. Using an appropriate Lyapunov function, the stability of the proposed observer and the adaptation laws are analyzed.

3.1 The Structure of the Proposed Observer

Let the dynamical equation of the system be in the following form:

$$\begin{aligned} \dot{x} &= Ax + B[u + f(y)] \\ y &= Cx \end{aligned} \tag{4}$$

In which $x \in \mathbb{R}^n$ is the n-dimentional state vector, $A \in \mathbb{R}^{n \times n}$ is the known state matrix, $B \in \mathbb{R}^{n \times 1}$ is the known input matrix, $C \in \mathbb{R}^{1 \times n}$ is the known output matrix, $u \in \mathbb{R}$ is the input signal and $f(y)$ is an unknown Lipschitz function of y. The structure of the proposed observer is:

$$\dot{\hat{x}} = A\hat{x} + B[u + \sum_{i=1}^{m} \theta_i h_i(y)] + LCe \tag{5}$$

in which $\hat{x} \in \mathbb{R}^n$ is the estimated value of x, $L \in \mathbb{R}^{n \times 1}$ is the observation gain and $\sum_{i=1}^{m} \theta_i h_i(y)$ is the output of the fuzzy system with m being the number of rules used to estimate $f(y)$. The normalized firing strength of the i-th rule of this fuzzy system is $h_i(y)$, the parameters of the consequent part are θ_i and

$$e = \hat{x} - x \tag{6}$$

is the observation error.

3.2 The Dynamics of the Observation Error

We have:

$$\dot{e} = (A + LC)e + B[\sum_{i=1}^{m} \theta_i h_i(y) - f(y)] \tag{7}$$

It is supposed that the there exist optimal parameters θ_i^* for the fuzzy system such that:

$$f(y) = \sum_{i=1}^{m} \theta_i^* h_i(y) + \varepsilon(y) \tag{8}$$

in which $\varepsilon(y)$ is the approximation error. It can be proved [23] that if $f(y)$ is a Lipschitz function and the control signal u is bounded the time derivative of $f(y)$ is also bounded and we have:

$$\sup_t |\frac{df(y)}{dt}| < \sigma_f \tag{9}$$

In which σ_f is a positive constant. It is assumed that the time derivative of the approximation error ε is bounded so that:

$$\sup_t |\dot{\varepsilon}(y)| = \sup_t \left| \frac{d\left(\sum_{i=1}^{m} \theta_i^* h_i(y)\right)}{dt} - \frac{df(y)}{dt} \right| < \sigma_\varepsilon \tag{10}$$

where σ_ε is apositive constant. It follows that:

$$\sup_t \left| \frac{dh_i(y)}{dt} \right| \le \eta \tag{11}$$

in which η is a positive constant. The observation error dynamics can be expressed as follows.

$$\dot{e} = (A + LC)e + B\left[\sum_{i=1}^{m} \tilde{\theta}_i h_i(y) - \varepsilon(y)\right] \tag{12}$$

In above $\tilde{\theta}_i = \theta_i - \theta_i^*$. Since the fuzzy systems are proved to be general function approximators, by considering enough number of the rules for the fuzzy system, we have:

$$\sigma_\varepsilon = \sup_t |\varepsilon| \tag{13}$$

It is also assumed that there exists positive definite matrices P and Q such that:

$$\begin{aligned} &(A + LC)^T P + P(A + LC) = -Q \\ &PB = C^T \end{aligned} \tag{14}$$

3.3 The Optimal Adaptation Law for the Observer

In order to design an optimal adaptation law for the observer, the following cost function is defined for the observer.

$$J = \frac{1}{2}\int_0^{t_f} (e-\Delta)^T Q(e-\Delta)dt \tag{15}$$

in which $\Delta = e(t_f)$ and t_f is the final time. In addition, $Q \in \mathbb{R}^{n\times n}$ is a user defined positive definite matrix. This cost function is quite similar to the cost function used in [23]. The difference is that the cost function defined in [23] includes the tracking error while the cost function introduced here includes the observation error. This is an optimal observer design problem and its solution can be obtained by Pontryangin's maximum principle. To solve this optimal problem a Hamiltonian is defined as:

$$H\left(e, \sum_{i=1}^{m} \tilde{\theta}_i h_i\right) = \frac{1}{2}(e-\Delta)^T Q(e-\Delta) + p^T\left((A+LC)e + B\sum_{i=1}^{m}\tilde{\theta}_i h_i + B\sigma_\varepsilon\right) \tag{16}$$

in which p is an adjoint variable, the adjoint equation is given by:

$$\dot{p} = -\nabla H_e^T = -Q(e-\Delta) - (A+LC)^T p \tag{17}$$

The adaptation law for θ_i can be obtained by gradient method as:

$$\dot{\tilde{\theta}}_i = -\gamma_i h_i \nabla H_{\tilde{\theta}_i h_i} = -\gamma_i h_i p^T B \tag{18}$$

in which $\gamma_i > 0$ is the learning rate. The transversality condition requires that [26]:

$$p(t_f) = 0 \tag{19}$$

By letting $p = Pe + S\sum_{i=1}^{m}\theta_i h_i(y)$ and considering (17) we have:

$$\begin{aligned} &\dot{P}e + P\left(\left(A+LC\right)e + B\sum_{i=1}^{m}\tilde{\theta}_i h_i + B\sigma_\varepsilon\right) + \dot{S}\sum_{i=1}^{m}\theta_i h_i \\ &+S\frac{d\left(\sum_{i=1}^{m}\theta_i h_i\right)}{dt} \geq -Q(e-\Delta) - (A+LC)^T\left(Pe + S\sum_{i=1}^{m}\theta_i h_i\right) \end{aligned} \tag{20}$$

This is called sweeping method [26], [27]. In addition, assuming that the adaptation law is stable (the stability analysis will be considered in the next section) we have:

$$\sup_t \left| \frac{d\left(\sum_{i=1}^m \theta_i h_i\right)}{dt} \right| = \sup_t \left| \frac{d\left(\sum_{i=1}^m \tilde{\theta}_i h_i\right)}{dt} + \dot{\varepsilon}(y) + \frac{df(y)}{dt} \right|$$

$$\leq \sup_t \left| \sum_{i=1}^m \dot{\tilde{\theta}}_i h_i + \sum_{i=1}^m \tilde{\theta}_i \dot{h}_i \right| + \sigma_\varepsilon + \sigma_f \tag{21}$$

$$\leq \sup \left| -B^T p \sum_{i=1}^m \gamma_i h_i^2 \right| + \sup_t \left| \sum_{i=1}^m \tilde{\theta}_i \dot{h}_i \right| + \sigma_\varepsilon + \sigma_f$$

The first term of the right hand of (21) is bounded bacause p must be a stable solution to the optimal control problem and h_i is bounded because it is the firing strength of the fuzzy system. The second term must be bounded if the adaptation law of $\tilde{\theta}$ is stable and $\dot{h}_i$ is also bounded. Therefore we have:

$$\sup_t \left| \frac{d\left(\sum_{i=1}^m \theta_i h_i\right)}{dt} \right| < \sigma_t \tag{22}$$

From (20) it follows that:

$$\Delta = Q^{-1} \left[PB\left(\varepsilon - \sum_{i=1}^m \theta_i^* h_i\right) + S \frac{d\sum_{i=1}^m \theta_i h_i}{dt} \right] \tag{23}$$

and also:

$$\begin{aligned} &\dot{P} + P(A + LC) + (A + LC)^T P + Q = 0 \\ &\dot{S} + PB + (A + LC)^T S = 0 \end{aligned} \tag{24}$$

subject to:

$$P(t_f) = 0 \text{ and } S(t_f) = 0 \tag{25}$$

Considering the infinite horizon optimal control $t_f \to \infty$, P and S are in their steady state value ($\dot{P} = 0$, $\dot{S} = 0$) we have:

$$\begin{aligned} &P(A + LC) + (A + LC)^T P = -Q \\ &S = -(A + LC)^{-T} PB \end{aligned} \tag{26}$$

Furthermore:

$$p = Pe - (A + LC)^{-T} PB \sum_{i=1}^m \theta_i h_i \tag{27}$$

By substituting (27) in (18) it follows that:

$$\dot{\tilde{\theta}}_i = -\gamma_i h_i \left(e^T P - \vartheta \sum_{i=1}^{m} \theta_i h_i B^T P (A+LC)^{-1} \right) B \tag{28}$$

Considering (14) we obtain:

$$\dot{\tilde{\theta}}_i = -\gamma_i h_i e_y + \gamma_i h_i \vartheta \sum_{i=1}^{m} \theta_i h_i B^T P (A+LC)^{-1} B \tag{29}$$

In which $e_y = C(\hat{x} - x)$ and $\vartheta > 0$ is a design parameter.

3.4 Stability Analysis of the Proposed Observer

In order to analyze the stability of the proposed observer the following Lyapunov function is introduced.

$$V = e^T P e + \sum_{i=1}^{m} \frac{1}{\gamma_i} \theta_i^2 \tag{30}$$

In which P is the solution of (14). The time derivative of the Lyapunov function is obtained as:

$$\dot{V} = \dot{e}^T P e + e^T P \dot{e} + \sum_{i=1}^{m} \frac{2}{\gamma_i} \theta_i \dot{\theta}_i \tag{31}$$

Considering (12) and (29) we have:

$$\begin{aligned} \dot{V} \le\ & e(A+LC)^T P e + e^T P (A+LC) e + 2 e^T P B \sum_{i=1}^{m} \tilde{\theta}_i h_i (y) \\ & + 2 \left| e_y \right| \sigma_\varepsilon + 2 \sum_{i=1}^{m} \theta_i h_i \left(e^T P - \vartheta \sum_{i=1}^{m} \theta_i h_i B^T P (A+LC)^{-1} \right) B \end{aligned} \tag{32}$$

$P(A+LC)^{-1}$ can be decomposed into a symmetric part (M) and anti-symmetric part (N) such that [23]:

$$M = \frac{1}{2} \left(P(A+LC)^{-1} + (A+LC)^{-T} P \right) = -\frac{1}{2} (A+LC)^{-T} Q (A+LC)^{-1} < 0 \tag{33}$$

$$N = \frac{1}{2} \left(P(A+LC)^{-1} - (A+LC)^{-T} P \right) \tag{34}$$

Using the property of anti-symmetric matrix (N) we have:

$$B^T N B = 0 \tag{35}$$

So that:

$$\begin{aligned}&\dot{V} \le -e^T Qe + 2e^T PB\left[\sum_{i=1}^{m} \tilde{\theta}_i h_i(y)\right] + 2\,|\,e_y\,|\,\sigma_\varepsilon \\ &+2\sum_{i=1}^{m} \theta_i h_i e^T PB - \vartheta \sum_{i=1}^{m} \theta_i h_i B^T (A+LC)^{-T} Q(A+LC)^{-1} B\end{aligned} \tag{36}$$

furthermore:

$$\begin{aligned}&\dot{V} \le -e^T Qe + 2e^T PB\sum_{i=1}^{m} \theta_i^* h_i(y) + 2\,|\,e_y\,|\,\sigma_\varepsilon \\ &-\vartheta \sum_{i=1}^{m} \theta_i h_i B^T (A+LC)^{-T} Q(A+LC)^{-1} B\end{aligned} \tag{37}$$

and

$$\begin{aligned}&\dot{V} \le -\lambda_{\min}(Q)\|e\|^2 + 2\lambda_{\max}(P)\|B\|\|e\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \|\sigma_\varepsilon\|\right] \\ &-\vartheta\lambda_{\min}(Q)\left\|(A+LC)^{-1} B\sum_{i=1}^{m} \theta_i h_i\right\|^2\end{aligned} \tag{38}$$

In which $\lambda_{\min}(Q)$ is the smallest eigenvalue of (Q). It follows that in order to have $\dot{V} \le 0$ we should have:

$$-\lambda_{\min}(Q)\|e\|^2 + 2\lambda_{\max}(P)\|B\|\|e\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \|\sigma_\varepsilon\|\right] \le 0 \tag{39}$$

This equally means that:

$$\frac{2\lambda_{\max}(P)\|B\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \|\sigma_\varepsilon\|\right]}{\lambda_{\min}(Q)} \le \|e\| \tag{40}$$

Thus V decreases inside a compact set S where:

$$S = \left\{ e \in \mathbb{R}^n \,|\, \frac{2\lambda_{\max}(P)\|B\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \|\sigma_\varepsilon\|\right]}{\lambda_{\min}(Q)} \le \|e\| \right\} \tag{41}$$

The following theorem summarizes the forthgoing optimal adaptation law and its stability analysis.

Theorem 1. If there exists a positive definite matrix P satisfying (14), the observer given by (5) for the nonlinear dynamical system of (4) with the adaptation law of (29) which is the optimal solution of the cost function (15) ensures that the state estimation error and the estimated values of the fuzzy system θ_i are uniformly bounded. Furthermore, the estimation error can be made to approach an arbitrarily small value by choosing appropriate values for the design constants $\lambda_{\max}(P), \lambda_{\min}(Q)$ and sufficient number of rules for the fuzzy system.

4 The Design of Indirect Model Reference Fuzzy Controller Based on Proposed Observer and its Stability Analysis

In the previous section, the stability of the observer is considered using an appropriate Lyapunov function. In this section the stability of the control system is analyzed and a stable control signal is derived. The goal of the model reference fuzzy controller is to derive the system such that it follows the model reference system in the form of:

$$\dot{x}_m = A_m x_m + B_r r,\ A_m = A + LC + Ba_m^T \tag{42}$$

in which $x_m \in \mathbb{R}^n$ is the n-dimentional state vector of the reference system, $A_m \in \mathbb{R}^{n\times n}$ is the state matrix of the reference system, $B_r \in \mathbb{R}^{n\times 1}$ is the input matrix, $a_m \in \mathbb{R}^{n\times 1}$ is a user defined matrix which determines the dynamics of the reference model. The Lyapunov function is considered as:

$$V_1 = \hat{e}_m^T P_1 \hat{e}_m + e^T P e + \sum_{i=1}^{m} \frac{1}{\gamma_i}\theta_i^2 \tag{43}$$

In which P_1 is the solution of:

$$\begin{aligned} &(A + LC + Ba_m^T)^T P_1 + P_1(A + LC + Ba_m^T) = -Q_1 \\ &P_1 B = C^T \end{aligned} \tag{44}$$

and Q_1 is a positive definite matrix. In addition, $\hat{e}_m$ is the observed tracking error defined as $\hat{e}_m = \hat{x} - x_m$. The Lyapunov function now includes both the observation error and the observed tracking error and it is possible to use it to analyze stability of the tracking error too. Considering (30) we have:

$$V_1 = \hat{e}_m^T P_1 \hat{e}_m + V \tag{45}$$

The time derivative of the Lyapunov function is obtained as:

$$\dot{V}_1 = \dot{\hat{e}}_m^T P_1 \hat{e}_m + \hat{e}_m^T P_1 \dot{\hat{e}}_m + \dot{V} \tag{46}$$

since

$$\dot{\hat{x}} = A\hat{x} + B[u + \sum_{i=1}^{m} \theta_i h_i(y)] + LCe \tag{47}$$

By subtracting (42) from (47) we have:

$$\dot{\hat{e}}_m = A\hat{e}_m + B[u + \sum_{i=1}^{m} \theta_i h_i(y) - a_m^T \hat{x} + a_m^T \hat{x} - a_m^T x_m - b_r r] + LC\hat{e}_m - LCx \tag{48}$$

In which $\hat{e}_m = \hat{x} - x_m$ and:

$$\dot{\hat{e}}_m = A_m \hat{e}_m + B[u + \sum_{i=1}^{m} \theta_i h_i(y) - a_m^T \hat{x} - b_r r] - LCx \tag{49}$$

furthermore:

$$\begin{aligned} \dot{V}_1 &= \hat{e}_m^T \left(P_1 A_m + A_m^T P_1 \right) \hat{e}_m \\ &+2\hat{e}_m^T P_1 B[u + \sum_{i=1}^{m} \theta_i h_i(y) - a_m^T \hat{x} - b_r r] - 2\hat{e}_m^T P_1 LCx + \dot{V} \end{aligned} \tag{50}$$

Since for any $\alpha \geq 0$ we have:

$$2\hat{e}_m^T P_1 BLCx \leq \alpha (\hat{e}_m^T P_1 BL)^2 + \frac{1}{\alpha} y^2 \tag{51}$$

We obtain the following.

$$\dot{V}_1 \leq -\hat{e}_m^T Q_1 \hat{e}_m + 2\hat{e}_m^T P_1 B[u + \sum_{i=1}^{m} \theta_i h_i(y) - a_m^T \hat{x} - b_r r] + \alpha (\hat{e}_m^T P_1 L)^2 + \frac{1}{\alpha} y^2 + \dot{V} \tag{52}$$

Considering the indirect model reference fuzzy control signal as:

$$u = a_m^T \hat{x} - \sum_{i=1}^{m} \theta_i h_i(y) - \rho \frac{\hat{e}_{my}}{\hat{e}_{my}^2 + \delta} y^2 + b_r r \tag{53}$$

in which $\rho > 0$ is a design parameter and $B_r = b_r B$. One gets:

$$\dot{V}_1 \leq -\hat{e}_m^T Q_1 \hat{e}_m - 2\rho \hat{e}_m^T P_1 B \frac{\hat{e}_{my}}{\hat{e}_{my}^2 + \delta} y^2 + \alpha (\hat{e}_m^T P_1 L)^2 + \frac{1}{\alpha} y^2 + \dot{V} \tag{54}$$

And further:

$$\dot{V}_1 \leq -\hat{e}_m^T Q_1 \hat{e}_m + \alpha (\hat{e}_m^T P_1 L)^2 + \frac{1 - 2\rho\alpha}{\alpha \left(\hat{e}_{my}^2 + \delta \right)} \hat{e}_{my}^2 y^2 + \frac{\delta}{\alpha \left(\hat{e}_{my}^2 + \delta \right)} y^2 + \dot{V} \tag{55}$$

Taking:

$$0.5\rho^{-1} \leq \alpha \tag{56}$$

and:

$$2\alpha L^T P_1^T P_1 L \leq \lambda_{\min}(Q_1) \tag{57}$$

One obtains:

$$\dot{V}_1 \leq -\frac{1}{2} \lambda_{\min}(Q_1) \left\| \hat{e}_m \right\|^2 + \frac{\delta}{\alpha \left(\hat{e}_{my}^2 + \delta \right)} y^2 + \dot{V} \tag{58}$$

and:

$$\dot{V}_1 \le -\frac{1}{2}\lambda_{\min}(Q_1)\left\|\hat{e}_m\right\|^2 + \frac{2\delta}{\alpha\left(\hat{e}_{my}^2+\delta\right)}y_m^2 + \frac{2\delta}{\alpha\left(\hat{e}_{my}^2+\delta\right)}e_y^2 + \frac{2\delta}{\alpha\left(\hat{e}_{my}^2+\delta\right)}\hat{e}_{my}^2 + \dot{V} \tag{59}$$

Considering (38) one obtains:

$$\begin{aligned}\dot{V}_1 &\le -\frac{1}{2}\lambda_{\min}(Q_1)\left\|\hat{e}_m\right\|^2 + \frac{2\delta}{\alpha\left(\hat{e}_{my}^2+\delta\right)}y_m^2 + \frac{2\delta}{\alpha\left(\hat{e}_{my}^2+\delta\right)}e_y^2 + \frac{2\delta}{\alpha\left(\hat{e}_{my}^2+\delta\right)}\hat{e}_{my}^2 \\ &-\lambda_{\min}(Q)\|e\|^2 + 2\lambda_{\max}(P)\|B\|\|e\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \|\sigma_\varepsilon\|\right] \\ &-\vartheta\lambda_{\min}(Q)\left\|(A+LC)^{-1}B\sum_{i=1}^{m}\theta_i h_i\right\|^2\end{aligned} \tag{60}$$

Since:

$$\frac{2\delta}{\alpha\left(\hat{e}_{my}^2+\delta\right)} \le \frac{2}{\alpha} \tag{61}$$

We have:

$$\begin{aligned}\dot{V}_1 &\le -\frac{1}{2}\lambda_{\min}(Q_1)\left\|\hat{e}_m\right\|^2 + \frac{2}{\alpha}y_m^2 + \frac{2}{\alpha}e_y^2 + \frac{2}{\alpha}\hat{e}_{my}^2 \\ &-\lambda_{\min}(Q)\|e\|^2 + 2\lambda_{\max}(P)\|B\|\|e\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \|\sigma_\varepsilon\|\right] \\ &-\vartheta\lambda_{\min}(Q)\left\|(A+LC)^{-1}B\sum_{i=1}^{m}\theta_i h_i\right\|^2\end{aligned} \tag{62}$$

By taking α as:

$$\max\left\{\frac{8C^TC}{\lambda_{\min}(Q_1)}, \frac{4C^TC}{\lambda_{\min}(Q)}\right\} < \alpha \tag{63}$$

One obtains:

$$\begin{aligned}\dot{V}_1 &\le -\frac{1}{4}\lambda_{\min}(Q_1)\left\|\hat{e}_m\right\|^2 + \frac{2}{\alpha}y_m^2 \\ &-\frac{1}{2}\lambda_{\min}(Q)\|e\|^2 + 2\lambda_{\max}(P)\|B\|\|e\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \|\sigma_\varepsilon\|\right] \\ &-\vartheta\lambda_{\min}(Q)\left\|(A+LC)^{-1}B\sum_{i=1}^{m}\theta_i h_i\right\|^2\end{aligned} \tag{64}$$

If

$$\frac{2\lambda_{\max}(P)\|B\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \|\sigma_\varepsilon\|\right]}{\lambda_{\min}(Q)} \le \|e\| \tag{65}$$

and

$$y_m^2 \le \frac{\alpha}{8}\lambda_{\min}(Q_1)\left\|\hat{e}_m\right\|^2 \tag{66}$$

We have $\dot{V}_1 \le 0$ thus $V_1(t)$ decreases inside a compact set S_1 where:

$$S_1 = \left\{ \hat{e}_m, e \;\middle|\; y_m^2 \le \frac{\alpha}{8}\lambda_{\min}(Q_1)\left\|\hat{e}_m\right\|^2 \text{ and } \frac{2\lambda_{\max}(P)\|B\|\left[\sum_{i=1}^{m}\left\|\theta_i^* h_i(y)\right\| + \left\|\sigma_\varepsilon\right\|\right]}{\lambda_{\min}(Q)} \le \|e\| \right\} \tag{67}$$

It follows that it is possible to make the observed tracking error ($\hat{e}_m$) and the state estimation error (e) to approach an arbitrarily small value by choosing appropriate values for the design constants α, $\lambda_{\max}(P)$, $\lambda_{\min}(Q)$, $\lambda_{\max}(P_1)$ $\lambda_{\min}(Q_1)$ and sufficient number of rules for the fuzzy system. But the main concern is to make the tracking error ($e_m = x - x_m$) to approach any small value. Since $e_m = \hat{e}_m - e$ and considering the fact that $\hat{e}_m$ and e can be made arbitrarily small it follows that it is possible to make e_m as small as desired. The following theorem summarizes the forthgoing stability analysis.

Theorem 2. If there exists positive definite matrixes P and P_1 satisfying (14) and (45), the control signal given by (53) together with the observer given by (5) for the nonlinear dynamical system of (4) with the adaptation law of (29) which is the optimal solution of the cost function (15) ensures that the nonlinear dynamical system of (4) follows the reference model of (42) with bounded error. In addition, the state estimation error and the estimated values of the fuzzy system θ_i are uniformly bounded. Furthermore, the tracking error and the state estimation error can be made to approach an arbitrarily small value by choosing appropriate values for the design constants α, $\lambda_{\max}(P)$, $\lambda_{\min}(Q)$, $\lambda_{\max}(P_1)$ $\lambda_{\min}(Q_1)$ and sufficient number of rules for the fuzzy system.

5 Simulation Results

In this section we use the well-known chaotic system of Chua's circuit to depict the design procedure and verify the effectiveness of the proposed algorithm. The control of the nonlinear chaotic Chua's circuits is an important topic for numerous practical applications since this circuit exhibits a wide variety of nonlinear dynamic phenomena such as bifurcations and chaos [28]. This chaotic circuit possesses the properties of simplicity and universality, and has become a standard prototype for investigation of chaos. In this section we will use the proposed method to control the nonlinear chaotic Chua's circuit.

5.1 Dynamical Equations of Chua's Circuit

The modified Chua's circuit is described by the following dynamical system [28]:

$$\begin{aligned} \dot{x}_1 &= p(x_2 - \frac{1}{7}(2x_1^3 - x_1)) + u_1, \\ \dot{x}_2 &= x_1 - x_2 + x_3 + u_2, \\ \dot{x}_3 &= q.x_2 + u_3, \end{aligned} \tag{68}$$

in which u_1, u_2 and u_3 are the external inputs and x_1, x_2 and x_3 are the states of the system. Considering $q = -\frac{100}{7}$ (as used in many references as [28]) and $u_2 = u_3 = 0$ [28], one obtains the state space equations of the system as:

$$\begin{aligned} \dot{x}_1 &= p(x_2 - \frac{1}{7}(2x_1^3 - x_1)) + u_1, \\ \dot{x}_2 &= x_1 - x_2 + x_3, \\ \dot{x}_3 &= -\frac{100}{7}.x_2, \end{aligned} \tag{69}$$

in which $p = 10$ [28].

5.2 Control of Chua's Circuit

The dynamical equation of Chua's circuit (69) can be viewed as the nonlinear dynamical system in the form of (4) in which

$$A = \begin{bmatrix} 0 & p & 0 \\ 1 & -1 & 1 \\ 0 & -\frac{100}{7} & 0 \end{bmatrix} B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \tag{70}$$

and

$$f(x) = -\frac{1}{7} p(2x_1^3 - x_1), g(x) = 1 \tag{71}$$

In order to model $f(x)$ in the interval of [-2, 2], TS membership functions labeled as *about(-2)*, *about(-1)*, *about(0)*, *about(1)* and *about(2)* are considered. These labels correspond to fuzzy membership functions as: $exp(-(x_1 + 2)^2 / 0.42^2)$, $exp(-(x_1 + 1)^2 / 0.42^2)$, $exp(-x_1^2 / 0.42^2)$, $exp(-(x_1 - 1)^2 / 0.42^2)$ and $exp(-(x_1 - 2)^2 / 0.42^2)$, respectively. The following rules for the TS fuzzy model are considered.

Rule 1: If x_1 is $about(-2)$ then $\dot{x} = Ax + B(a_1^T x + b_1 u)$

Rule 2: If x_1 is $about(-1)$ then $\dot{x} = Ax + B(a_2^T x + b_2 u)$

Rule 3: If x_1 is $about(0)$ then $\dot{x} = Ax + B(a_3^T x + b_3 u)$

Rule 4: If x_1 is $about(1)$ then $\dot{x} = Ax + B(a_4^T x + b_4 u)$

Rule 5: If x_1 is $about(2)$ then $\dot{x} = Ax + B(a_5^T x + b_5 u)$

in which A, B and C are defined as in (70) and:

$$a_1 = a_5 = \begin{bmatrix} -\frac{23}{7} p & 0 & 0 \end{bmatrix}^T, a_2 = a_4 = \begin{bmatrix} -\frac{5}{7} p & 0 & 0 \end{bmatrix}^T,$$
$$a_3 = \begin{bmatrix} \frac{1}{7} p & 0 & 0 \end{bmatrix}^T, b_1 = b_2 = b_3 = [1,0,0]^T \tag{72}$$

are estimated using linearization around the mean point. It should be noted that these values are reported for demonstration and are not used in the design of the controller. The gain of reference model a_m and gain of the observer L are taken as:

$$a_m = [0.01,0,0], L = \begin{bmatrix} -12.5 & -0.14 & -0.12 \end{bmatrix}^T, \tag{73}$$

respectively. The gain of the reference model a_m correspond to the reference model (A_m) as:

$$A_m = \begin{bmatrix} -12.5 & 10 & 0 \\ -0.14 & -1 & 1 \\ -0.12 & -14.29 & 0 \end{bmatrix} \tag{74}$$

whose eigenvalues are: $\{-12.39, -0.55 \pm 3.77i\}$. The design parameter Q is taken as

$$Q = \begin{bmatrix} 24.98 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix} \tag{75}$$

in which $I_{3\times 3}$ is the identity matrix. The positive definite matrix P which is the solution of (14) is obtained as:

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 76.43 & -5 \\ 0 & -5 & 5.7 \end{bmatrix} \tag{76}$$

Taking Q_1 as:

$$Q_1 = \begin{bmatrix} 25 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix} \tag{77}$$

the positive definite matrix P_1 which is the solution of (44) is obtained as:

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 76.43 & -5 \\ 0 & -5 & 5.7 \end{bmatrix} \tag{78}$$

The state matrix of the reference model is set as $A_m = A + Ba_m^T + LC$. Using these design parameters, the regulation performances of the proposed control scheme are tested. Figures 1(a)-1(d) show the results of the regulation of the system as well as the state estimation performance of the observer.

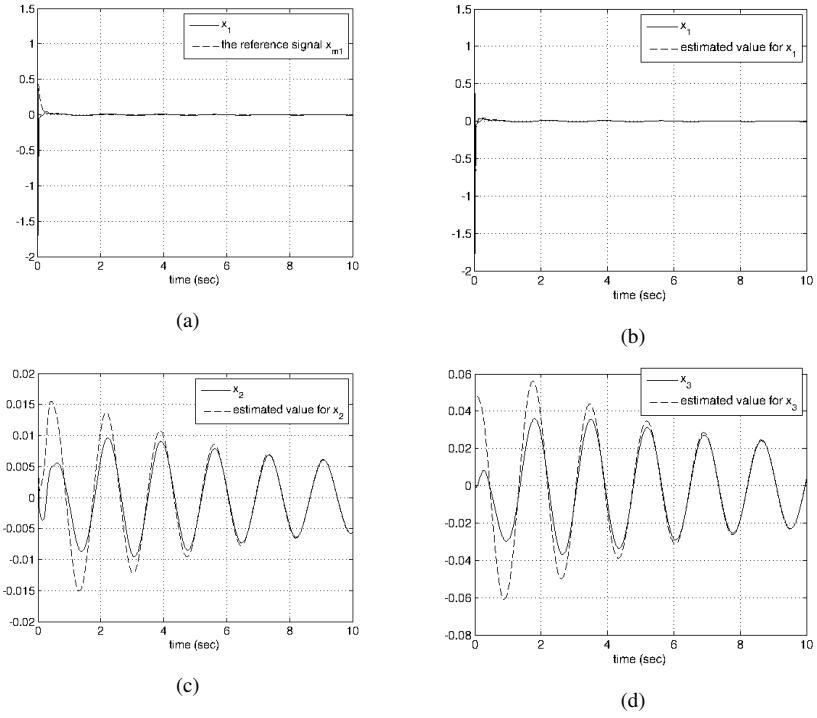


Figure 1

The regulation performance of the proposed observer and the controller when applied to Chua's chaotic system, a) The regulation response of Chua's chaotic system for x_1, b) The performance of the observer for x_1, c) The performance of the observer for x_2, d) The performance of the observer for x_3

The initial values for the states of the system, the observer and the reference model are considered as: $x = [0.5,0,0]^T$, $\hat{x} = [0.4,0,0.05]^T$ and $x_m = [0.55,0,0.1]$ respectively. In addition the tracking performance of the proposed controller is tested. Figures 2(a)-2(d) depict the tracking performance of the controller and the response of the observer. As can be seen from the figures, the tracking performance of the system under control is quite satisfactory.

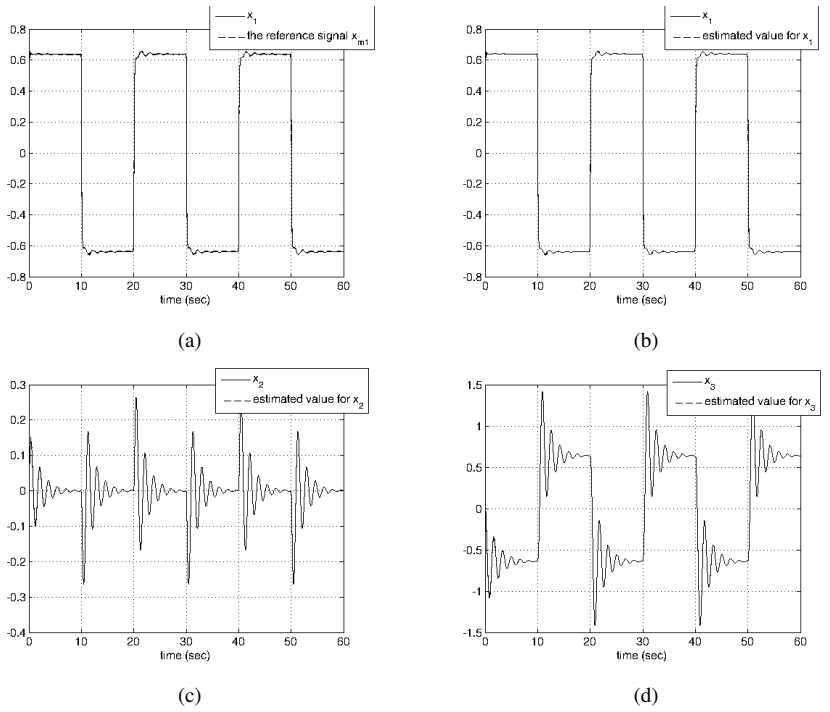


Figure 2

The tracking performance of the proposed observer and the controller when applied to Chua's chaotic system, a) The tracking response of Chua's chaotic system for x_1, b) The performance of the observer for x_1, c) The performance of the observer for x_2, d) The performance of the observer for x_3

Conclusions

This paper describes the design of an indirect model reference adaptive fuzzy controller based on an optimal observer for use with nonlinear dynamical systems. The proposed method adopts a TS fuzzy model to represent the dynamics of the system in hand and the adaptive model reference controller. The main contribution of the current work with respect to the previous studies in the field of model reference fuzzy controllers is that the current approach benefits from an optimal observer and therefore full state measurement is no longer needed. Its additional superiority over the observer-based TS adaptive fuzzy controllers seen in the

literature is that it is capable of making the system follow a reference model rather than just regulation of the system to zero. Moreover the proposed algorithm calculates the parameters of the TS fuzzy model from data using an optimal adaptation law. The stability of the approach is automatically accomplished with the derivation of the adaptive law by the Lyapunov theory. Lastly, through the application to a Chua's circuit, the applicability of the design to the practical problems of control of chaotic systems is verified. It is shown that the current approach is capable of controlling Chua's circuit with high performance.

References

[1] T. Procyk, E. Mamdani, "A Linguistic Self-Organizing Process Controller," Automatica, Vol. 15, No. 1, pp. 15-30, 1979

[2] K. M. Passino, S. Yurkovich, Fuzzy Control, Addison-Wesley, USA, 1998

[3] Chih-Chiang Cheng, Shih-Hsiang Chien, Adaptive Sliding Mode Controller Design Based on T–S Fuzzy System Models, Automatica, Volume 42, Issue 6, pp. 1005-1010, 2006

[4] Martin Kratmüller, The Adaptive Control of Nonlinear Systems Using the T-S-K Fuzzy Logic, Acta Polytechnica Hungarica Vol. 6, No. 2, pp. 5-16, 2009

[5] Tai-Zu Wu, Yau-Tarng Juang, Adaptive Fuzzy Sliding-Mode Controller of Uncertain Nonlinear Systems, ISA Transactions, Volume 47, Issue 3, pp. 279-285, 2008

[6] Martin Kratmüller, Adaptive Fuzzy Control Design, Acta Polytechnica Hungarica Vol. 6, No. 4, pp. 29-49, 2009

[7] Chun-Fei Hsu, Chih-Min Lin, Fuzzy-Identification-based Adaptive Controller Design via Backstepping Approach", Fuzzy Sets and Systems, Vol. 151, No. 1, pp. 43-57, 2005

[8] C.-W. Park, M. Park, Adaptive Parameter Estimator Based on t-s Fuzzy Models and its Applications to Indirect Adaptive Fuzzy Control Design, Information Sciences, Vol. 159, pp. 125-139, 2004

[9] K. J. Astrom, B. Wittenmark, Adaptive Control, Dover Publications, 2008

[10] Young-Wan Cho, Chang-Woo Park, Mignon Park, "An Indirect Model Reference Adaptive Fuzzy Control for SISO Takagi-Sugeno Model", Fuzzy Sets and Systems, Vol. 131, No. 2, pp. 197-215, 2002

[11] Cho, Y.-W., Park, C.-W., Kim, J.-H. and Park, M., Indirect Model Reference Adaptive Fuzzy Control of Dynamic Fuzzy-State Space Model, Control Theory and Applications, IEE Proceedings, Vol. 148, No. 4, pp. 273-282, 2001

[12] Koo, T. J., Stable Model Reference Adaptive Fuzzy Control of a Class of Nonlinear Systems, Fuzzy Systems, IEEE Transactions on, Vol. 9, No. 4, pp. 624-636, 2001

[13] Park, C.-W., Cho, Y.-W., Adaptive Tracking Control of Flexible Joint Manipulator Based on Fuzzy Model Reference Approach, Control Theory and Applications, IEE Proceedings, Vol. 150, No. 2, pp. 198-204, 2003

[14] Youngwan Cho, Yangsun Lee, Kwangyup Lee, and Euntai Kim, A Lyapunov Function-based Direct Model Reference Adaptive Fuzzy Control, Knowledge-based Intelligent Information and Engineering Systems, Vol. 3214, 202-210, 2004

[15] Young-Wan Cho, Eung-Sun Kim, Ki-Chul Lee and Mignon Park, Tracking Control of a Robot Manipulator Using a Direct Model Reference Adaptive Fuzzy Control, Intelligent Robots and Systems, 1999. IROS '99. Proceedings, Vol. 1, pp. 100-105, 1999

[16] Khanesar, M. A., Kaynak, O., Teshnehlab, M., "Direct Model Reference Takagi–Sugeno Fuzzy Control of SISO Nonlinear Systems," Fuzzy Systems, IEEE Transactions on, Vol. 19, No. 5, pp. 914-924, 2011

[17] Han Ho Choi, LMI-based Nonlinear Fuzzy Observer-Controller Design for Uncertain MIMO Nonlinear Systems, Fuzzy Systems, IEEE Transactions on, Vol. 15, No. 5, pp. 956-971, 2007

[18] Teixeira, M. C. M., Assuncao, E., Avellar, R. G., On Relaxed LMI-based Designs for Fuzzy Regulators and Fuzzy Observers, Fuzzy Systems, IEEE Transactions on, Vol. 11, No. 5, pp. 613-623, 2003

[19] Kuang-Yow Lian, Chian-Song Chiu, Tung-Sheng Chiang and Liu, P., LMI-based Fuzzy Chaotic Synchronization and Communications, Fuzzy Systems, IEEE Transactions on, Vol. 9, No. 9, pp. 539-553, 2001

[20] Yih-Guang Leu, Tsu-Tian Lee, Wei-Yen Wang, Observer-based Adaptive Fuzzy-Neural Control for Unknown Nonlinear Dynamical Systems, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, Vol. 29, No.5, pp. 583-591, 1999

[21] Yih-Guang Leu, Wei-Yen Wang, Tsu-Tian Lee, Observer-based Direct Adaptive Fuzzy-Neural Control for Nonaffine Nonlinear Systems, Neural Networks, IEEE Transactions on, Vol. 16, No. 4, pp. 853-861, 2005

[22] Chang-Ho Hyun, Chang-Woo Park, Seung woo Kim, Takagi-Sugeno Fuzzy Model-based Indirect Adaptive Fuzzy Observer and Controller Design, Information Sciences, Vol. 180, No. 11, pp. 2314-2327, 2010

[23] N. Nguyen, K. Krishnakumar, J. Boskovic, “An Optimal Control Modification to Model Reference Adaptive Control for Fast Adaptation,” in AIAA Guidance, Navigation and Control Conference, Honululu, HI, 2008

[24] Mohammad Ababneh, Ahmed Muhammed Almanasreh, Hani Amasha, Design of Digital Controllers for Uncertain Chaotic Systems Using Fuzzy Logic, Journal of the Franklin Institute, Vol. 346, No. 6, pp. 543-556, 2009

[25] Takagi, T., M. Sugeno, Fuzzy Identification of Systems and its Applications to Modeling and Control. IEEE Trans. System, Man and Cybernetics, Vol. 15, No. 1, pp. 116-132, 1985

[26] Frank L. Lewis, Vassilis L. Syrmos, Optimal Control, Wiley-IEEE, 1995

[27] Bryson, A. E., Jr and Y. C. Ho, Applied Optimal Control, New York:Hamisphere, 1975

[28] M. T. Yassen, Adaptive Control and Synchronization of a Modified Chua's Circuit System, Applied Mathematics and Computing, Vol. 135, pp. 113-128, 2003

ERP Project Implementation: Evidence from the Oil and Gas Sector

Alok Mishra, Deepti Mishra

Department of Computer Engineering, Atilim University
Incek 06836, Ankara Turkey
alok@atilim.edu.tr, deepti@atilim.edu.tr

Abstract. Enterprise Resource Planning (ERP) systems provide integration and optimization of various business processes, which can lead to improved planning and decision quality, and a smoother coordination between business units, resulting in higher efficiency and a quicker response time to customer demands and inquiries. This paper reports the challenges and opportunities and the outcome of an ERP implementation process in the Oil & Gas exploration sector. This study will facilitate the understanding of the transition, constraints, and implementation process of ERP in this sector and will also provide guidelines from lessons learned in this regard.

Keywords: case study; ERP; implementation; oil and gas exploration; SAP

1 Introduction

ERP implementation poses major challenges to organizations, as many of them fail in their early stages or substantially exceed the project cost [1]. ERP systems differ qualitatively from prior large scale Information Technology (IT) implementations in three ways [2]: 1) ERP impacts the whole organization, 2) employees may be learning new business processes in addition to new software, and 3) ERP is often a business led initiative, rather than IT led. ERP is an integrated set of subsystems that integrates all facets of the business, including planning, manufacturing, logistics, sales and marketing. ERP systems originated to serve the information needs of manufacturing companies. Over time though, they have grown to serve other industries, including financial services, consumer goods sector, supply chain management and the human resources sector. These systems provide integration and optimization of various business processes and this was what the companies looked for [3] along with tangible and intangible business benefits to organizations [4]. Effective integration is the key because if one of these links fail, the organization's performance may suffer and may not meet the expectations of its customers or the service level of its competitors [5]. It

is not wrong to say that ERP systems gained importance as they arrived at a time when process improvement and accuracy of information became critical strategic issues [6]. With this growth, ERP systems, which first ran on mainframes before migrating to client-server systems, are now migrating to the Web and include numerous applications. ERP is a product that helps automate a company's business process by employing an integrated user interface, an integrated data set, and an integrated code set. Hunter and Lippert [7] forecasted the ERP market to reach USD 1 trillion by 2010. A summer 2005 survey of members of the Society for Information Management showed that ERP is among the top application and technology developments of its members [8]. ERP systems are complex, and implementing one can be challenging, time-consuming and an expensive project for any company [9]. Motwani et al. [10] emphasized that ERP adoption involves initiating appropriate business process changes as well as information technology changes to significantly enhance performance, quality, costs, flexibility, and responsiveness. ERP systems are widely adopted in a diverse range of organizations and define the business model on which they operate [11]. An ERP implementation can take many years to complete and costs tens of millions of dollars for a moderate size firm and more than $100 million for large organizations [12]. Implementing an ERP system is a major undertaking. It is well known that the implementation of an ERP system is a very expensive and complex task and implementation tasks include consulting, process design, data conversion, training, integration and testing [13]. About 90% of ERP system implementations are late or over budget [14] and the success rate of ERP systems implementation is only about 33% [15] [16]. The relative invisibility of the ERP implementation process is also identified as a major cause of ERP implementation failures [17]. Such invisibility is attributed to the unpredictably complex social interaction of IT and organization [18]. Volkoff [19] suggested that the critical challenge of ERP implementation is believed to be the mutual adaptation between IT and user environment. The inclusion of today's strategic choices into the enterprise systems may significantly constrain future action. By the time the implementation of an ERP system is completed, the strategic context of the firm may have changed [11]. Nicolaou [20] reports that ERP implementation success often results from a number of factors, such as user participation and involvement in software development, the assessment of business needs, the processes during the analysis phase of the project and the level of data integration designed into the systems. ERP changes these processes, from designing a custom system to accommodate the existing business processes of a firm to selecting a business application system that best meets the firm's needs. Mabert et al. [3] suggested that case studies and interviews facilitate to obtain reliable and detailed information on the current status of ERP practice and ERP implementations. They further argued that most implementation projects are unique in many ways, in spite of many common underlying issues, activities and strategies. To meet time deadlines alongside budget targets, ERP projects must be planned very carefully and managed very efficiently [3]. Moreover, a lack of understanding and time and budget pressures

budget pressures make it difficult for system and maintenance personnel to identify and remove unused modifications during a release change [21].

In the context of ERP project implementation, challenges represent major pitfalls which, if not addressed, may cause the failure of a project. Therefore, it is important to understand the real-life implementations, problems and related scenarios in detail.

Furthermore, to the best of our knowledge, very few real-life ERP implementations in the oil and gas sector are documented in the literature. Therefore, this paper will facilitate the understanding of the constraints, problems, success and pitfalls of implementation in this sector.

This paper is organized as follows: First, the relevant ERP implementation literature is reviewed. The next section follows a real-life ERP implementation as a case study, followed by discussions. Section 5 summarizes the conclusions.

2 Literature Review

ERP systems, similar to other management information systems, are often perceived as very complex and difficult to implement [22], [23]. System implementation success depends on many factors: the ERP system evaluation, vendor selection, the ERP consultant, the implementation plan and execution are all critical to the success of implementing an ERP system [24]. The inability of some firms to successfully implement and utilize enterprise systems to increase organizational outcomes has been a source of concern for both practitioners and academia [25]. The evidence of enterprise implementation failures go back to the late 1990s [26], [27], [9]. For many organizations, ERP systems are the largest systems they have worked with in terms of the financial resources invested, the number of people involved and the scale of implementation [24]. Several cases of ERP system implementation have experienced considerable difficulties [28], [29], [30], [23]. The failure rate of ERP implementation is very high [31]. Numerous examples of failed and abandoned implementation projects are cited in the literature, such as Fox-Meyer Drug, Mobile Europe, Dell and Applied Materials [9]. Wah [30] cites failures at Whirlpool, Hershey, Waste Management, Inc. and W.L. Gore & Associates. The University of Massachusetts-Amhrest [32] and Indiana University [33], have also experienced revenue loss, wasted time, cost overruns and delays in ERP implementation projects. The Chaos Chronicles mentioned that only 34% of IT projects undertaken by Fortune 500 companies are successfully completed [34]. Nike's ERP implementation is included in a listing of "infamous failures in IT project management" because of a major inventory problem which resulted in a profit drop of USD 100 million in the 3rd quarter of 2000 [35].

Muscatello and Parente [36] cite ERP failure rates to be as high as 50%. Among other obstacles, technical problems and people obstacles have been cited as the major barriers [37], [29]. The types of problems and issues that arise from the implementation of ERP systems range from specific issues and problems that can come up during the installation of an ERP to behavioural, procedural, political and organisational changes, among others, that manifest themselves once the system is installed. In the case of ERP, successful implementation is imperative, since the costs and risks of these technology investments rival their potential pay-offs [38]. The failure of ERP system implementation projects may lead to bankruptcy [9], [39], [40], [41]. A study of 100 projects by Sirkin and Dikel [42] found that their sponsors considered them successful in only one-third of the cases, and that tangible financial impact was achieved in only 37% of cases. Markus et al. [43] suggest that ERP systems are inherently flexible, which means that stakeholders have many opportunities to influence the form of technology during the initial decision-making, development, implementation and also the use of the system. They further argued that many problems related to ERP-implementation are related to a mismatch of the system to characteristics of the organization. This is supported by Davenport [9], that "ERP tends to impose its own logic on a company's strategy, culture, and organization which may or may not fit with existing organizational arrangements". Although ERP systems are functionally rich, standardizing organizational processes with these systems is often difficult [44]. It is found that many firms that have experienced success with ERP have comprehensively reengineered their organizational processes and structures as a method for enterprise–wide transformation [45]. In the case of implementing an ERP system we should put more effort in customizing ERP modules to comply with the existing workflow, report formats and data needs [24]. Involving users as early as possible in system implementation is generally a good strategy [46]. As an enterprise system, the success of ERP implementation requires close cross-functional cooperation [10]. Further evidence from literature shows that, although many organizations are using some modules of an ERP system, they do not see themselves as equipped with ERP [47], [48], [49].

In particular, IT integrators that specialize in energy are seeing more opportunities in what is termed as the "upstream" segment of the oil and gas sector. Upstream includes oil and gas exploration and the drilling and operation of wells. Drilling companies deal with large assets and work crews that move about a country or different ocean sites. Such companies use ERP to make sure their resources are deployed effectively. ERP solutions also help companies track equipment maintenance and keep tabs on employee certification and training. Drilling personnel may need certification to operate certain types of equipment [50]. Mergers and acquisitions are common in the upstream space, and integrators find opportunity in consolidation. This trend got underway a few years ago and continues apace. Consolidation begets complexity and generates interest in ERP. Defining a reasonable (i.e., smaller) system scope by phasing in software functionality over a series of sequential implementation phases is an important

means of decreasing complexity [51]. Moore [50] further suggests that, as the oil and gas sector companies absorb others, operations may span several countries, each with its own statutory reporting requirements. Companies crossing international boundaries also need to deal with multiple currencies. Overall, combined organizations face rationalizing financial and accounting systems which require ERP implementation.

Case study research is "an empirical inquiry that investigates a contemporary phenomenon within its real-life context" [52]. Since research is more interested in the process aspects of ERP implementation, a case study has the potential of providing an in-depth investigation of these issues in a real-life context (Yin, 2003). Generally, the case study method is a preferred strategy when "how" and "why" questions are being posed and the researcher has little control over events [52]. The case study method, a qualitative and descriptive research method, looks intensely at an individual or small number of participants, drawing conclusions only about the participants or group and only in the specific context [52]. The case study method is an ideal methodology when a holistic, in-depth investigation is required [53]. The case study method has been proven a useful tool in investigating the problems of ERP implementation [54], [55], [56], [10].

3 Case Study

3.1 Background of the Company

The company was established in the early 1970s to handle the drilling operations required for exploration and field development as well as undertaking work-over and maintenance operations in both onshore and offshore areas. It has successfully carried out all the requirements of drilling operations and played an important role in the discovery of oil and gas. The main functions of the company are:

- Operations: This function includes two main divisions: Onshore and Offshore – each handles drilling operations. A logistics division is also included under this function and is responsible for providing logistics support in terms of transportation and civil equipment.
- Technical: This is mainly responsible for providing technical support to the operations function. The key divisions under this function are commercial (procurement, inventory, tendering, warehouse, etc), engineering & projects, maintenance, business support and a newly established division under the name of new services. The field support services, two warehouses and two workshops, are under commercial and maintenance divisions respectively.

- Administration: The role of this function is to provide administrative support including human resources (HR), finance, IT and general services. All of these divisions are located in the head office.

3.2 Information Technology Infrastructure Setup

The Information Systems & Technology (IS&T) department was formally established in the early 1990s with the mandate of providing computer and networking services to employees at the head office. At that time, the company was running Novell Netware and desktop computers were primarily used by finance and payroll services. The structure of the IS&T consisted of a networking unit and applications unit. The total number of IT staff, including network engineers, programmers and customer support staff was under 20. The following in-house data-based applications were being used:

- **Financial Applications:** General Ledger, Accounts Payable, Accounts Receivable, Payroll
- **Material Management:** Inventory Management, Fixed Assets
- **Miscellaneous:** Employee Database, Maintenance Work Order, Historical Database

Most of the above applications were developed by third parties and later on supported, maintained and enhanced by the internal development team of IS&T. Each application was dedicated to a particular group (department or process) and the data exchange among these applications was very limited. The standard management reports were incorporated in the applications and were printed and distributed to the management or the concerned staff on a periodic or on-request basis. Management had to rely on the availability of the existing data and most of the decision making required a lot of manual information from various resources.

Initially the computers were only available to financial analysts, data entry operators and managers. During the mid 90s, PC-based computing became popular and gradually all employees were provided with PC workstations with a Windows operating system using word processing tools and other office applications. After all of the PCs were networked, the company decided to centralize the electronic files and hence the storage system (merely a dedicated file server) was added to the data centre.

3.3 Weaknesses of IT Applications

The following problems were faced in the old IT setup [57]:

- Only a few functions / processes were automated using database applications.

- All the applications were working in silos without any exchange or integration among them.
- The maintenance of these applications was very difficult due to the lack of documentation of source code, process information among the development team, etc.
- Most business areas were not automated – hardly any decision-making was fully supported by the existing applications.
- Most of the company's processes were cross-functional, e.g. Material Requirement Planning, Procurement, Inventory, Maintenance, Invoices and Payments, Operations Planning, etc. However, the existing applications only supported a small portion of the cross-functional process, so the value generated by these applications was offset by the subsequent manual flow of the information.
- The architecture of the applications itself was weak. The system controls were inappropriate, allowing human error during data entry. As a result, the management had little confidence in the reports generated from the system, resulting in a forced parallel-run of the manual registers and files for reconciliation and validation purposes.
- The core business areas were handled by manual processes. For example, more than 80% of staff were working in operations (onshore and offshore), 10% were based in the Head office and the remaining 10% were deployed in field support services (workshops, warehouses, base camps, etc) – none of these areas had IT systems to support their processes.
- Long-employed staff with built-up tacit knowledge of the company became the only source of information. Lack of process documentation aggravated the problem and a few key positions held most of the process knowledge, creating critical organizational risk.

3.4 ERP Implementation

3.4.1 Objectives Setting

In order to define clear goals and a set of expectations, the taskforce arranged a workshop with the management team to obtain their viewpoint. Participants agreed on the following points:

Timeframe – the implementation should not take a long time to complete.

Cost – learning from industry experience, it was a general concern that any such implementation typically took 3 times the initially estimated cost; the taskforce was asked to focus on the cost variance of the project.

3.4.2 ERP Selection

The first task was to finalize the selection of a particular ERP system. The task force had the following options to evaluate:

i) Single ERP (System Application Product - SAP or Oracle) or

ii) Best of breed (selecting the best module for each of its functional area)

Option (ii) was discarded quickly as it required more cost, time and skills to implement. In addition, it required building a comprehensive skill set for a variety of applications, which was extremely difficult at the time. Therefore, the option to go for a single ERP was selected. The next question was to choose between SAP and Oracle as these two ERP packages were amongst the most popular choices in that region and industry sector (i.e. Oil & Gas). Again the taskforce had the following options to consider:

i) Conducting a self-study and choose between SAP and Oracle or

ii) Hiring a consultant to study the company's requirements and propose a particular ERP system.

After evaluating both options, the taskforce dismissed the second option as it required extra time (the tendering process itself could take many weeks) and cost. Therefore, it was decided to arrange meetings with other sister companies who had already implemented an ERP to obtain their view point and lessons learned. It was also decided to arrange volunteers from each functional area to study the high-level features of a particular module of both ERPs. After conducting the self-study and meetings with other operating companies, the task force agreed to proceed with SAP. The recommendation was presented to management and it was accepted.

The task force then conducted market research to find out the range of costs and timeframe. The initial data collected was not very encouraging as the minimum cost identified was USD 8 million (software license, hardware and implementation cost). The average implementation time ranged from 18 months to 3 years; which was beyond the initial estimations, as the company was aiming to complete the transition in 12 months.

3.4.3 Scoping and Approach Definition

The taskforce then moved to the Scoping and Planning phase in which a team of focal points (from each of the functional areas) was created to jointly develop a business requirements document for the ERP implementation. The focal points were selected based on their experience and knowledge of functional areas of the company. These focal points were required to allocate 80% of their business hours to work on this task as the deadline was in four weeks. Since most of the focal points were new to this type of work, they started working on their individual

areas in their own style – the consolidated set of requirements produced by the team were clearly lacking in quality and consistency, as the requirements were either too high-level/generic or too detailed. The team took another two weeks to refine those requirements further. An organization's strategic decision on ERP customization or business process adaptation during planning can have a profound impact on the practices used to support the system during maintenance and support [9]. Here it is also important to note that IT management may poorly define goals, have an overly simplistic project plan, use unrealistic deadlines and budgets, and fail to set and manage the expectations on the product (the software being developed) and the project (the development process) to gain support from users, developers, and functional managers [58].

It was planned to implement the following SAP modules in the first round of implementation:

Table 1
Implemented SAP Modules

Financial Accounting	**Controlling**	**Asset Management**	**Human Resources**	**Plant Maintenance**	**Material Management**
General Ledger	Cost Elements	Purchase	Employment History	Labour	Requisitions
Accounts Receivable	Cost Centres	Sale	Payroll	Material	Purchase Orders
Accounts Payable	Activity Based Costing (ABS)	Depreciation	Succession Planning	Downtime and Outages	Goods Receipt
Book Close	Profit Centres	Tracking	Career Management		Inventory Management
Consolidation	*Interface development with Oil and Gas applications	*Oil and Gas Control report system			Bill of Material

** Specific to Oil and Gas sector requirements*

The taskforce had to address some of the strategic options:

Big-Bang vs. Phased Approach: One of the questions was to finalize the implementation approach – whether to implement all modules in parallel or use a phased approach where each module would be implemented in a sequential manner. The later approach seemed to take a longer time than big-bang, and therefore the team proposed to adopt a big-bang approach.

Third Party vs. In-house Implementation: Where the first question mainly addressed the timeframe, this question concerned cost as well. The taskforce evaluated various options and the most suitable appeared to be the hiring of SAP consultants on a contract (as short-term employees), along with an experienced SAP project manager whose core responsibility would be to

manage the SAP contract staff to deliver in the agreed time frame. Most of the SAP consultants were recruited from a body-shop (Indian resource costing maximum 20% of any SAP implementation consultancy firm). For ERP implementations in particular, in-house expertise is often lacking, and companies often turn to external consultants in implementing the system [59] but the outsourcing of jobs does not transfer the ultimate management responsibility for their successful completion [60]. They further argued that poor management of outsourcing responsibilities can increase risks and create integration problems across products and processes.

During this phase, the new SAP project manager was recruited and a team of 10 SAP consultants was hired as contract employees. These included six functional resources specialized in different SAP modules, two SAP Advanced Business Application Programming (ABAP) developers as technical resources, one SAP GUI and security administrator and one database administrator. At that time, SAP 4.6C version was bought. The license agreement included all SAP modules along with 200 initial user licenses.

3.4.4 Business Blueprints

The newly recruited project manager formed a functional team including the focal points from each of the business areas and the SAP functional consultants. The team was given the task of preparing the detailed business blueprints which were mainly detailed definitions of the company's processes and their mapping with the best practice-based processes existing, defined in SAP. In most areas, the company agreed to adopt the built-in processes of SAP as it gave an opportunity to implement the best practices simultaneously. The HR and payroll modules, however, required some customization, as certain local personnel policies were governed by government regulations and changing them was out of the question.

The task took eight weeks. With some known and unknown weaknesses in the blueprint document, the team decided to move to the next phase.

3.4.5 Design and Development

In IT projects, design and implementation decisions made at an early stage can have an impact on activities undertaken at a later stage [60]. During the design phase, the complete definition of SAP GUI screens, transaction details, input/output layout and reporting formats were prepared. As most of the existing processes were manual, the major part of the design phase was actually aiming towards a vanilla implementation of SAP. The design phase started in the 13th week of the project (measured from the Scoping and Approach Definition Phase), and it took nearly eight weeks to complete. As time elapsed, the team was feeling a sense of urgency to complete the tasks-in-hand. As a result, some of the areas such as detailed reporting of requirements, test criteria, test cases and others did

not get the attention they required. Nonetheless, the team produced a detailed design document at the end of the design phase. The role of the focal points was merely to review and sign-off the design document.

During the design phase, the technical team had completed the hardware sizing and specification. The platform choices were left open for the company and, based on the long-term relations with the existing hardware vendors, a combination of Compaq and Dell servers were acquired. The backend database server was also kept open for the company to choose from and the existing relationships with Microsoft business partners were leveraged to cut the deal for a Microsoft SQL Server as the backend database server. Clearly the company's platform choice was Windows, as all the PCs were equipped with Windows O/S, Microsoft Office, and Windows NT/2000 as network operating systems. The company-wide email was supported by Microsoft Exchange server.

Towards the end of the design phase, the project team moved to the development phase. During this phase, the following activities were carried out:

a) Hardware set up

b) MS SQL Server installation and configuration on the database server

c) Installation and configuration of development and testing environment on separate servers

d) Preparation for the test user machines

e) Configuration of the SAP applications

f) Data migration and conversion for the existing applications

At the end of this phase, the project was completed in 32 weeks and the overall management was satisfied with the progress.

3.4.6 Specific ERP Implementation Issues with Oil and Gas Sector

An oil and gas control report system was installed, whose purpose was to maintain records to assess product quantities. A production scheduling component (software) for this sector must have the capability to record the movement of raw materials and intermediate products from one unit to the other. Most of the ERPs do not provide the features to capture such specific information particular to this sector. The organization developed their in-house systems for oil and gas control report system, operations management and production scheduling. During ERP implementations, if these specific information requirements are not correctly captured during the requirements analysis stage, it results in ERP implementation failure because the new ERP is not in a position to satisfy the information requirements of all stake holders [61]. Further, the development of interfaces and their testing requires more resources, effort and time, and problems in poorly-designed interfaces result in the failure of the entire project [61].

3.4.7 Implementation

Once the configuration of the SAP interfaces was completed, the initial user acceptance testing was conducted. As suggested by its name, stakeholders, IT managers and users play main roles in this stage [62]. The same team of focal points was used with a few added divisional users. Not much time was given for this testing as it was assumed that unchanged processes in SAP were already tested and confirmed. A list of target users was prepared for the system training in their respective areas. The project team struggled during this phase as the availability of the users was only 50% in all of the training sessions, despite management instructions to give full time to these sessions. The project adopted a 'train-the-trainers' approach where it was assumed that the selected users would train the rest of the staff in their divisions.

The system finally rolled-out in the 40th week. The whole SAP team's contract was extended for another year to provide continuous technical and functional support until the system matured. The company had great expectations for SAP and was aiming to collect immediate benefits after the implementation.

4 Discussions

The overall project achieved both of the primary goals – timeline and cost. However, post-implementation progress did not occur as the company expected. Many areas remained 'out of SAP', data residing in SAP was questionable in its accuracy, certain controls were still missing in SAP, and transactions were taking more time to complete in SAP, compared to the previous applications or manual processes.

When these issues were realized at top-management level, an SAP review committee was formed to conduct an assessment of the current situation and to develop an action plan. The team started working on the task and after assessing the situation and meeting with key staff, the following was presented to the management:

- The overall project lacked appropriate change management during its implementation. The SAP was definitely a transformational project for the company where its scope involved the company-wide processes and almost all the head office based employees were expected to use the system. Since ERP is a major investment of an organization and the implementation may involve substantial organizational changes, top management support was found to be a key success factor of success; but more importantly, top management needs to develop a shared vision and to communicate it to the employees so that expectations are clear [24], [46], [63]. A case study of 12

manufacturers found that a common characteristic of ERP projects which finished on time and on/under budget was the involvement of senior executives who also established clear priorities [3]. Laughlin [64] posits top management support as the first order of business for ERP; but what degree of involvement is appropriate? Jarvenpaa and Ives [65] found that "executive involvement" (a psychological state) is more strongly associated with the firm's progressive use of IT than executive participation (actual behaviour) in IT activities. In a survey of SAP users, Bradford and Florin [66] found that top management support was directly related to perceived organizational performance and user satisfaction. Thus, the expectation of both peers and top management may influence the behaviour of the ERP users [24]. However, in this case, very little effort was spent in planning the transition from its legacy/manual processes to a sophisticated ERP arena. The project's core focus remained on the timely completion of the project within the budget, rather than achieving the results. Mabert et al. [3] also found in their case study that because of the investment required for an ERP project, both in terms of resources and the resulting organizational changes, companies are very sensitive about implementation times and budgets.

- Another factor which was not considered was the employees' perception of SAP. The rumour had already been spread in the company that after SAP, the warehouse staff would be truncated to just 20% of the original staff. Similarly, the support staff in other areas like Finance, HR, and Material Management had a similar impression. Focal points that were a part of the project team were aware of the uncertain climate and may not have proactively quelled fears and rumours. As a result, the design phase remained weak and certain controls in SAP remained open. This caused the system to accept inaccurate data in some of the transactions, which created doubts about the integrity of the system. Compatibility between the new system and the existing business procedures and data format are the major issues reported by companies [67], [68]. Reimers [69] also observed that implementing an ERP system implies that master data are maintained in one department but are actually used by other departments; smooth master data maintenance involves a high degree of cross-functional collaboration and also understanding, which might be lacking in state-owned enterprises. Since ERP contains various modules that are intricately linked with each other, data should be managed properly to ensure their accuracy [70]. Here it is important to note that implementing an ERP will bring changes in the way people work within the organization, processes will change and there may be job cuts and rationalization of responsibilities within departments [71].
- The third very important factor was the reduced training time for the end users. Umble et al. [72] supported education/training as the most widely recognized critical success factor. A change management consultant observes that while shortening planned training may be the "fastest and least

expensive way" it may be "counter-productive in the long run" [30]. Here the project team wanted to complete the implementation phase and made an unfairly optimistic assumption about the 'train the trainers' approach. In order to provide a smooth access to ERP systems, a large number of elements must work closely together. These elements include support in hardware, software, training and information provision [24]. Reimers [69] also identified training as one of the critical success factors in ERP implementation. Bradford and Florin [66] surveyed SAP users and found a relationship between training, perceived organizational performance and user satisfaction. Bradley and Lee [73] found a positive relationship between user satisfaction with ERP training and user satisfaction with the installed ERP system at a university. A Gartner Group study indicates that 25% of the ERP budget should be dedicated to training users [74]. Yet, a study by Benchmarking Partners found that training averaged 8% of total project cost, but varied from 1% to 30% [75]. Somers and Nelson [76] found user training ranked 14 on the list of Critical Success Factors (CSFs) developed from a survey of senior IS executives involved in both on-going and completed implementations. It is a significant measure of successful ERP implementation to provide training to most employees to understand and use end-to-end business processes using the enterprise systems [25]. Incorrect mapping of business processes with application features may result in complete ERP failure because the system will not be able to capture all business processes according to company requirements. Management faces a dilemma between reducing the use of costly consultants and the lack of internal skills and knowledge to implement ERP [77].

- There was a need for developing many interfaces and their testing required more resources, effort and time. Legacy systems do not work in an integrated ERP environment. Due to lack of capabilities to record the oil and gas control specific information during implementation, it becomes difficult to fill all the required fields of new systems, due to which the data conversion stage faced a lot of dilemmas.
- The company had a mix of many nationalities and cultures, and not all employees had influence over others to train or convince them in their respective areas. Moreover, some of the trained employees viewed their new status as one of increased power within the company, and were reluctant to pass their new-found knowledge to their colleagues.

Conclusions

This study provides valuable insights into understanding ERP implementations and significant factors influencing their success. Various case studies provide different findings which are unique to ERP implementations because of the integrative characteristics of ERP systems. Alignment of the standard ERP processes with the company's business process has been considered as an

important step in the ERP implementation process [37]. After almost a year of implementation, the company has mixed results in this case. Certain areas have seen great improvements after the implementation of SAP (e.g. Procurement, Maintenance, Financial) where as certain areas remain weak (e.g. Employee Records, Contract Administration, Integrated Planning). From this implementation experience, it can be seen that it is not a particular technology platform or software application that can transform a company. Instead, it is the way the company implements the technology that makes it successful.

Acknowledgement

We would like to thank executive editor Dr. Péter Tóth and reviewers for their valuable comments to improve the quality of this paper.

References

[1] Lui, K. M., Chan, K. C. C. (2008) Rescuing Troubled Software Projects by Team Transformation: A Case Study with an ERP Project, IEEE Transaction Engineering Management, 55(1), 171-184

[2] Milford, M., Stewart, G. (2000) Are ERP Implementations Qualitatively Different from Other Large System Implementations?; Americas Conference on Information Systems, Long Beach California, 2000, 951-940

[3] Mabert, V. A., Soni, A., Venkataramanan (2003) Enterprise Resource Planning: Managing the Implementation Process. European Journal of Operational Research, 146, 302-314

[4] Mishra, A. (2008) Achieving Business Benefits from Enterprise Systems in Enterprise Resource Planning for Global Economies: Managerial Issues and Challenges, Carlos Ferran and Ricardo Salim, IGI Global, USA, ISBN: 1599045311, Chapter V, 76-91

[5] Mishra, A., Mishra, D. (2010) ERP System Implementation in FMCG sector, *Technical Gazette* 17, 1(2010), 115-120

[6] Yen, H. R., Sheu, C. (2004) Aligning ERP Implementation with Competitive Priorities of Manufacturing Firms: an Exploratory Study, International Journal of Production Economics, 92, 207-220

[7] Hunter, M. G., Lippert, S. K. (2007) Critical Success Factors of ERP Implementation. Information Resource Management Conference Vancouver, BC, Canada. Hershey, PA: IGI Publishing

[8] Luftman, J., Kempaiah, R., Nash, E. (2006) Key Issues for IT Executives 2005, MIS Quarterly Executive 2006; 5(2), 81-99

[9] Davenport, T. H. (1998) Putting the Enterprise into the Enterprise System. Harvard Business Review, 76(4), 121-32

[10] Motwani, J., Mirchandani, D., Madan, M., Gunasekaran, A. (2002) Successful Implementation of ERP Projects: Evidence from Two Case Studies, International Journal of Production Economics, 75(1-2), 83-96

[11] Mishra, A. (2009) Enterprise Resource Planning Systems: Effects and Strategic Perspectives in Organizations in Handbook of Research on Enterprise Systems, J. N. D. Gupta, S. K. Sharma, M. A. Rashid, IGI Global, USA, ISBN:978-1-59904-859-8., Chapter V, 57-66

[12] Mabert, V. A., Soni, A., Venkataramanan (2000) Enterprise Resource Panning Survey of US Manufacturing Firms. Production and Inventory Management Journal, 41(20), 52-58

[13] Mendelson, H. (1999) ERP Overview: Stanford Business School, Mimeo (1999)

[14] Martin (1998) An ERP Strategy. Fortune. v2. 95-97

[15] Zhang, M. K. O., Lee, L. (2003) Critical Success Factors of Enterprise Resource Planning Systems Implementation Success in China. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences

[16] Arif, M., Kulonda, D., Jones, J., Proctor, M. (2005) Enterprise Information Systems: Technology First or Process First? Business Process Management Journal 11(1), 5-21

[17] Griffith, T. L., Zammuto, R. F., Aiman-Smith, L. (1999) Why New Technologies Fail? Industrial Management, 29-34

[18] Markus, M. L., Robey, D. (1988) Information Technology and Organizational Change: casual Structure in Theory and Research, Management Science, 34, 583-598

[19] Volkoff, O. (1999) Enterprise System Implementation: A Process of Individual Metamorphosis, American Conference on Information Systems

[20] Nicolaou, A. I. (2004) Quality of Post Implementation Review for Enterprise Resource Planning Systems, International Journal of Accounting Information System, 5 (2004), 25-49

[21] Sekatzek E. P., Krcmar H. (2009) Measurement of the Standard Proximity of Adapted Standard Business Software. Business Information System Engineering, 1(3):234-244

[22] Liang, H., Saraf, N., Hu, Q., Xue, Y. (2007) Assimilation of Enterprise Systems: The effect of Institutional Pressures and the Mediating Role of Top Management. MIS Quarterly, 31(1), 59-87

[23] Xue, Y., Liang, H., Boulton, W. R., Snyder, C. A. (2005) ERP Implementation Failures in China: Case Studies with Implications for ERP Vendors. International Journal of Production Economics, 97(3), 279-295

[24] Chang, M. K., Cheung, W., Cheung, C-H, Yeung, J. H. Y. (2008) Understanding ERP System Adoption from the User's Perspective, International Journal of Production Economics, 113(2008), 928-942

[25] Kumar, V., Movahedi, B., Kumar, U., Lavassani, M. (2008) A Comparative Study of Enterprise System Implementations in Large North American Corporations. W. Abramowicz and D. Fansel (Eds.): BIS 2008, LNBIP 7, 390-398

[26] Hayes, S. (2007) Providing Enterprise Systems. Practical Accountant 40(2), SR11-SR11

[27] Hendricks, K. B., Singhal, V. R., Stratman, J. K. (2007) The Impact of Enterprise Systems on Corporate Performance: A Study of ERP, SCM, and CRM System Implementations. Journal of Operations Management, 25(1), 65-82

[28] Goldberg, A. (2000) The ERP Trap. Upside, 12(11), 32

[29] Krasner, H. (2000) ERP Experiences and Evolution. Communications of the ACM, 43(4), 22-26

[30] Wah, L. (2000) Give ERP a Chance. Management Review, 89(3), 20-24

[31] Yeh, T. M., Yang, C. C., Lin, W. T. (2007) Service Quality and ERP Implementation: A Conceptual and Empirical Study of Semiconductor-related Industries in Taiwan. Computers in Industry, 58(8-9), 844-854

[32] Bray, H. (2004) Computer Woes Cause Registration 'Nightmare' at University of Massachusetts, Knight Ridder Business News, Washington: Sept14, 2004:1

[33] Songini, M. (2004) ERP System Doesn't Make the Grade in Indians. Computerworld, p. 1, Sep13, 2004

[34] Nelson, R. R. (2005) Project Retrospectives: Evaluating Project Success, Failure, and Everything between. MIS Quarterly Executive, 4(3), 361-372

[35] Nelson, R. R. (2007) IT Project Management: Infamous Failures, Classic Mistakes, and Best Practices, MIS Quarterly executive, 6(2), 67-78

[36] Muscatello, J. R., Parente, D. H. (2006) Enterprise Resource Planning (ERP): A Postimplementation Cross-Case Analysis, Information Resource Management Journal, 19(3), 61-80

[37] Botta-Genoulaz, V., Millet, P. (2006) A Survey on the Recent Research Literature on ERP Systems. Computers in Industry, 95(2), 510-522

[38] Boonstra, A. (2006) Interpreting an ERP-implementation project from a stakeholder perspective. International Journal of Project Management, 24(2006), 38-52

[39] Fowler, A., Gilfillan, M. (2003) A Framework for Stakeholder Integration in Higher Education Information System Projects. Technol Anal Strategic Manage, 15(4), 467-89

[40] Markus, M. L., Tanis, C. (2000) Multisite ERP Implementations. Communications of ACM, 43(4), 26-42

[41] McAfee, A. (2003) When too Much IT Knowledge is a Dangerous Thing. Sloan Management Review, 44(2), 83-9

[42] Sirkin, H., Diekel, K. (2001) Getting Value from Enterprise Initiatives, Boston: Boston Consulting Group

[43] Markus, M. L., Axline, S., Petrie, D., Tanis, C. (2000) Learning from Adopters' Experiences with ERP: Problems Encountered and Success Achieved. Journal of Information Technology, 15(4), 245-265

[44] Genoulaz, V. B., Millet, P. A. (2006) An Investigation into the Use of ERP Systems in the Service Sector. International Journal of Production Economics, 99, 202-221

[45] Mische, R., Bennis, W. (1996) Reinventing through Reengineering. Information Systems Management, 13, 58-65

[46] Tchokogue, A., Bareil, C., Duguay, C. R. (2005) Key Lessons from the Implementation of an ERP at Pratt & Whitney Canada. International Journal of Production Economics, 95(2), 151-163

[47] Keil, M., Tiwana, A. (2006) Relative Importance of Evaluation Criteria for Enterprise Systems: A Conjoint Study. Information Systems Journal, 16(3), 237-262

[48] Rikhardsson, P., Kraemmergaard, P. (2006) Identifying the Impacts of Enterprise System Implementation and Use: Examples from Denmark. International Journal of Accounting Information Systems 7(1), 36-49

[49] Choi, J., Ashokkumar, S., Sircar, S. (2007) An Approach to Estimating Work Effort for Enterprise Systems Software Projects. Enterprise Information Systems, 1(1), 69-87

[50] Moore, J. (2008) Oil and Gas Sector Generates Big Business for Systems Integrators, SearchITChannel.com, available at http://searchitchannel.techtarget.com/news/article/0,289142,sid96_gci1334850,00.html

[51] Mishra, A, Mishra D. (2009) Customer Relationship Management – Implementation Process Perspective in *Acta Polytechnica Hungarica,* Vol. 6, Issue 4, 83-99

[52] Yin, R. K. (2003) Case Study Research: Design and Methods, 3rd Edition, Sage Publications, Thousands Oaks, CA

[53] Feagin, J., Orum, A., Sjoberg, G. (Eds.) (1991) A Case for Case Study, University of North Carolina Press, Chapel Hill, NC

[54] Sheu, C., Chae, B., Yang, C. L. (2004) National Differences and ERP Implementation: Issues and Challenges, Omega, 32(5), 361-37

[55] Sarker, S., Lee, A.S. (2003) Using a Case Study to Test the Role of Three Key Social Enablers in ERP Implementation, Information & Management, 40(8), 813-829

[56] Voordijk, H., Leuven, A. V., Laan, A. (2003) Enterprise Resource Planning in Large Construction Firm: Implementation Analysis, Construction Management & Economics, 21(5), 511-521

[57] Mishra, A., Mishra, D. (2009) ERP System Implementation: An Oil and Gas Exploration Sector Perspective, F. Bomarius et al. (Eds.): PROFES 2009, LNBIP 32, 416-428

[58] Jurison, J. (1999) Software Project Management: The Managers View, Communications of AIS, 2(17), 1-56

[59] Piturro, M. (1999) How Mid-Size Companies are Buying ERP, Journal of Accountancy, 188 (3), 41-48

[60] Chen, C. C., Law, Chuck C. H., Yang, S. C. (2009) Managing ERP Implementation Failure: A Project Management Perspective, IEEE Transaction on Engineering Management, 56(1), 157-170

[61] Nazir, M. M. (2005) ERP Implementation in Oil Refineries, *Business Recorder*, available at http://www.mubashirnazir.org/RM/R0009-00-ERP%20Implementation%20in%20Oil%20Refineries.htm

[62] Mishra, D, Mishra A. (2010) Improving Baggage Tracking, Security and Customer Services with RFID in the Airline Industry, *Acta Polytechnica Hungarica,* Volume 7 (2), 139-154

[63] Motwani, J., Subramanian, R., Gopalakrishna, P. (2005) Critical Factors for Successful ERP Implementation: Exploratory Findings from Four Case Studies. Computers in Industry, 56(6), 524-544

[64] Laughlin, S. P. (1999) An ERP Game Plan, Journal of Business Strategy, 20(1), 32-37

[65] Jarvenpaa, S. L., Ives, B. (1991) Executive Involvement and Participation in the Management of Information Technology, MIS Quarterly, 5(2), 205-227

[66] Bradford, M., Florin, J. (2003) Examining the Role of Innovation Diffusion Factors on the Implementation Success of Enterprise Resource Planning Systems, International Journal of Accounting Information Systems, 4(3), 205-225

[67] Soh, C., Kien, S. S., Tay-Yap, J. (2000) Cultural Fits and Misfits: Is ERP a Universal Solution? Communications of the ACM, 43 (4), 47-51

[68] Van Everdingen, Y. (2000) ERP Adoption by European Midsize Companies, Communications of the ACM, 43 (4), 27-31

[69] Reimers, K. (2002) Implementing ERP Systems in China, Proceedings of the 35th Hawaii International Conference on System Sciences, IEEE Computer Society

[70] Ngai, E. W. T., Law, C. C. H., Wat, F. K. T. (2008) Examining the Critical Success Factors in the Adoption of Enterprise Resource Planning, Computers in Industry, 59(2008), 548-564

[71] Otieno, J. O. (2008) Enterprise Resource Planning (ERP) Systems Challenges: A Kenyan Case Study. W. Abramowicz and D. Fansel (Eds.): BIS 2008, LNBIP 7, 399-409

[72] Umble, E. J., Haft, R. R., Umble, M. M. (2003) Enterprise Resource Planning: Implementation Procedures and Critical Success Factors, European Journal of Operations Research, 146(2003), 241-257

[73] Bradley, J., Lee, C. C. (2007) ERP Training and User Satisfaction: A Case Study, International Journal of Enterprise Information Systems, 3(4), 33-50

[74] Coetzer, J. (2000) Survey- Enterprise Resource Planning – Analyse before Implementing, Business Day, 2000 Sept. 18, p. 20

[75] Wheatly, M. (2000) ERP Training Stinks, CIO Magazine, June 1, 2000:13, 86-96

[76] Somers, T. M., Nelson, K. (2001) The Impact of Critical Success Factors across the Stages of Enterprise Resource Planning Implementations. The proceedings of the 34th Hawaii International Conference on System Science

[77] Haines, M. N., Goodhue, D. L. (2000) ERP Implementations: the Role of Implementation Partners and Knowledge Transfer. 11th Information Resource Management International Conference

Non-Conventional Approaches to Feature Extraction for Face Recognition

Jozef Ban[1], Matej Féder[2], Miloš Oravec[2], Jarmila Pavlovičová[1]

[1] Department of Telecommunications, [2] Department of Applied Informatics and Information Technology, Faculty of Electrical Engineering and Information Technology, STU in Bratislava, Slovak Republic
jozef.ban@stuba.sk, matej.feder@stuba.sk, milos.oravec@stuba.sk, jarmila.pavlovicova@stuba.sk

Abstract: This paper deals with human face recognition based on the use of neural networks, such as MLP (multi-layer perceptron), RBF (radial basis function) network, and SVM (support vector machine) methods. The methods are tested on the MIT (Massachusetts Institute of Technology) face database. We use non-conventional methods of feature extraction for the MIT face images in order to improve the recognition results. These methods use so called HLO and INDEX images. HLO images are generated by feature extraction of the MLP neural network in auto-association mode, and INDEX images are formed by a self-organized map used for image vector quantization. We propose novel methods based on HLO and INDEX images with SVM classifier. We also analyze the impact of adding noise to the learning process.

Keywords: human face recognition; feature extraction; multilayer perceptron; RBF networks; support vector machines

1 Introduction

Biometric face recognition [1] is a difficult and developing task even for the most advanced computer technology. While humans can recognize a familiar face in various lights and perspectives, there remain barriers to effective recognition by computers.

In this paper, we use and compare three methods of biometric face recognition. The first method is a neural network - multilayer perceptron (MLP), the second one is a radial basis function (RBF) neural network, and the third one is a support vector machine (SVM). For input data we used images of 64x60 pixels taken from the MIT (the Massachusetts Institute of Technology) database. The images were divided into two sets. The first set was designed for training, and the second (more extensive) set was designed for the recognition test itself. We attempted to achieve

the best results of biometric face recognition by setting the appropriate parameters for the MLP, RBF and SVM methods. For the MLP and RBF methods we used Matlab7 (http://www.mathworks.com/) software (Neural Network Toolbox). For the SVM method we used the Libsvm (A Library for Support Vector machines) freeware (http://www.csie.ntu.edu.tw/~cjlin/libsvm/).

The aim of the work is an evaluation of achieved results on the unmodified images taken from the MIT database and the analysis of the impact of proposed non-conventional methods of image feature extraction through HLO and INDEX images. Appropriate feature extraction [2], [3] reduces storage requirements and improves the computational and time complexity of used methods. We also evaluated the impact of adding Gaussian white noise to the training set of images, and its effect on the recognition results.

2 Neural Networks

A neural network [4], [5], [6], [7] is a massive parallel processor generally used to store experimental information for later use. It simulates the human brain in two aspects:

- a neural network obtains the information from an environment by a learning process,
- connections among neurons (synaptic weights SW) are used for information storage.

2.1 Multilayer Perceptron

The multilayer perceptron (MLP) [5], [6] is the first method used for face recognition in this paper.

This type of network consists of an input layer of elements which distribute input signals to the network, one or more hidden layers and one output layer of computational elements.

2.1.1 Multilayer Perceptron Characteristics

The existence of one (Fig. 1) or more hidden layers allows the network to learn complicated tasks, because it selects the most important features from the input samples. The term “hidden neuron” refers to a neuron that is not a part of either the input or output layer, thus inaccessible from the outside world. An important feature of this method is a high degree of network interconnection determined by network synapses. Another significant characteristic is that the model of each neuron in the network contains nonlinearity in its output. This nonlinearity must be smooth, i.e. differentiable everywhere.

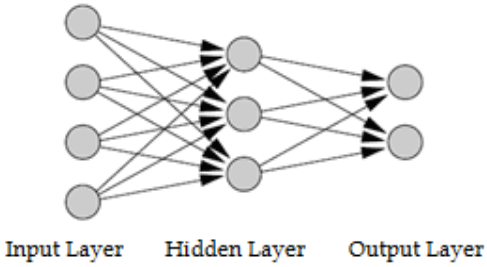


Figure 1
MLP with one hidden layer

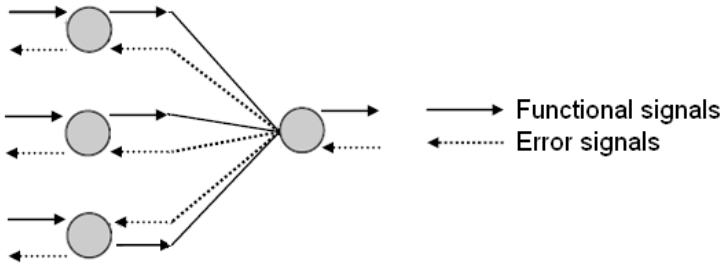


Figure 2
Functional and error signals in MLP

In the given part of the multilayer perceptron (Fig. 2), it is possible to notice the functional and error signals. The functional signal propagates forward starting at the network input and ending at the network output as an output signal. The error signal originates in the output neurons during learning and propagates backward.

Hidden and output neurons perform two computations during their training. The first is the computation of a functional signal which is the output of the neuron. It is expressed as a nonlinear function of the input signals and synaptic weights connected to this neuron. The second computation is an estimation of the instantaneous gradient vector, i.e. the error surface gradient with regard to the weights which are connected to the given neuron. This vector is necessary for the back propagation phase of errors.

2.1.2 Backpropagation Algorithm

The MLP is trained in a supervised manner (by a teacher). It uses a back propagation algorithm [5], [6]. The error signal at the output of neuron j for n-th training sample (iteration) is defined as:

$$e_j(n) = d_j(n) - y_j(n) \tag{1}$$

supposing that neuron j is an output neuron, d_j(n) is a desired and y_j(n) actual response.

The internal activity of neuron *j* is:

$$v_j(n) = \sum_{i=0}^{p} w_{ji}(n) y_i(n) \tag{2}$$

where p is the number of inputs of the neuron j and w_{ji}(n) is a synaptic weight connecting the output of neuron i to the input of neuron j in iteration n. The functional signal at the output of the neuron j with the activation function $\varphi_j()$ for iteration n is:

$$y_j(n) = \varphi_j(v_j(n)) \tag{3}$$

The adjustment Δw_{ji}(n) of the weight connecting neurons i and j is defined by delta rule:

$$\begin{pmatrix} \text{adjustment} \\ \text{of weight} \\ \Delta w_{ji}(n) \end{pmatrix} = \begin{pmatrix} \text{parameter} \\ \text{of fast learning} \\ \eta \end{pmatrix} * \begin{pmatrix} \text{local} \\ \text{gradient} \\ \delta_j(n) \end{pmatrix} * \begin{pmatrix} \text{input signal} \\ \text{of neuron j} \\ y_i(n) \end{pmatrix} \tag{4}$$

Thus, the backpropagation algorithm adjusts weight w_{ji}(n) by value Δw_{ji}(n), which is proportional to the instantaneous gradient $d\varepsilon$(n)/dw_{ji}(n).

2.1.3 Stopping Criterion for Learning

For the backpropagation algorithm, several stopping criteria exist. It is possible to set a maximum number of training cycles or some kind of a maximal output error. We used a cross-validation as the stopping criterion. It consists of dividing the available data to a training and a test sets. After a period of training, the network is tested on a test set in order to examine its generalization properties. The process of learning stops when there is an increase of error for the test set (this method is also known as the early stopping method of training). At that time the network has reached its maximum ability to generalize.

2.2 Radial Basis Function Network

A radial basis function network (RBF network) is a neural network in which the hidden neurons represent a set of functions forming a base for input vector transformation into the hidden neuron space. These functions are called RBF – radial basis functions [6], [8], [9], [10], [11]. The learning process is equivalent to searching the space which best approximates the training data. Generalization is then equivalent to a utilization of this space for an interpolation of testing data.

In many cases, better results are achieved by RBF networks than by networks with sigmoidal activation functions (e.g. MLP – multilayer perceptron). One possible

reason is that RBF activation functions respond better to the receptive fields of real neurons [6], [8].

An example of an RBF network is shown in Fig. 3. Formally, an RBF network can be described as follows [12]:

$$f(x) = w_0 + \sum_{i=1}^{m} w_i h_i(\mathbf{x}) \tag{5}$$

where **x** is a parameter of RB activation function h_i and w_i are weights. The output of the network is a linear combination of RBFs.

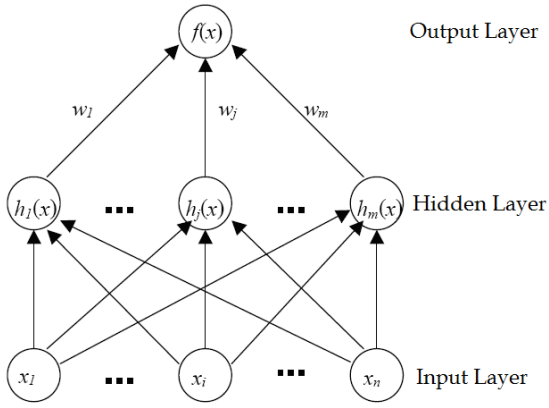


Figure 3
RBF network [12]

The training set for an RBF network consists of N input-output pairs $(\boldsymbol{x}_k, \boldsymbol{d}_k)$, $k=1,...,N$, where $\boldsymbol{x}_k \in R^p$ is input and $\boldsymbol{d}_k \in R^q$ is a desired response.

The RBF net is trained in three steps [6]:

- The first step is a determination of the hidden elements centers, which are represented by the weights between input and hidden layer. An easy solution is a random selection of n_0 points of the inputs, which we set as the centre values. Another approach generates uniform distribution of the centers in the input space.

  However, it is desirable that the centers comply with the structure of the input data – this requires the use of more complex algorithms in order to set up the centers. This task falls within a category of self-organizing learning.

- The second (optional) step is the setup of additional parameters of RBFs. Let us mention the often-used (even in our experiments) Gaussian radial basis function:

$$\phi(x) = \exp\left(-\frac{\|x - c\|^2}{\sigma^2} \right) \tag{6}$$

Its parameter σ represents a width of function ϕ which determines a radial space around the centre *c* in which a hidden element has a rational response. Widths of the functions influence the generalizing capacity of the net – the smaller width, the worse generalization is to be expected. Parameter σ was the object of our tests.

- In the third step of learning we set up values *w*, through the minimization of the error function *E*:

$$E(\mathbf{W}) = \frac{1}{2}\sum_{l=1}^{N} \|\mathbf{d}_l - f(\mathbf{x}_1)\|^2 = \frac{1}{2}\sum_{l=1}^{N}\sum_{k=1}^{q} (d_{kl} - f_k(\mathbf{x}_l))^2 \tag{7}$$

where $f_k(\mathbf{x}_l) = \mathbf{y}_l$ is a response (output) of the network to input $\mathbf{x}_l$, $f_k(\mathbf{x}_l) = y_{kl}$ is a response (output) of *k*-th output element to input $\mathbf{x}_l$, $\mathbf{W}$ is matrix $[w_{sr}]$ – i.e. with units w_{sr}, $s = 1,...,q$ and $r = 1,...,n_0$.

A solution obtained by $\frac{\partial E_3}{\partial w_{sr}} = 0$ is as follows:

$$\mathrm{w}_{\mathrm{sr}} = \sum_{j=1}^{n_0} [\Phi^+]_{jr} \left(\sum_{l=1}^{N} \phi_j(\mathbf{x}_l) d_{kl} \right) \tag{8}$$

where matrix $\mathbf{\Phi}$ is defined as

$$[\Phi]_{jr} = \sum_{l=1}^{N} \phi_j(\mathbf{x}_l)\phi_r(\mathbf{x}_l) \tag{9}$$

and $\mathbf{\Phi}^+$ is a pseudoinverse matrix.

3 Support Vector Machine

Support vector machines (SVM) [13], [14], [15], [16], [17] are based on the concept of decision planes that define optimal boundaries. The optimal boundary separates two different sets (sets of objects having different class membership) and is located as to achieve the largest distance between these sets. In the case of linearly nonseparable sets, a SVM uses kernel methods.

3.1 Kernel Methods

The soft margin method is an extension of a SVM within linear methods [18]. The kernel method is a method of finding non-linear thresholds. The basic concept of the kernel method is the transformation of vector space into a high-dimensional space. Let us consider the linearly non-separable example shown in Fig. 4(a) [13]. If the two-dimensional space is transformed into a three-dimensional space (Fig. 4(b)), the black and white vectors become linearly separable.

Φ is the transformation into a multidimensional space. The space which is to be transformed should match the distance defined in the transformed space and is related to the distance in the original space. Kernel function $K(\mathbf{x},\mathbf{x}')$, which meets both conditions is defined as follows:

$$K(\mathbf{x},\mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}') \tag{10}$$

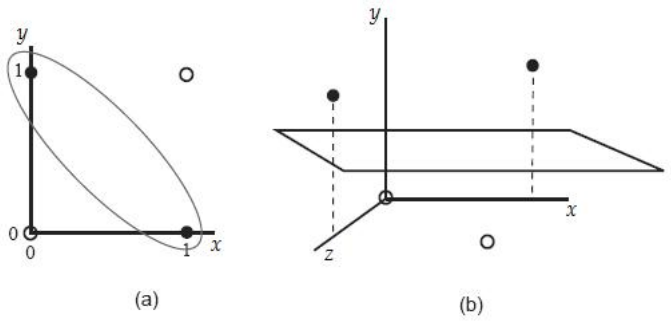


Figure 4
Linearly non-separable vector space (a), linearly separable vector space (b) [13]

This equation indicates that kernel function is equivalent to the distance between ***x*** and ***x'*** measured in high-dimensional space, transformed by Φ . If we measure the margin by the kernel function and perform an optimization, a nonlinear boundary is obtained.

The transformed space's threshold can be found by

$$\mathbf{w}^{\mathbf{T}}\Phi(\mathbf{x}) + b = 0 \tag{11}$$

It can be then formulated as:

$$\sum_i \alpha_i y_i \Phi(\mathbf{x}_\mathbf{i}^{\mathbf{T}})\Phi(\mathbf{x}) + b = \sum_i \alpha_i y_i K(\mathbf{x}_\mathbf{i}, \mathbf{x}) + b = 0 \tag{12}$$

Optimization function in transformed space is also obtained by substitution of $\mathbf{x}_i^T \mathbf{x}_j$ to $K(\mathbf{x}_i, \mathbf{x}_j)$. These results prove the fact that all the calculations can be

obtained just by using $K(\mathbf{x}_i, \mathbf{x}_j)$, without the exact formulation of Φ and the transformed space. K has to be positive definite (the sufficient condition).

Several examples of kernel function have been known, e.g.:

Polynomial kernel $K(\mathbf{x},\mathbf{x}') = (\mathbf{x}^{\mathbf{T}}\mathbf{x}'+1)^p$ (13)

Gaussian kernel $K(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{\sigma^2}\right)$ (14)

4 Used Data and Methods

4.1 The MIT Face Database

For the simulation of the biometric methods of face recognition we used a face database developed at MIT (the Massachusetts Institute of Technology) (http://web.mit.edu/). The MIT database contains 432 images. It consists of 16 subjects (Fig. 5), while each subject is represented by 27 images (Fig. 6). Each subject is distinguished by various head tilts, illumination and distances from the lens of the camera. Tests of individual methods use 256 level grayscale images of dimensions 64x60 pixels.

Figure 5
Face database subjects

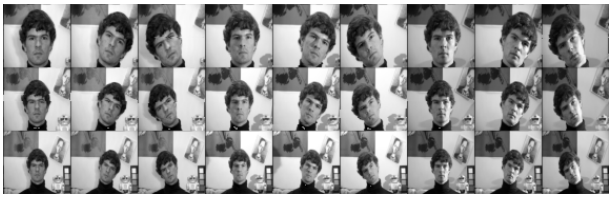


Figure 6
Various samples of the same subject

4.2 Novel Methods Based on Non-conventional Approaches to Feature Extraction

In order to improve recognition results, we use non-conventional methods of feature extraction for the MIT face images. These methods use so called HLO and INDEX images. HLO images are generated by feature extraction using the MLP neural network in auto-association mode and INDEX images are formed by a self-organized map (SOM) used for image vector quantization. We propose novel methods based on HLO and INDEX images with SVM classifier. The formation of HLO and INDEX images is described in sections 4.3 and 4.4. An overview of the methods used (two proposed methods using feature extraction and standard methods of direct classification) is shown in Fig. 7.

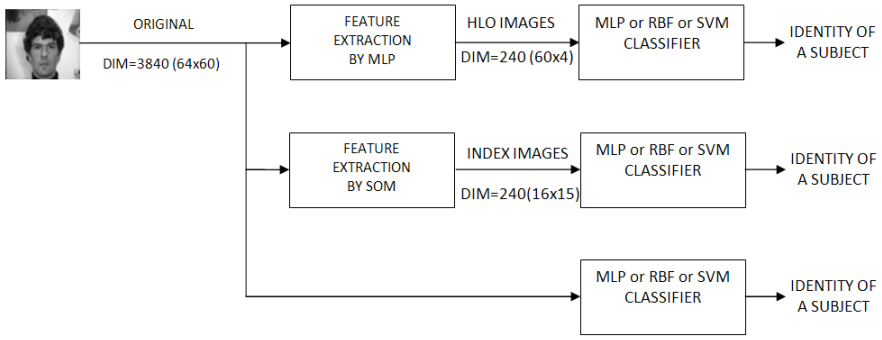


Figure 7

Overview of methods used for simulations:

- two proposed methods using feature extraction by MLP (top) and SOM (middle)
- standard methods of direct classification used for comparison purposes (bottom)

4.3 HLO Images

HLO (Hidden layer outputs) images [19] are the result of the feature extraction method by the MLP in auto-associative mode. Its configuration and utilization for feature extraction is shown in Fig. 8. Input face images of dimensions 64x60 pixels are divided into 16x15 blocks. The MLP configuration is 16x15-15-16x15 (i.e. 240 input and output neurons and 15 hidden neurons). For the MLP 240-15-240 and 64x60 images divided into blocks 16x15 pixels we obtain 16 blocks from each image and each block is represented by 15 hidden layer outputs. Each face image is thus represented by 240 hidden layer outputs, which are used for HLO image formation – resulting in 60x4 HLO images. This corresponds to a bit rate of 0.5 bit/pixel compared to 8 bit/pixel in the original images.

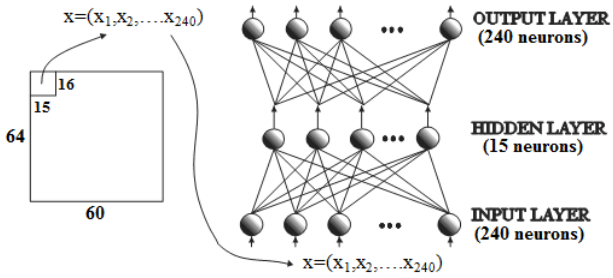


Figure 8
64x60 face image is divided into 16x15 blocks and then fed into the MLP network 240-15-240

Fig. 9 represents a sample of HLO images for 16 subjects of the MIT face database from Fig. 5.

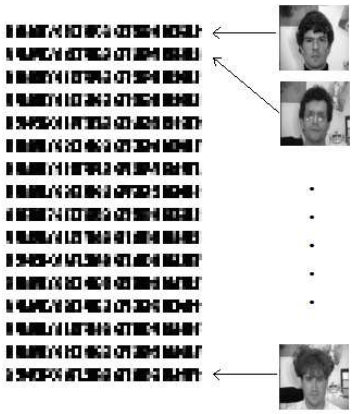


Figure 9
60x4 HLO images for 16 MIT face database subjects

4.4 INDEX Images

INDEX images result from feature extraction method based on the vector quantization (VQ) of images using Kohonen self-organizing map for codebook design. This method is described in [20]. Vector quantization is performed on 64x60 face images dividing the originals into 4x4 blocks. After vector quantization, all 240 indexes corresponding to all blocks of the face image are formed into a two-dimensional array of dimensions 16x15. For the image vector quantization, the configuration of the self-organizing map with 16x16 neurons with 16-dimensional weight vectors was used. It again corresponds to a bit rate of 0.5 bit/pixel. INDEX images corresponding to face images from Fig. 5 are shown in Fig. 10.

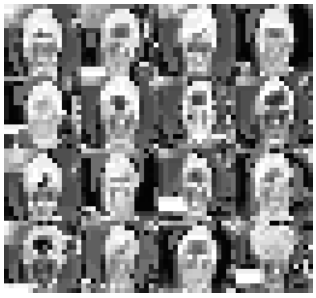


Figure 10
MIT database INDEX images (each of 16 images is 16x15 pixels)

4.5 Noise Modification of Images

We attempted to analyze the impact of adding noise to the learning process. Adding new training samples modified by the noise can lead to better convergence for some methods. For the modification of the MIT database images, Gaussian white noise [21] was used. It was applied with variance *v=0.02* and zero mean *m=0*. Gaussian white noise for a pixel *x* is defined as follows:

$$p_q(x) = (2\pi)^{-1/2} e^{-(x-m)^2/2v^2} \tag{15}$$

An example of the original and the noisy image is shown in Fig. 11.

Figure 11
Original and noisy MIT image 64x60 pixels, Gaussian white noise, v=0.02

4.6 Training and Test MIT Sets

For our simulations, two main training sets *a11* and *a11+112+113* were created. These training sets are described in Fig. 12. The set *a11* consists of all 111, 211 and 311 images. The set *a11+112+113* consists of a11 set plus all 112 and 113 images. These two main sets were extended by images modified by noise (according to Section 4.5). The same holds also for HLO and INDEX training sets.

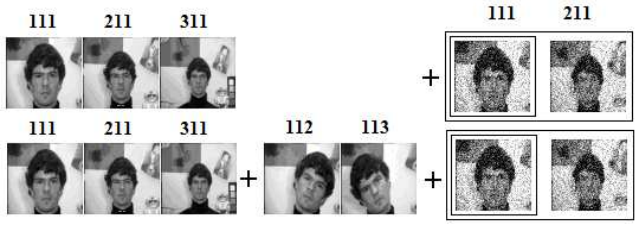


Figure 12
Illustration of the MIT database images division into training and test sets - a11 (111, 211, 311), number of images is 48 (3 views for each of the 16 subjects), a11+112+113 represented by 80 images (5 views for each of the 16 subjects)

4.7 Parameters of MLP, RBF and SVM Methods

Simulation results are influenced by the proper settings of the parameters of the individual methods:

- **MLP**

  The topology of the neural network, the learning rate η and training method (gradient descent back propagation, gradient descent with momentum and adaptive learning rate back propagation) are the parameters affecting the training process.

- **RBF**

  A decisive impact on the success of human face recognition for the RBF method is influenced by the number of neurons in the hidden layer of the network, and in our case also by a variance parameter of Gaussian curve σ .

- **SVM**

  The success of human face recognition in the SVM method depends on the setting of value *C* as a parameter of the SVM method, and on the optimization of the value γ (the parameter of the used kernel function).

5 Simulation Results

Three standard face recognition methods, MLP, RBF and SVM, used for the direct classification (the bottom method in Fig. 7) of the input images (i.e. without feature extraction), were compared. Figure 13 and Table 1 show the recognition results achieved by these methods. The SVM is the most successful method for face recognition (without feature extraction), with an average value of recognition

efficiency of 78.32%. The comparable results were achieved also by a method using the RBF network with an average value of 70.03%. The MLP method in comparison to the SVM and RBF achieved unstable and insufficient results with an average value of 44.60%.

We also analyzed the impact of extending training sets by modified images (with Gaussian noise). This expansion helped in several simulations with the methods MLP, SVM, and significantly increased the success of recognition with the RBF method.

The essential goal of the paper was to propose the methods using feature extraction through HLO and INDEX images followed by SVM classifier (the top and the middle method in Fig. 7). Interesting results were achieved by simulations for the proposed approaches. In Figure 13, the simulations on a smaller training set (a11) are denoted by the color black. The value of 72.14% (black underlined) obtained on the 64x60 pixel images by the RBF method has been overcome in five cases using HLO and INDEX images (SVM classifier was used in four cases). In the case of the larger training set (a11+112+113, brown underlined), the best value of 92.33% has been overcome only once - with the use of HLO images. In this case, however, the best recognition efficiency was achieved (93.18%).

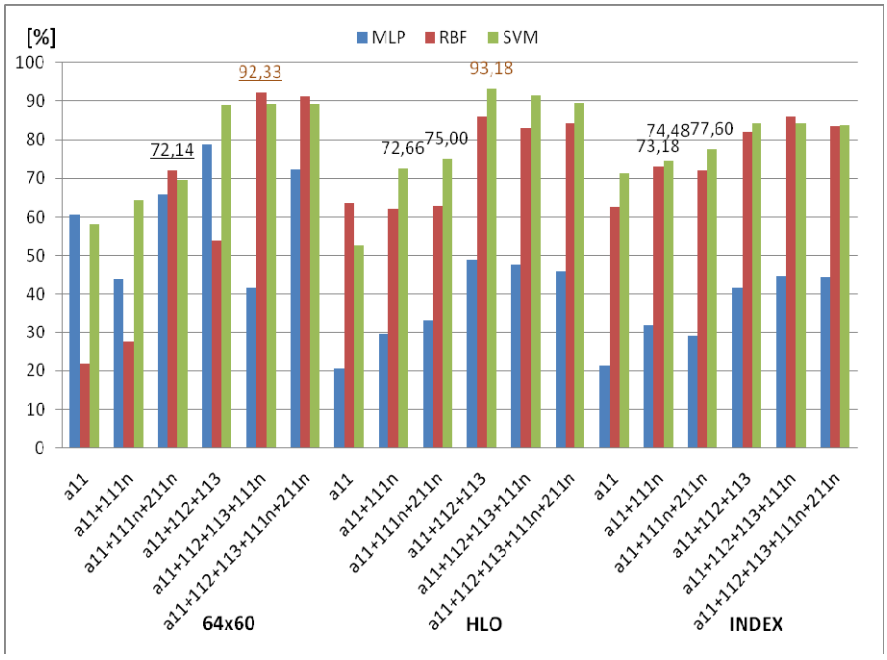


Figure 13

Graphic comparison of the results by methods MLP, RBF and SVM

Table I
The results of the MLP, RBF and SVM methods

		Recognition efficiency [%]		
Image	Training set	MLP	RBF	SVM
64x60	a11	60,6771	21,875	58,0729
64x60	a11+111n	44,03	27,6042	64,3229
64x60	a11+111n+211n	65,9	72,1354	69,5313
64x60	a11+112+113	78,6932	53,9773	88,9205
64x60	a11+112+113+111n	41,7	92,3295	89,2405
64x60	a11+112+113+111n+211n	72,4	91,1932	89,2405
HLO	a11	20,5729	63,5417	52,6042
HLO	a11+111n	29,4271	62,2396	72,6563
HLO	a11+111n+211n	33,3333	62,7604	75
HLO	a11+112+113	48,8636	86,0795	93,1818
HLO	a11+112+113+111n	47,7273	82,9545	91,4773
HLO	a11+112+113+111n+211n	46,0227	84,375	89,4886
INDEX	a11	21,3542	62,5	71,3542
INDEX	a11+111n	32,0313	73,1771	74,4792
INDEX	a11+111n+211n	28,9063	72,1354	77,6042
INDEX	a11+112+113	41,7614	82,1023	84,375
INDEX	a11+112+113+111n	44,7917	86,0795	84,375
INDEX	a11+112+113+111n+211n	44,5313	83,5227	83,8068

Conclusion

In this paper, we proposed novel methods for face recognition based on non-conventional feature extraction followed by the SVM classifier. We used MLP for the formation of HLO (hidden layer output) images and Kohonen SOM for the formation of INDEX images. Such representation was fed into MLP, RBF and SVM classifiers. The presented results show excellent recognition efficiency and they were compared to results of three standard methods using direct classification by MLP, RBF and SVM. For a small training set, results on the proposed methods using HLO and INDEX images overcame in five cases results of direct classification. For a larger training set, the method using HLO images achieved the best results of all. SVM was the most successful classifier for HLO and INDEX images. The advantages of the proposed methods are as follows: the reduction of storage requirements, the speed-up of the learning process and the improvement of computational and time complexity.

Acknowledgement

Research described in the paper was done within the grants No. 1/0214/10 and 1/0961/11 of the Slovak Grant Agency VEGA.

References

[1] Jain, A. K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition, IEEE Trans. Circuits and Systems for Video Technology, Vol. 14, No. 1, Jan. 2004, pp. 4-20

[2] Puyati, W., Walairacht, S., Walairacht, A.: PCA in Wavelet Domain for Face Recognition, The 8th International Conference Advanced Communication Technology, p. 455, Phoenix Park, ISBN: 89-5519-129-4, 2006

[3] Prema, R., Thirunadanasikamani, K., Suguna, R.: A Novel Feature Extraction Scheme for Face Recognition, 2010 International Conference on Signal and Image Processing (ICSIP), pp. 502-505, 15-17 Dec. 2010

[4] Bishop, C. M.: Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1996

[5] Haykin, S.: Neural Networks - A Comprehensive Foundation, New York: Macmillan College Publishing Company, 1994

[6] Oravec, M., Polec, J., Marchevský, S.: Neural Networks for Digital Signal Processing (in Slovak), Bratislava, 1998

[7] Oravec, M., Petráš, M., Pilka, F.: Video Traffic Prediction Using Neural Networks, Acta Polytechnica Hungarica, Journal of Applied Sciences, Hungary, Vol. 5, Issue 4, pp. 59-78, 2008

[8] Hlaváčková, K., Neruda, R.: Radial Basis Function Networks, Neural Network World, No. 1, 1993, pp. 93-102

[9] Poggio, T., Girosi, F.: Networks for Approximation and Learning, Proc. of the IEEE, Vol. 78, No. 9, Sept. 1990, pp. 1481-1497

[10] Zhang, Y., Xue, Z. M.: RBF Neural Network Application to Face Recognition, 2010 International Conference on Challenges in Environmental Science and Computer Engineering (CESCE), Vol. 2, pp. 381-384, 6-7 March 2010

[11] Benyó, B., Somogyi, P., Paláncz, B.: Classification of Cerebral Blood Flow Oscillation, Acta Polytechnica Hungarica, Journal of Applied Sciences, Hungary, Vol. 3, Issue 1, pp. 159-174, 2006

[12] Mark, J. L. Orr: Introduction to Radial Basis Function Networks. Centre for Cognitive Science, University of Edinburgh, April 1996

[13] Asano, A.: Pattern Information Processing, (lecture 2006 Autumn semester), Hiroshima University, Japan, 2006, http://laskin.mis.hiroshima-u.ac.jp/Kougi/06a/PIP/PIP12pr.pdf

[14] Hsu, C. W., Chang, C. C. Lin, C. J.: A Practical Guide to Support Vector Classification. Department of Computer Science and Information Engineering, National Taiwan University, 2003

[15] Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B.: An Introduction to Kernel-based Learning Algorithms. IEEE Transactions on Neural Networks, Vol. 12, No. 2, March 2001, pp. 181-201

[16] Boser, B. Guyon, I. Vapnik, V.: A Training Algorithm for Optimal Margin Classifiers, In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM Press, 1992, pp. 144-152

[17] Faruqe, M. O., Al Mehadi Hasan, M.: Face Recognition Using PCA and SVM, 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication, ASID 2009, pp. 97-101, 20-22 Aug. 2009

[18] Horváth, G.: Kernel CMAC: an Efficient Neural Network for Classification and Regression, Acta Polytechnica Hungarica, Journal of Applied Sciences, Hungary, Vol. 3, Issue 1, pp. 5-20, 2006

[19] Oravec, M., Pavlovičová, J.: Feature Extraction by Multilayer Perceptron - Visualization of Internal Representation of Input Data, The Seventh IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP 2007), August 29-31, 2007, Palma de Mallorca, Spain, ISBN Hardcopy: 978-0-88986-691-1 / CD: 978-0-88986-692-8, pp. 112-117

[20] Oravec, M.:A Method for Feature Extraction from Image Data by Neural Network Vector Quantization, Proc. of the 6th International Workshop on Systems, Signals and Image Processing, June 2-4, 1999, Bratislava, Slovakia, pp. 73-76

[21] Bovik, C. A.: Handbook of Image and Video Processing 2nd edition. Elsevier Academy Press, 2005, pp. 397-409

A Fully-coupled Thermo-Hydro-Mechanical Model for the Description of the Behavior of Swelling Porous Media

Hanifi Missoum, Nadia Laredj, Karim Bendani, Mustapha Maliki

Construction, Transport and Protection of the Environment Laboratory (LCTPE)
Université Abdelhamid Ibn Badis de Mostaganem, Algeria
hanifimissoum@yahoo.fr, nad27000@yahoo.fr, bendanik@yahoo.fr, mus27000@yahoo.fr

Abstract: Although many numerical models have been proposed for unsaturated porous media, unrealistic assumptions have been made, such as the non-deformable nature of media, constant material properties, the neglecting of convection heat flow transfer, and static air phase. Most of these conditions are not justifiable for porous media with low hydraulic permeability and high swelling activity. In the present article, a fully coupled thermo-hydro-mechanical model is proposed that takes into account nonlinear behavior, including both the effects of temperature on dynamic viscosity of liquid water and air phase, and the influence of temperature gradient on liquid and air flows. Fully coupled, nonlinear partial differential equations are established and then solved by using a Galerkin weighted residual approach in space domain and an implicit integrating scheme in time domain. The obtained model is finally validated by means of some case tests for the prediction of the thermo-hydro-mechanical behaviour of unsaturated swelling soils.

Keywords: multiphase flow; water transfer; unsaturated porous media; finite element; conservation; simulation

1 Introduction

Swelling porous materials are commonly found in nature as well as developed in industry. They are studied in many disparate fields including in soil science, in hydrology, in forestry, in geotechnical, chemical and mechanical engineering, in condensed matter physics, in colloid chemistry and in medicine. This article specifically focuses on unsaturated swelling clays, which are widely distributed in nature. In agriculture, water adsorption by the clay determines the ability of soils to transport and supply water and nutrients. Compacted bentonites play a critical role in various high level nuclear waste isolation scenarios and in barriers for

commercial landfills [1, 2]. In engineering and construction, swelling and compaction of clayey soils induce stresses which are very troublesome in foundation and structure buildings.

The engineering behaviour of unsaturated soil has been the subject of numerous experimental and theoretical investigations [3-8]. Numerous researches have been undertaken on both the experimental and theoretical aspects of thermo-hydro-mechanical transfer processes in porous media. Firstly, many of those studies have analysed the coupling behaviour on saturated media [9-11] based on Biot theory. Some of these investigations were founded on small temperature gradients assumptions and non-convective heat flow; further, no phase fluid change was taken into account and the physical material properties were constant.

Finite element solutions for non-isothermal two phase flows of deformable porous media were developed by [12], where the pore-air pressure is atmospheric condition [13]. In most of the studies [9-12], temperature effect on dynamic viscosity and permeability were neglected.

In the present article, a fully coupled thermo-hydro-mechanical model is proposed, which takes into account nonlinear behaviour, including the effects of temperature on the dynamic viscosity of both liquid water and air phases, as well as the influence of temperature gradient on liquid and air flows. The fully coupled, nonlinear partial differential equations are established and then solved by using a Galerkin weighted residual approach in space domain and an implicit integrating scheme in time domain. The obtained model has been finally validated by means of some case tests for the prediction of the thermo-hydro-mechanical behaviour of unsaturated swelling soils.

2 Theoretical Formulation

In this work a three-phase porous material consisting of solid, liquid and air requires consideration. A set of coupled governing differential equations are presented below to describe coupled multiphase flow in the soil. The model is based on combinations of equations or derivations from conservation principles and the classical laws of known physical phenomena for the coupled flow. Governing differential equations for pore water, pore air and heat transfer in unsaturated soil are derived as follows:

2.1 Heat Transfer

Considering heat transfer by means of conduction, convection and latent heat of vaporization effects, and applying the principle of conservation of energy, the following equation is derived:

$$\frac{\partial \phi}{\partial t} = -\nabla . Q \tag{1}$$

Where ϕ is the heat content of the soil and Q is the total heat flux, defined as:

$$\begin{aligned} Q = &-\lambda_T \nabla T - \left(v_v \rho_v + v_a \rho_v\right) L + \\ &\left(C_{pl} v_l \rho_l + C_{pv} v_v \rho_l + C_{pv} v_a \rho_v + C_{pda} v_a \rho_{da}\right)\left(T - T_r\right) \end{aligned} \tag{2}$$

$$\phi = H_c \left(T - T_r\right) + L n S_a \rho_v \tag{3}$$

where H_c is the specific heat capacity of the soil, T is the temperature, T_r is the reference temperature, L is the latent heat of vaporization of soil water, C_{pl}, C_{pv} and C_{pa} are the specific heat capacity of soil water, soil vapour and soil dry air respectively and λ_T is the coefficient of thermal conductivity of the soil.

Three modes of heat transfer are included: thermal conduction, sensible heat transfer associated with liquid, vapour and air flow and latent heat flow with vapour.

2.2 Moisture Transfer

For the moisture transfer, the mass transfer balance equation, accommodating both liquid and vapour, can be expressed as:

$$\frac{\partial(\rho_l n S_l)}{\partial t} + \frac{\partial(\rho_v n(S_l - 1))}{\partial t} = -\rho_l \nabla . v_l - \rho_l \nabla . v_v - \rho_v \nabla . v_a \tag{4}$$

where n is the porosity, ρ is the density, S_l is the degree of saturation, t is time and v the velocity. The subscripts l, a and v refer to liquid, air and water vapour respectively.

In this simulation, a generalized Darcy's law is used to describe the velocities of pore water and air:

$$v_l = -\frac{K_l}{\gamma_l}(\nabla u_l + \gamma_l \nabla z) \tag{5}$$

$$v_a = -K_a \nabla\left(\frac{u_a}{\gamma_a}\right) \tag{6}$$

where K_l and K_a are the hydraulic conductivities of liquid and air, respectively, γ_l is the unit weight of liquid, γ_a is the unit weight of air, u_l is the pore water pressure, u_a is the pore air pressure, z is the elevation and ∇ is the gradient operator.

The hydraulic conductivities of water and air through soil may be expressed in terms of the saturation degree or water content as follows:

$$K_l = K_l(S_l) \tag{7}$$

$$K_a = K_a(S_l, \eta_a) \tag{8}$$

where η_a is the dynamic viscosity of air.

The water vapour density ρ_v is evaluated from the thermodynamic assumptions, and when liquid and vapour phases are in equilibrium, it can be evaluated by the following relationship [14]:

$$\rho_v = \rho_0 . h_t \tag{9}$$

where h_t is the total relative humidity calculated by the following expression:

$$h_t = \exp\left(\frac{u_l - u_a}{\rho_l R_v T}\right) \tag{10}$$

and ρ_0 is the total saturated water vapour defined as [14]:

$$\rho_0 = \left[194.4 \exp\left(-0.06374(T-273) + 0.1634.10^3 (T-273)^2\right)\right]^{-1} \tag{11}$$

R_v is the gas constant for water vapour end T is the temperature.

2.3 Pore Air Mass Transfer

Using Henry's law to take account of dissolved air in the pore water, the following equation is derived for the dry air phase from the principle of mass conservation:

$$\frac{\partial\left[n\rho_a(S_a + HS_l)\right]}{\partial t} = -\nabla.\left[\rho_a(v_a + Hv_l)\right] \tag{12}$$

where, H is Henry's volumetric coefficient of solubility and ρ_a is the dry air density.

The dry air density ρ_a can be evaluated from Dalton's law as:

$$\rho_a = \frac{u_a}{RT} - \frac{R_v}{R}\rho_v \tag{13}$$

R and R_v are the gas constants for dry air and water vapour respectively, and T is the temperature.

2.4 Constitutive Stress-Strain Relationship

For problems in unsaturated swelling porous media, the total strain ε is assumed to consist of components due to suction, temperature and stress changes. This can be given in an incremental form as:

$$d\varepsilon = d\varepsilon_{\sigma} + d\varepsilon_{s} + d\varepsilon_{T} \tag{14}$$

where the subscripts σ, s and T refer to net stress, suction and temperature contributions.

The stress-strain relationship can therefore be expressed as:

$$d\sigma'' = D(d\varepsilon - d\varepsilon_{s} - d\varepsilon_{T}) \tag{15}$$

where

$$[\sigma''] = [\sigma_x \;\; \sigma_y \;\; \sigma_z \;\; \tau_{xy} \;\; \tau_{yz} \;\; \tau_{xz}] \tag{16}$$

where σ'' is the net stress and D is the elastic matrix. A number of constitutive relationships can be employed, for example an elasto-plastic constitutive relationship [15].

2.5 Coupled Equations

This leads to a set of coupled, nonlinear differential equations, which can be expressed in terms of the primary variables T, u_l, u_a and u of the model as energy balance:

$$\begin{aligned} C_{Tl}\frac{\partial u_l}{\partial t} + C_{TT}\frac{\partial T}{\partial t} + C_{Ta}\frac{\partial u_a}{\partial t} + C_{Tu}\frac{\partial u}{\partial t} \\ = \nabla[K_{Tl}\nabla u_l] + [K_{TT}\nabla T] + [K_{Ta}\nabla u_a] + J_T \end{aligned} \tag{17}$$

Mass balance:

$$\begin{aligned} C_{ll}\frac{\partial u_l}{\partial t} + C_{lT}\frac{\partial T}{\partial t} + C_{la}\frac{\partial u_a}{\partial t} + C_{lu}\frac{\partial u}{\partial t} \\ = \nabla[K_{ll}\nabla u_l] + [K_{lT}\nabla T] + [K_{la}\nabla u_a] + J_l \end{aligned} \tag{18}$$

$$C_{al}\frac{\partial u_l}{\partial t} + C_{aT}\frac{\partial T}{\partial t} + C_{aa}\frac{\partial u_a}{\partial t} + C_{au}\frac{\partial u}{\partial t} = \nabla[K_{al}\nabla u_l] + [K_{aa}\nabla u_a] + J_a \tag{19}$$

Stress equilibrium:

$$C_{ul}du_l + C_{uT}du_T + C_{ua}du_a + C_{uu}du + db = 0 \tag{20}$$

where K_{ij} and C_{ij} represent the corresponding terms of the governing equations ($i, j = l, T, a, u$)

3 Discretisation Techniques

The numerical solution of the theoretical models commonly used in geo-environmental problems is often achieved by a combination of numerical discretisation techniques. For the example presented in this paper, the finite element method is employed for the spatial discretisation and a finite difference time stepping scheme for temporal discretisation.

In particular, the Galerkin weighted residual method [16] is used to formulate the finite element discretisation. An implicit mid-interval backward difference algorithm is implemented to achieve temporal discretisation since it has been found to provide a stable solution for highly non-linear problems [17]. With appropriate initial and boundary conditions the set of typically nonlinear coupled partial differential equations can be solved.

Applying a Galerkin formulation of the finite element method, we obtain a system of matrix equations are as follows:

$$[K]\{\varphi\}+[C]\{\dot{\varphi}\}+\{J\}=0 \tag{21}$$

where

$$[K]=\begin{bmatrix} K_{TT} & K_{Tl} & K_{Ta} & 0 \\ K_{lT} & K_{ll} & K_{la} & 0 \\ 0 & K_{al} & K_{aa} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},\ [C]=\begin{bmatrix} C_{TT} & C_{Tl} & C_{Ta} & C_{Tu} \\ C_{lT} & C_{ll} & C_{la} & C_{lu} \\ C_{aT} & C_{aL} & C_{aa} & C_{au} \\ C_{uT} & C_{ul} & C_{ua} & C_{uu} \end{bmatrix} \tag{22}$$

$$\{\varphi\}=\begin{pmatrix} T & u_l & u_a & u \end{pmatrix}^T,\ \{\dot{\varphi}\}=\begin{pmatrix} \frac{dT}{dt} & \frac{du_l}{dt} & \frac{du_a}{dt} & \frac{du}{dt} \end{pmatrix}^T,$$

$$\{J\}=\begin{pmatrix} J_T & J_l & J_a & J_u \end{pmatrix}^T \tag{23}$$

To solve equation (21), a general form of fully implicit mid-interval backward difference time stepping algorithm is used to discretise the governing equation temporally. Therefore equation (21) can be rewritten as:

$$K\left(\varphi^n\right)\left[1-\theta\right]\left\{\varphi^{n+1}\right\}+\theta\left\{\varphi^n\right\}+C\left(\varphi^n\right)\left[\frac{\left\{\varphi^{n+1}\right\}-\left\{\varphi^n\right\}}{\Delta t}\right]+J\left(\varphi^n\right)=\{0\} \tag{24}$$

$\left(\varphi^n\right)$ is the level of time at which the matrices *K*, *C* and *J* are to be evaluated and is given by:

$$\left(\varphi^n\right)=\omega\left\{\varphi^{n+1}\right\}+(1-\omega)\left\{\varphi^n\right\} \tag{25}$$

where ω is integration factor, which defines the required time interval ($\omega \in [0,1]$) and $\theta = 0, 0.5, 1$ for backward, central and forward difference schemes, respectively.

For a mid-interval backward difference scheme, $\omega = 0.5$ and $\theta = 0$. Therefore, equation (24) reduces to:

$$K\left(\frac{\{\varphi^{n+1}\}+\{\varphi^{n}\}}{2}\right)\{\varphi^{n+1}\}+C\left(\frac{\{\varphi^{n+1}\}+\{\varphi^{n}\}}{2}\right)\left(\frac{\{\varphi^{n+1}\}-\{\varphi^{n}\}}{\Delta t}\right) + J\left(\frac{\{\varphi^{n+1}\}+\{\varphi^{n}\}}{2}\right)=\{0\} \tag{26}$$

Eq. (26) may be rewritten in alternate form as:

$$K^{n+0.5}\{\varphi^{n+1}\}+C^{n+0.5}\left(\frac{\{\varphi^{n+1}\}-\{\varphi^{n}\}}{\Delta t}\right)+J^{n+0.5}=\{0\} \tag{27}$$

A solution for $\{\varphi^{n+1}\}$ can be obtained provided the matrices *K, C* and *J* at time interval ($n+0.5$) can be determined. This is achieved by the use of a predictor-corrector iterative solution procedure.

4 Applications and Results

4.1 Example 1

4.1.1 Problem Definition

The following example is investigated to demonstrate swelling pressure calculation by using the back hydro-mechanical model. The simulation begins with isothermal two phase flow coupled with deformation. Free extension is allowed at the first stage to calculate the free extension displacement on the boundary. The geometric set-up and boundary conditions are as shown in Fig. 1. The example is a compacted bentonite block, 0.025 m length and 0.024 m height. The element discretization is $\Delta x = 0.00125\,m$ and $\Delta y = 0.00124\,m$, eight nodded composed elements. The initial conditions of the system are: atmospheric air pressure and liquid saturation $S_l = 0.357$. A water solution enters the sample from the bottom under pressure described by the curve in Fig. 2. The corresponding part on the boundary is fully saturated.

The material properties for this example are based on data in the literature [18, 1] and summarized in Table 1. The deformation under free extension conditions is assumed to be non-linear elastic.

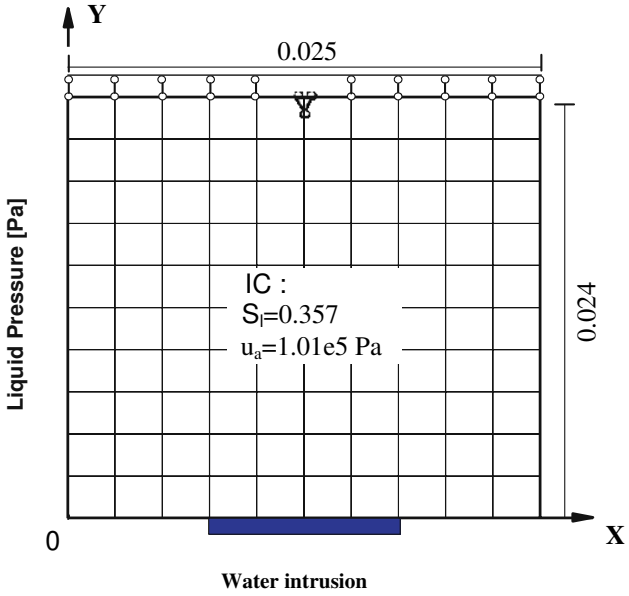


Figure 1
Model set-up of the example

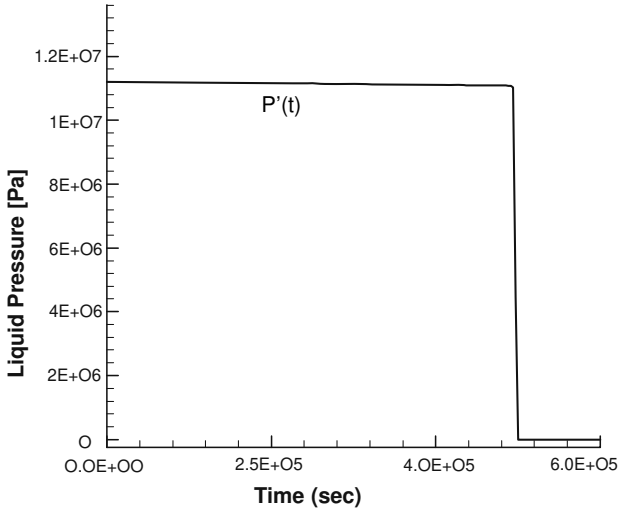


Figure 2
Curve of the liquid pressure on the boundary

Symbols	functions and constants	Units
ρ_l	1000	Kg/m^3
ρ_g	1.26	Kg/m^3
ρ_s	1600	Kg/m^3
S_l	$1-0.85\left[1-\exp\left(-1.20\times10^{-7}\left(u_a-u_l\right)\right)\right]$	
μ_l	1.20×10^{-3}	Pa s
μ_g	1.80×10^{-5}	Pa s
K_l	$\frac{1.2\times10^{-12}}{\mu_l[1+1.3\times10^{-10}\left(u_a-u_l\right)^{1.7}]}$	m/s
K_a	$1.3\times10^{-19}\frac{\gamma_a}{\mu_a}\left[e\left(1-S_l\right)\right]^{3.0}$	m/s
n	0.37	
S_0	31.80	m^2/g
E	3.5	MPa
ν	0.3	
T_r	293	°K

Table 1
Porous medium properties [18, 19]

4.1.2 Results and Discussion

The simulation results of the free extension processes after $5.8\times10^5\,s$ (6.7 days) are shown in Figs. 3 and 4. With the intrusion of water from the bottom, the saturation process starts. This phenomenon can clearly be seen from the saturation evolution profile along the vertical symmetric axis (Fig. 3). At the early stage, the value of liquid saturation increases quite fast. After $4.9\times10^5\,s$ (5.7 days) the liquid pressure on the bottom sinks at $5.0\times10^5\,s$ (5.8 days) reaches zero. Because of the low permeability of bentonite, the liquid pressure at the centre of the specimen sinks slower, which results in the higher pressure region in the specimen after 6.7 days as shown in Fig. 4a.

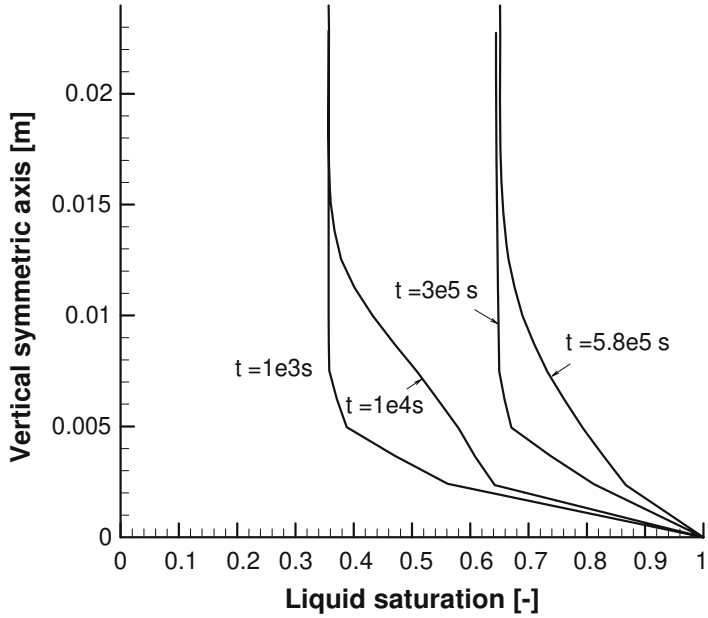


Figure 3
Computed profiles of liquid saturation along the vertical symmetric axis

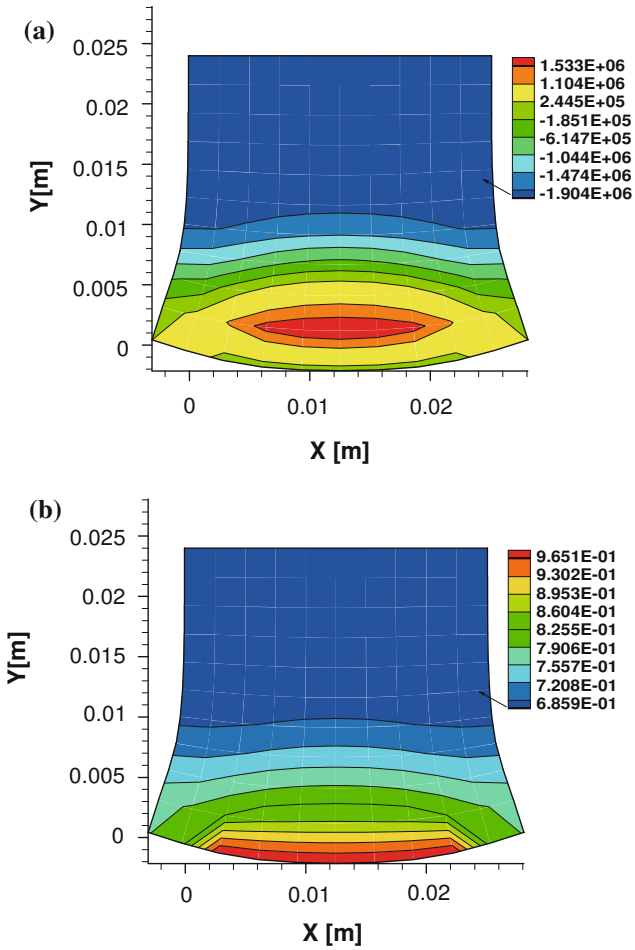


Figure 4
Simulation results of the free extension process
(a) Distribution of liquid pressure, (b) Distribution of liquid saturation

With the intrusion of water, the sample begins to expand. After 6.7 days, the shape of the specimen should be as shown in Fig. 5. The maximal width of the specimen increases about 20% (Fig. 5). In this example, experimental data from free swelling tests for compacted bentonites were used [18].

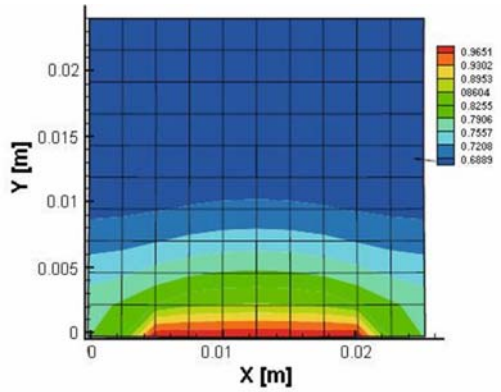


Figure 5
Simulated shape of the sample, and distribution of liquid saturation for material at $t = 5.8 \times 10^5$ s

4.2 Example 2

4.2.1 Problem Definition

The model has been verified through simulating a test problem of bentonite tested in laboratory conditions [20]. The compacted bentonite was planned to be used to limit the flow velocity of groundwater in the near field of a geological repository for radioactive waste. The sample of bentonite has a height of 203 mm. To minimize heat losses, the cell was insulated with a heat-proof envelope. Heat was applied at the bottom plate of the cylinder while the temperature at the other end was kept constant and equal to 20°c. A maximum temperature of 150° was applied. A constant water pressure was applied to the end opposite the one where the temperature variation was prescribed. Constant volume conditions were ensured in the test. The main variables that were measured during the test include temperature, relative humidity and total axial stress. The model geometry is the same as the sample with the height equal to 203 mm. The temperature variation at the bottom end of the model is as specified during the test; it was raised in steps until reaching 150°c. The temperature at the top end of the specimen was kept constant at 20°c. The other surfaces are adiabatic.

According to the test procedure, the hydraulic boundary condition is impermeable for all surfaces of the model. The gas pressure is equal to the atmospheric pressure. The initial value of porosity is 0.3242, initial temperature is 20°c, initial gas pressure is 0.10132 MPa and the initial stress of sample is 0.5 MPa. The initial values were obtained according to the measured data from the experiments [19].

According to the measured data from the bentonite [20, 21, 22], the intrinsic permeability is 1.0E-21 m^2 and the specific heat capacity of solids is 920 J/Kg.

4.2.2 Results and Discussion

Fig. 6 shows the comparison between measured temperatures as a function of time. At the bottom end of the sample, the measured and simulated temperatures agree well. At the top end of the sample, the simulation underestimates the temperature slightly at the initial heating stage, but trends generally agree well.

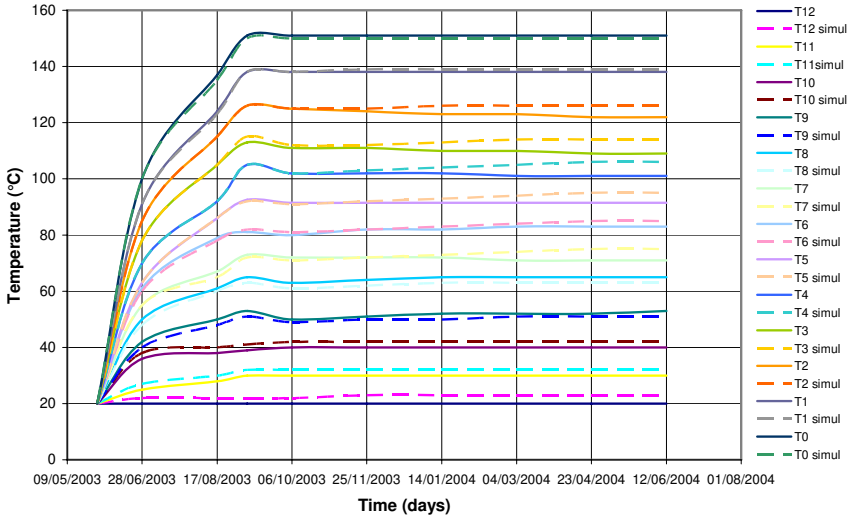


Figure 6

Comparison of results between the simulated and measured temperature at different locations (heights) of the sample as function of time

The good agreement between the simulated and measured temperatures indicates that the model can simulate the thermal response of processes having high temperature gradients. Numerical simulation shows that the thermal conductivity of bentonite, treated as a multiphase material, plays an important role in the thermal response of the whole medium. The numerical results can be further improved if the effects of the material heterogeneity can be quantified during testing.

The simulated results of the vapour pressure are shown in Fig. 7. Although there are no measured results with which to compare, it shows that the variation of vapour pressures is realistic.

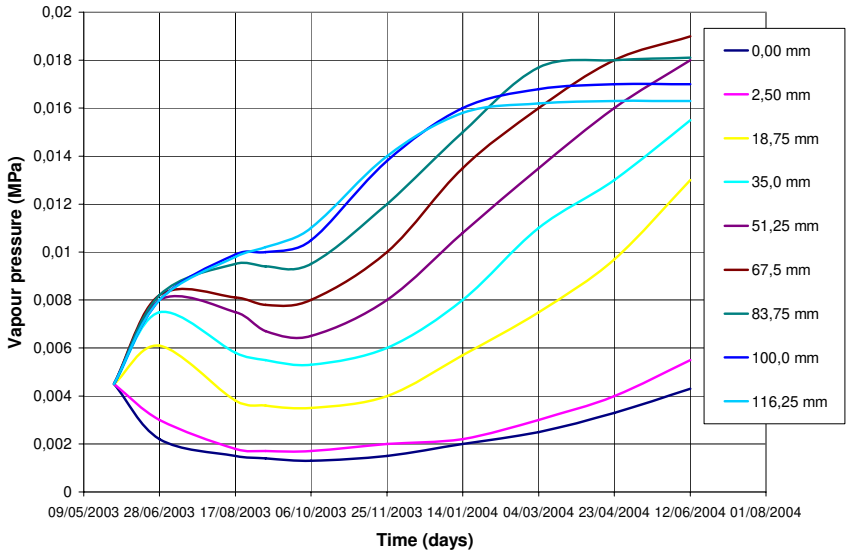


Figure 7
The simulated results of vapour pressure

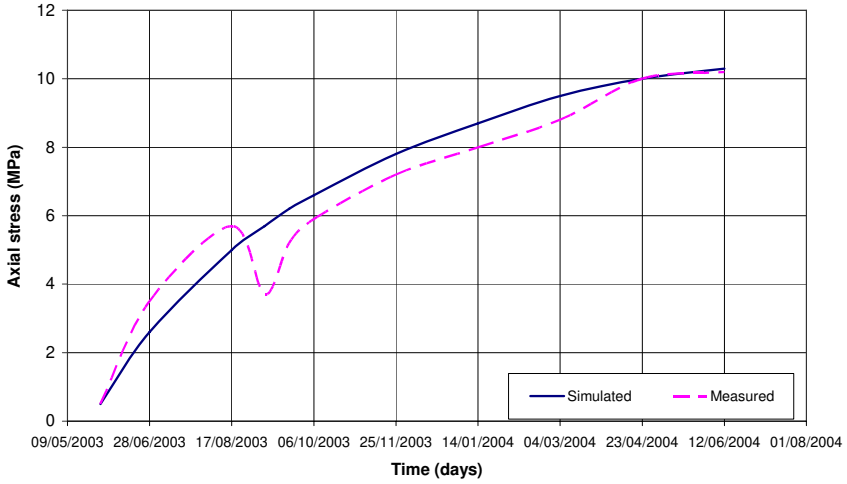


Figure 8
Comparison of results between simulated and measured axial stress

The evolution of the axial stress was also reasonably well simulated (Fig. 8), in trend and magnitude. Numerical calculation shows the results of stress strongly depend on the constitutive relationship between stress and strain. More comprehensive development of the material models is needed to further improve the numerical capability of the code, such as elastoplastic models for example.

Conclusion

A fully coupled thermo-hydro-mechanical model is proposed, which takes into account nonlinear behaviour including the effects of temperature on dynamic viscosity of both liquid water and air phases, and the influence of temperature gradient on liquid and air flows. A set of fully coupled, nonlinear partial differential equations were established and then solved by using a Galerkin weighted residual approach in the space domain and using an implicit integrating scheme in time domain. A range of simulation results has been presented detailing the hydraulic and thermal behaviour of expansive soil. The simulation results have been compared with the experimentally measured results and it was shown that a good correlation was found in the hydraulic regime and a reasonable correlation in the thermal field. The results of the validations indicate that the model is general and suitable for the analyses of many different problems in unsaturated soils.

References

[1] Gens A, Olivella S: Chemo-Mechanical Modelling of Expansive Materials. 6th International Workshop on Key Issues in Waste Isolation Research. Paris (France), 2001, pp. 463-495

[2] Andra : Dossier argile. Evaluation de la faisabilité du stockage géologique en formation argileuse, Rapport Andra C.RP.ADP.04.0002,2005

[3] Fredlund DG, Morgenstern NR: Stress State Variables for Unsaturated Soils. J. Geotch. Engng. Div., ASCE 103, 1977, pp. 447-466

[4] Alonso EE, Gens A, Josa A: A Constitutive Model for Partially Saturated Soils. Geotechnique 40, 1990, pp. 405-430

[5] Cui J, Delage P: Yielding and Plastic Behaviour of an Unsaturated Compacted Silt. Geotechnique, 1996, 46 (2): pp. 291-311

[6] Cui YJ, Yahia-Aissa M, Delage P: A Model for the Volume Change Behaviour of Heavily Compacted Swelling Clays. Engineering Geology, 2002, 64 (2): pp. 233-250

[7] Fleureau JM, Verbrugge JC, Huergo PJ, Correia AG, Kheirbek Saoud S: Aspects of the Behaviour of Compacted Clayey Soils on Drying and Wetting Paths. Can. Geotech. Engng., 2002, 39: pp. 1341-1357

[8] Saiyouri N, Tessier D, Hicher PY: Experimental Study of Swelling in Unsaturated Compacted Clay. Clay minerals, 2004, 39 (4): pp. 469-479

[9] Detournay E, Cheng A H D: Poroelastic Response of a Borehole in a Non-Hydrostatic Stress Field, International Journal of Rock Mechanics and Mining sciences and Geomechanics Abstracts, 1988, 25 (3): pp. 171-182

[10] Cheng A H D, Abousleiman Y, Roegiers J C: Review of Some Poroelatic Effects in Rock Mechanics. International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts, 1993, 30 (7): pp. 1119-1126

[11] Senjuntichai T: Green's Functions for Multi-layered Poroelastic Media and an Indirect Boundary Element Method, PhD Thesis. Winnipeg: University of Manitoba, 1994

[12] Britto A M, Savvidou C, Gunn M J: Finite Element Analysis of the Coupled Heat Flow and Consolidation around Hot BURIED objects. Soils and Foundations, 1992, 32 (1): pp. 13-25

[13] Chen W, Tan X, Yu H, Jia S: A Fully Coupled Thermo-Hydro-Mechanical Model for Unsaturated Porous Media. Journal of Rock Mechanics and Geotechnical Engineering, 2009, 1 (1): pp. 31-40

[14] Edlefsen NE, Anderson ABC: The Thermodynamics of Soil Moisture Hilgardia, 1943, pp. 31-299

[15] Thomas HR, He H: Modelling the Behaviour of Unsaturated Soil Using an Elastoplastic Constitutive Relationship. Geotechnique, 1998, 48 (5): pp. 589-603

[16] Zienkiewicz OC, Taylor RL: The Finite Element Method. Butterworth-Heinemann, 5th ed., Oxford, 2000

[17] Thomas HR, King SD: Simulation of Fluid Flow and Energy Processes Associated with High Level Radioactive Waste Disposal in Unsaturated Alluvium. Water Ressour. Res., 1991, 22 (5): pp. 765-775

[18] Agus S, Schanz T: Swelling Pressures and Wetting-Drying Curves of a Highly Compacted Bentonite-Sand Mixture, in: Schanz T (2003) Unsaturated Soils: Experimental Studies, Proceedings of the International Conference. From Experimental Evidence Towards Numerical Modelling of Unsaturated Soils. Weimar, Germany, September 18-19, 2003, Volume 1 Series: Springer Proceedings in Physics (93) Volume Package: Unsaturated Soils: Experimental Studies 2005, XIV, p. 533, Springer

[19] Poling BE, Prausnitz JM, O'Connell JP: The Properties of Gases and Liquids, Mc Graw Hill, 5th ed., New York, 2001

[20] Gatabin C, Billaud P: Bentonite THM mock up experiments. Sensor data report. CEA, Report NT-DPC/SCCME 05-300-A, 2005

[21] Loret B, Khalili N: A three phase model for unsaturated soils. Int. J. Numer. Anal. Meth. Geomech; 2000, 24: pp. 893-927

[22] Marshall TJ, Holmes JW: Soil physics. Bristol. Cambridge University Press, 2nd ed., New York, 1988

The Modeling and Simulation of an Autonomous Quad-Rotor Microcopter in a Virtual Outdoor Scenario

Aleksandar Rodić[1], Gyula Mester[2]

[1] University of Belgrade, Institute Mihajlo Pupin, Robotics Laboratory, Belgrade, Serbia, aleksandar.rodic@pupin.rs

[2] University of Szeged, Institute of Informatics, Department of Technical Informatics, Robotics Laboratory, Szeged. Hungary, gmester@inf.u-szeged.hu

Abstract: This paper presents the modeling and simulation of an autonomous quad-rotor microcopter in a virtual outdoor scenario. The main contribution of this paper focuses on the development of a flight simulator to provide an advanced R/D tool suitable for control design and model evaluation of a quad-rotor systems to be used for control algorithm development and verification, before working with a real experimental system. The main aspects of modeling of rotorcraft kinematics and rigid body dynamics, spatial system localization and navigation in a virtual outdoor scenario are considered in the paper. Some high-level control aspects are considered, as well. Finally, several basic maneuvers (examples) are investigated and simulated in the paper to verify the simulation software capabilities and engineering capabilities.

Keywords: modeling; simulation; autonomous quad-rotor microcopter; Xaircraft X 650; rotorcraft; virtual outdoor scenario; quad-rotor dynamics; spatial navigation; GPS coordinates; GPS navigation, simulation example

1 Introduction

Over the past several decades, a growing interest has been shown in robotics. In fact, several industries (automotive, medical, manufacturing, space, . . .) require robots to replace men in dangerous, boring or onerous situations. A wide area of this research is dedicated to aerial platforms. Several structures and configurations have been developed to allow 3D movements [1]-[11]. For example, there are blimps, fixed-wing planes, single rotor helicopters, bird-like prototypes, quad-rotors, etc. Each of these has advantages and drawbacks. The vertical take-off and landing requirements exclude some of the aforementioned configurations. However, the platforms which show these characteristics have a unique ability for vertical, stationary and low speed flight.

The quad-rotor architecture has been chosen for this research for its low dimension, good maneuverability, simple mechanics and payload capability. As the main drawback, the high energy consumption can be mentioned. However, the trade-off results are very positive. This structure can be attractive in several applications, in particular for surveillance, for imaging dangerous environments, and for outdoor navigation and mapping (Fig. 1).

Figure 1
Quad-rotor XAircraft X650

The study of the kinematics and dynamics helps to understand the physics of the quad-rotor and its behavior [1], [2]. Together with modeling, the determination of the control algorithm structure is very important for improving stabilization. The whole system can be tested thanks to a Matlab-Simulink program that is interfaced with the remote controller. This software provides a 3D graphic output as well as status data for debugging the system performance. The real platform is developed by creating a system (integration) of interconnected devices. Two types of sensors are used for measuring the robot attitude and for measuring its height from the ground. For the first, an Inertial Measurement Unit (IMU) was adopted, while the distance was estimated with a SOund Navigation And Ranging (SONAR) and an InfraRed (IR) modules. The data processing and the control algorithm are handled in the Micro Control Unit (MCU) which provides the signals to the motors. Actually, four motor driver boards are needed to amplify the power delivered to the motors. Their rotation is transmitted to the propellers which move the entire structure (see Fig. 1).

The paper is organized as follows: Section 1: Introduction. In Section 2, the modeling of the Quad-rotor aircraft and the control strategy are presented. In Section 3, the GPS navigation of the Quad-rotor is illustrated. In Section 4, the simulation results are illustrated. Conclusions are given in Section 5.

2 Modeling of the Quad-Rotor Aircraft

Rotary wing aerial vehicles have distinct advantages over conventional fixed wing aircrafts in surveillance and inspection tasks because they can take-off and land in limited spaces and easily fly above the target. A quad-rotor is a four rotor helicopter. An example of one is shown in Fig. 1. Helicopters are dynamically unstable and therefore suitable control methods are needed to make them stable. Although unstable dynamics is not desirable, it is good from the point of view of agility. The instability comes from changes in the helicopter parameters and from disturbances such as a wind gust or air density variation. A quad-rotor helicopter is controlled by varying the rotor speed, thereby changing the lift forces [1], [2]. It is an under-actuated dynamic vehicle with four input forces and six outputs coordinates. One of the advantages of using a multi-rotor helicopter is the increased payload capacity. Quad-rotors are highly maneuverable, which allows for vertical take-off/landing, as well as flying into hard-to-reach areas; but the disadvantages are the increased helicopter weight and increased energy consumption due to the extra motors. Since the machine is controlled via rotor-speed changes, it is more suitable to utilize electric motors. Large helicopter engines, which have a slow response, may not be satisfactory without incorporating a proper gear-box system.

Unlike typical helicopter models (and regular helicopters), which have variable pitch angles, a quad-rotor has fixed pitch angle rotors, and the rotor speeds are controlled in order to produce the desired lift forces. The basic motions of a quad rotor can be described using the model presented in Figs. 2 and 3.

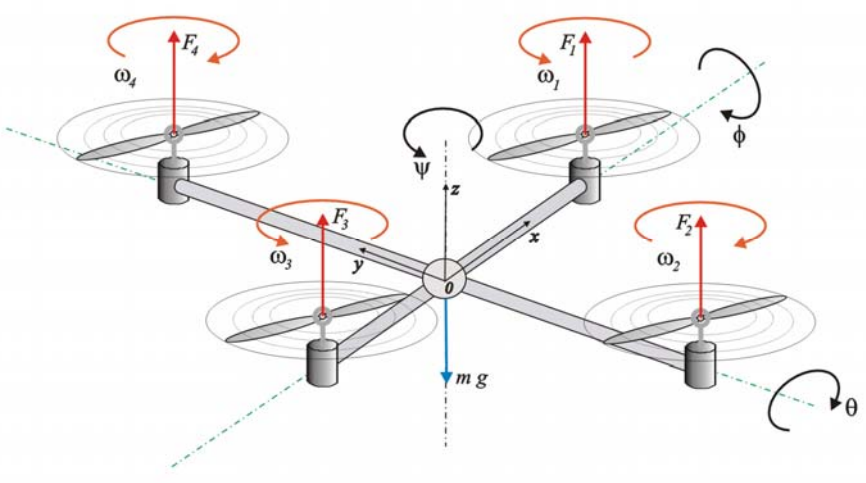


Figure 2
3 D motion, commonly used model of the quad-rotor

In the first method, the vertical motion of the helicopter can be achieved by changing all of the rotor speeds at the same time. Motion along the x-axis (Fig. 2) is related to tilt around the y-axis. This tilt can be obtained by decreasing the speed of propeller 1 and increasing corresponding speed of propeller 2. This tilt also produces acceleration along the x-axis. Similarly, y-motion is the result of the tilt around the x-axis. A good controller should be able to reach a desired yaw angle while keeping the tilt angles and the height constant.

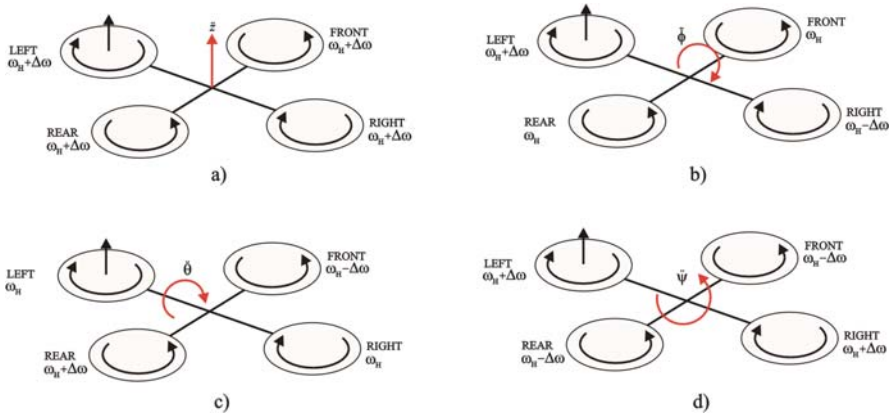


Figure 3
Definition of a) throttle, b) roll, c) pitch and d) yaw movements

2.1 Analysis of Possible Movements

The quad-rotor is satisfactorily modeled with four rotors in a cross configuration (Fig. 1). This cross structure is quite thin and light; however, it shows robustness via mechanically linking the motors (which are heavier than the structure). Each propeller is connected to the motor through the reduction gears. All the propellers have fixed and parallel axes of rotation. Furthermore, they have fixed-pitch blades and their air flows points downwards (to get an upward lift). These considerations point out that the structure is quite rigid, and the only things that can be varied are the propeller speeds. In this section, neither the motors nor the reduction gears are important to consider because the movements are directly related only to the propeller-velocities. Another neglected component is the electronic box. As in the previous instance, the electronic box is not important for understanding how the quad-rotor flies. It thus follows that the basic model for evaluating the movements of a quad-rotor is composed only of a thin cross structure with four propellers on the ends. The front and the rear propellers rotate counter-clockwise, while the left and the right ones turn clockwise. This configuration of pairs moving in opposite directions removes the need for a tail rotor (which is needed in the standard helicopter structure). Fig. 1 shows the structure model in hovering condition

where all the propellers have the same speed $\omega_i = \omega_H$, $i = 1,...,4$. In Fig. 2 all the propellers rotate at the same (hovering) speed ω_H (rad/s) to counterbalance the acceleration due to gravity. Thus, the quad-rotor performs stationary flight and no forces or torques move it from its position. Even though, the quad-rotor has 6 DOFs, it is equipped just with four propellers, hence it is not possible to reach a desired set-point for all the DOFs, but at maximum four. However, thanks to its structure, it is quite easy to choose the four best controllable variables and to decouple them to make the control task easier. The four quad-rotor targets are thus related to the four basic movements which allow the helicopter to reach a certain height and attitude. The description of these basic movements follows [3]:

2.1.1 The Throttle Movements

This command is provided by increasing (or decreasing) all the propeller speeds by the same amount. It leads to a vertical force with respect to body-fixed frame which raises or lowers the quad-rotor. If the helicopter is in a horizontal position, the vertical direction of the inertial frame and that one of the body-fixed frame coincide. Otherwise the provided thrust generates both vertical and horizontal accelerations in the inertial frame. Fig. 3a shows the throttle command in the quad-rotor sketch. The speed of the propellers ω_i, $i = 1,...,4$ in this case are equal to $\omega_H + \Delta\omega$ for each. The $\Delta\omega$ (rad/s) is a positive variable which represents an increment with respect to the constant value. The $\Delta\omega$ must not be too large because the model would eventually be influenced by strong non-linearities or saturations.

2.1.2 The Roll Movements

This command is provided by increasing (or decreasing) the speed of the left propeller and by decreasing (or increasing) the speed of the right one. This leads to torque with respect to the $x-$ axis (Fig. 2), which makes the quad-rotor turn. The overall vertical thrust is the same as in hovering, hence this command leads only to a roll angle acceleration (in the first approximation). Figure 3b shows the roll command on a quad-rotor sketch. The positive variable $\Delta\omega$ is chosen to maintain the vertical thrust unchanged. As in the previous case, they must not be too large because the model would eventually be influenced by strong non-linearities or saturations.

2.1.3 The Pitch Movements

This command is very similar to the roll and is provided by increasing (or decreasing) the speed of the rear propeller and by decreasing (or increasing) the speed of the front one. This leads to torque with respect to the $y-$ axis which makes the quad-rotor turn. The overall vertical thrust is the same as in hovering,

hence this command leads only to a pitch angle acceleration (in the first approximation). Figure 3c shows the pitch command on a quad-rotor sketch. As in the previous case, the positive variable $\Delta\omega$ is chosen to maintain the vertical thrust unchanged, and it cannot be too large.

2.1.4 The Yaw Movements

This command is provided by increasing (or decreasing) the front and rear propellers' speed and by decreasing (or increasing) that of the left-right couple. It leads to torque with respect to the $z-$ axis which makes the quad-rotor turn. The yaw movement is generated thanks to the fact that the left-right propellers rotate clockwise while the front-rear ones rotate counter-clockwise (Fig. 3d). Hence, when the overall torque is unbalanced, the helicopter turns on itself around z. The total vertical thrust is the same as in hovering, hence this command leads only to a yaw angle acceleration (in the first approximation). Figure 3d shows the yaw command on a quad-rotor sketch.

2.2 The Newton-Euler Model

To describe the motion of a 6 DOF rigid body it is usual to define two reference frames (Fig. 4):

- the earth inertial frame (E-frame), and
- the body-fixed frame (B-frame)

The equations of motion are more conveniently formulated in the B-frame because of the following reasons:

- The inertia matrix is time-invariant.
- Advantage of body symmetry can be taken to simplify the equations.
- Measurements taken on-board are easily converted to body-fixed frame.
- Control forces are almost always given in body-fixed frame.

The E-frame (*OXYZ*) is chosen as the inertial right-hand reference. *Y* points toward the North, *X* points toward the East, *Z* points upwards with respect to the Earth, and *O* is the axis origin. This frame is used to define the linear position (in meters) and the angular position (in radians) of the quad-rotor. The B-frame (*oxyz*) is attached to the body. *x* points toward the quad-rotor front, *y* points toward the quad-rotor left, *z* points upwards and *o* is the axis origin. The origin *o* is chosen to coincide with the center of the quad-rotor cross structure. This reference is right-hand, too. The linear velocity v (m/s), the angular velocity Ω (rad/s), the forces *F* (N) and the torques *T* (Nm) are defined in this frame. The linear position of the helicopter (*X, Y, Z*) is determined by the coordinates of the vector between the

origin of the B-frame and the origin of the E-frame according to the equation. The angular position (or attitude) of the helicopter (ϕ, θ, ψ) is defined by the orientation of the B-frame with respect to the E-frame. This is given by three consecutive rotations about the main axes which take the E-frame into the B-frame. In this paper, the "roll-pitch-yaw" set of Euler angles were used. The vector that describes the quad-rotor position and orientation with respect to the E-frame can be written in the form:

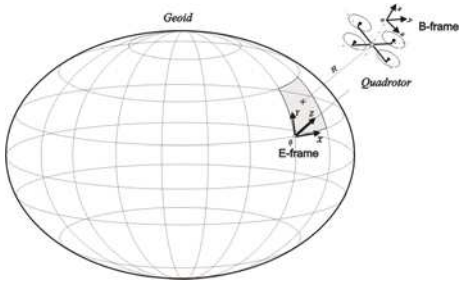


Figure 4
Earth- and Body-frame used for modeling of the quad-rotor system

$$s = \left[X\ Y\ Z\ \phi\ \theta\ \psi\right]^T \tag{1}$$

The rotation matrix between the E- and B-frames has the following form [3]:

$$R = \begin{bmatrix} c_\psi c_\theta & -s_\psi c_\phi + c_\psi s_\theta s_\phi & s_\psi s_\phi + c_\psi s_\theta c_\phi \\ s_\psi c_\theta & c_\psi c_\phi + s_\psi s_\theta s_\phi & -c_\psi s_\phi + s_\psi s_\theta c_\phi \\ -s_\theta & c_\theta s_\phi & c_\theta c_\phi \end{bmatrix} \tag{2}$$

The corresponding transfer matrix has the form:

$$T = \begin{bmatrix} 1 & s_\phi t_\theta & c_\phi t_\theta \\ 0 & c_\phi & -s_\phi \\ 0 & s_\phi / c_\theta & c_\phi / c_\theta \end{bmatrix} \tag{3}$$

In the previous two equations (and in the following) this notation has been adopted: $s_{(.)} = \sin(.)$, $c_{(.)} = \cos(.)$, $t_{(.)} = \tan(.)$. Now, the system Jacobian matrix, taking (2) and (3), can be written in the form:

$$J = \begin{bmatrix} R & 0_{3x3} \\ 0_{3x3} & T \end{bmatrix} \tag{4}$$

where 0_{3x3} is a zero-matrix. The generalized quad-rotor velocity in the B-frame has a form [3]:

$$v = \left[\dot{x}\ \dot{y}\ \dot{z}\ \dot{\phi}\ \dot{\theta}\ \dot{\psi}\right]^T \tag{5}$$

Finally, the kinematical model of the quad-rotor can be defined in the following way:

$$\dot{s} = J \cdot v \tag{6}$$

The dynamics of a generic 6 DOF rigid-body system takes into account the mass of the body m (kg) and its inertia matrix I ($Nm\ s^2$). Two assumptions have been done in this approach:

- The first one states that the origin of the body-fixed frame is coincident with the center of mass (COM) of the body. Otherwise, another point (COM) should be taken into account, which could make the body equations considerably more complicated without significantly improving model accuracy.
- The second one specifies that the axes of the B-frame coincide with the body principal axes of inertia. In this case the inertia matrix I is diagonal and, once again, the body equations become simpler.

The dynamic model of a quad-rotor can be defined in the following matrix form:

$$M_B\dot{v} + C_B(v)v - G_B = \Lambda \tag{7}$$

where M_B is the system Inertia matrix, C_B represents the matrix of Coriolis and centrifugal forces and G_B is the gravity matrix. The mentioned matrices have the known forms as presented in [3].

A generalized force vector Λ has the form [3]:

$$\Lambda = O_B(v)\Omega + E_B\Omega^2 \tag{8}$$

where:

$$O_B = J_{TP}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \dot{\theta} & -\dot{\theta} & \dot{\theta} & -\dot{\theta} \\ -\dot{\phi} & \dot{\phi} & -\dot{\phi} & \dot{\phi} \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{9}$$

is the gyroscopic propeller matrix and J_{TP} is the total rotational moment of inertia around the propeller axis. The movement aerodynamic matrix has the form [3]:

$$E_B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ b & b & b & b \\ 0 & -b \cdot l & 0 & b \cdot l \\ -b \cdot l & 0 & b \cdot l & 0 \\ -d & d & -d & d \end{bmatrix} \tag{10}$$

where b ($N\ s^2$) and d ($N\ m\ s^2$) are thrust and drag factors [3] and l (m) is the distance between the center of the quad-rotor and the center of the propeller. Equation (11) defines the overall propellers' speed (rad s^{-1}) and the propellers' speed vector (rad s^{-1}) used in equation (8).

$$\omega = -\omega_1 + \omega_2 - \omega_3 + \omega_4 \tag{11}$$

$$\Omega = \begin{bmatrix} \omega_1 & \omega_2 & \omega_3 & \omega_4 \end{bmatrix}^T \tag{12}$$

Equations (1)-(12) take into account the entire quad-rotor non-linear model including the most influential effects.

2.3 Control Strategy

In this section we present a control strategy to stabilize of the quad-rotor. Fig. 5 shows the block diagram of the quad-rotor control system.

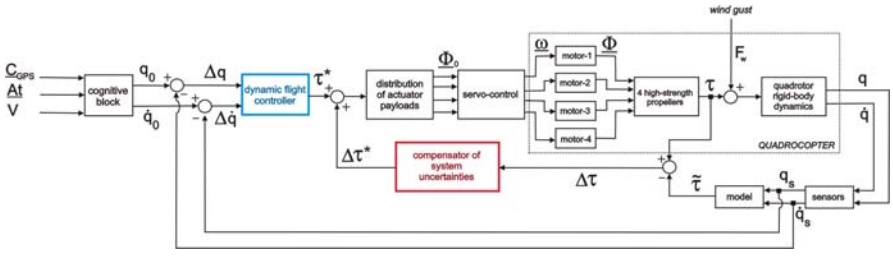


Figure 5
The block diagram of the quad-rotor control system

3 GPS Navigation of Quad-Rotor

The trajectory of the quad-rotor can be introduced by GPS coordinates (e.g. $\underline{P}_{GPS}(j)$) as shown in Fig. 6.

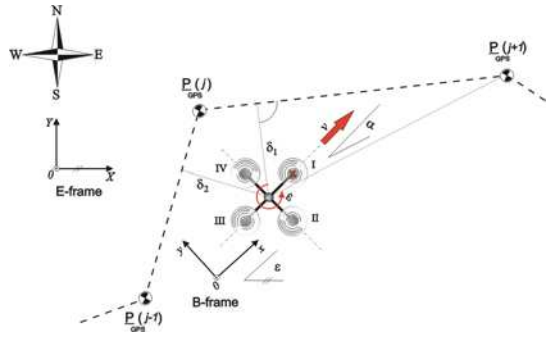


Figure 6
Quad-rotor localization and navigation with respect to the imposed GPS coordinates

The quad-rotor is requested to track the imposed trajectory between the particular points ($j = 1,...,n$) with satisfactory precision, keeping the desired attitude and height of flight. The quad-rotor checks for the current position (X and Y) by use of a GPS sensor and/or electronic compass. Also, the altitude is measured by a barometric sensor. An on-board microcontroller calculates the actual position deviation from the imposed trajectory given by successive GPS positions $\underline{P}_{GPS}(j)$. It localizes itself with respect to the nearest trajectory segment (by calculation of the distances δ_1 or δ_2). Using the gyroscope, the quad-rotor determines desired azimuth of flight α (Fig. 6) and keeps the desired direction of flight. The height of flight is also controlled to enable the performance of the imposed mission (task). One characteristic example of imposing quad-rotor trajectory by use of GPS coordinates is given in the next paragraph.

The corresponding Google Earth map is utilized to provide corresponding GPS coordinates of the quad-rotor trajectory as presented in Fig. 7.

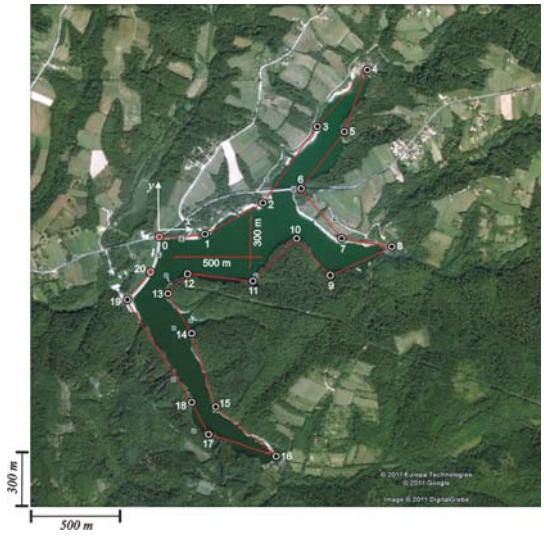


Figure 7

Google-Earth map of the Garaši lake used to define desired GPS trajectory of the quad-rotor aerial robot

GPS coordinates (longitude, latitude and altitude), defined in the map and given in the Fig. 8, are used to calculate quad-rotor trajectory in the E-frame.

	LONGITUDE			LATITUDE			ATTITUDE
point	degrees	minutes'	seconds"	degrees	minutes'	seconds"	feet
0	44	17	20.88	20	28	24.19	970
0'	44	17	20.88	20	28	24.19	970
1	44	17	21.44	20	28	36.04	963
2	44	17	27.11	20	28	50.67	964
3	44	17	40.87	20	29	4.07	970
4	44	17	51.45	20	29	17.09	968
5	44	17	39.87	20	29	11.14	987
6	44	17	29.61	20	29	0.2	980
7	44	17	20.63	20	29	10.35	970
8	44	17	19.05	20	29	23.15	991
9	44	17	14.02	20	29	7.53	1049
10	44	17	20.72	20	28	59.15	986
11	44	17	13.12	20	28	48.47	1045
12	44	17	14.54	20	28	31.86	983
13	44	17	11.05	20	28	26.98	987
14	44	17	3.83	20	28	32.98	970
15	44	16	50.71	20	28	39.13	1007
16	44	16	41.90	20	28	54.12	970
17	44	16	46.05	20	28	37.28	1036
18	44	16	53.58	20	28	31.58	1016
19	44	17	9.90	20	28	16.59	983
20	44	17	14.96	20	28	22.6	943
20'	44	17	14.95	20	28	22.6	943

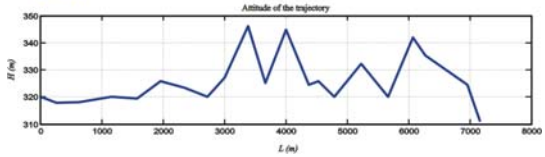


Figure 8

GPS coordinates acquired from the Google Earth map and used for the determination of the desired quad-rotor trajectory

Corresponding model of the trajectory given in E-frame is presented in Fig. 9.

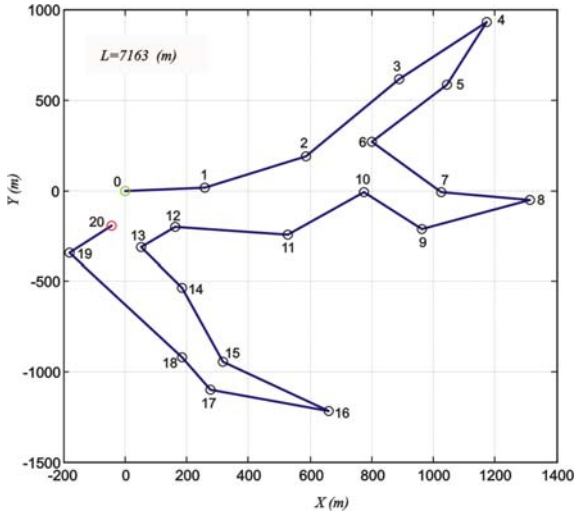


Figure 9
Multi-segment trajectory model of the quad-rotor determined in the E-frame

4 Simulation Example

Corresponding modeling and simulation software of rotorcraft aerial robots is developed in Matlab/Simulink. The open-loop simulation is performed and results are presented in the paper to verify model capabilities and system analysis. Aiming towards this goal, a quad-rotor flight is simulated by imposing corresponding particular propeller angular velocities presented in Fig. 10.

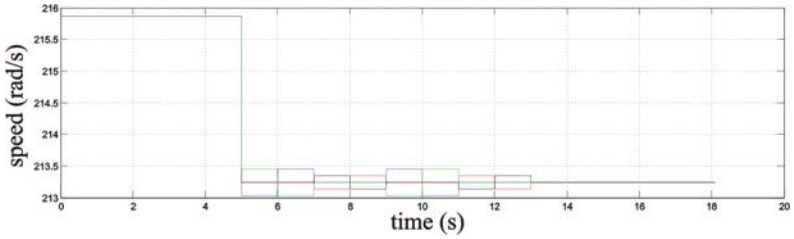


Figure 10
Example of the imposed propeller angular velocities that enable desired movements of quad-rotor

The imposed movements include several flight phases: (i) throttle movements in the vertical direction, (ii) counter-clockwise roll movements, (iii) tilt movement about the pitch axis, and (iv) hovering with a constant propeller speed. One sequence of the simulation test-flight is presented by the 3D-plot in Fig. 11.

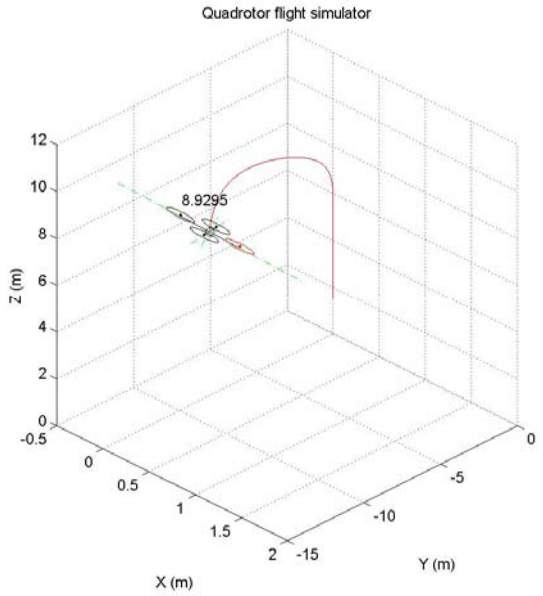


Figure 11
3D-plot of the quad-rotor simulation flight

Corresponding roll and pitch movements, due to the imposed variation of rotors' speeds, are presented in Fig. 12.

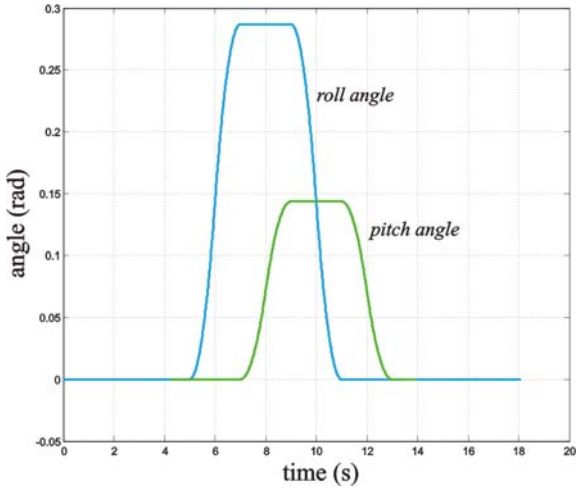


Figure 12
Roll and pitch movements due to the imposed propeller rotations given in Fig. 9

The quad-rotor trajectory projection in the X-Y plane is given in Fig. 13. Changing of the quad-rotor height during flight is given in Fig. 14.

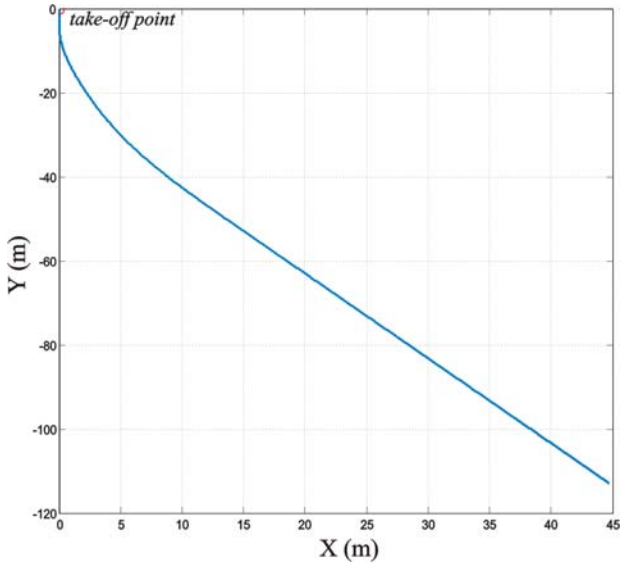


Figure 13
Quad-rotor movements in X- and Y-direction

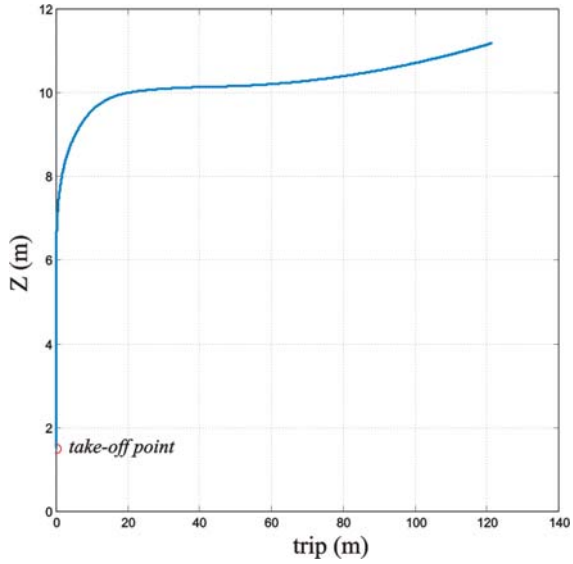


Figure 14
Quad-rotor movements in Z-direction

Conclusions

The paper considers the modeling and simulation of an autonomous quad-rotor microcopter in a virtual outdoor scenario. The main contribution of this paper focuses on the development of a flight simulator to provide an advanced R/D tool suitable for control design and model evaluation of a quad-rotor system to be used for control algorithm development and verification before working with real experimental systems. The main aspects of modeling of rotorcraft kinematics and rigid body dynamics, spatial system localization and navigation in virtual outdoor scenario are considered in the paper. Finally, several basic maneuvers are investigated and simulated in the paper to verify the simulation software capabilities and engineering capabilities.

Acknowledgement

This work was supported by the innovation project 'Research and Development of Ambientally Intelligent Service Robots', TR-35003, 2011-2014, funded by the Ministry of Science of the Republic Serbia and partially supported by the TÁMOP-4.2.2/08/1/2008-0008 program of the Hungarian National Development Agency.

References

[1] C. Lebres, V. Santos, N. M. Fonseca Ferreira and J. A. Tenreiro Machado: Application of Fractional Controllers for Quad Rotor, Nonlinear Science and Complexity, Part 6, DOI: 10.1007/978-90-481-9884-9_35, Springer, 2011, pp. 303-309

[2] J. Coelho, R. Neto, C. Lebres, V. Santos: Application of Fractional Algorithms in Control of a Quad Rotor Flight, Proceedings of the 2nd Conference on Nonlinear Scienceand Complexity, Porto, Portugal, July 28-31, 2008, pp. 1-12

[3] Tommaso Bresciani, Modelling, Identification and Control of a Quadrotor Helicopter, Department of Automatic Control, Lund University, ISSN 0280-5316, ISRN LUTFD2/TFRT/5823.SE, October 2008

[4] B. Siciliano and O. Khatib, Eds., Handbook of Robotics, Springer, ISBN: 978-3-540-23957-4, 2008, pp. 391-410

[5] Barnes W. and McCormick, W., Aerodynamics Aeronautics and Flight Mechanics. New York: Wiley, 1995

[6] Gordon Leishman, J., Principles of Helicopter Aerodynamics, Second Edition, Cambridge University Press, 1995

[7] Etkin, B. and Reid L. R., Dynamics of Flight- Stability and Control. John Wiley & Sons. New York, 1996

[8] Castillo, P. Dzul, A. Lozano, R. Stabilization of a Mini Rotorcraft Having Four Rotors, Control Systems Magazine, Vol. 25, No. 6, pp. 45-55, December 2005.Authors: Title

[9] Aircraft X650 Quad-rotor, http://www.infmetry.com/coolstuff/xaircraft-x650-quadcopterquadrotor/

[10] Aleksandar Rodic, Gyula Mester, "Modeling and Simulation of Quad-Rotor Dynamics and Spatial Navigation", Proceedings of the SISY 2011, 9th IEEE International Symposium on Intelligent Systems and Informatics, pp 23-28, ISBN: 978-1-4577-1973-8, Subotica, Serbia, September 8-10, 2011

[11] Ján Lábun, František Adamčík, Ján Pil'a, Ladislav Madarász: Effect of the Measured Pulses Count on the Methodical Error of the Air Radio Altimeter, in Acta Polytechnica Hungarica, Vol. 7, No. 1, 2010, pp. 41-49

Effect of Thermo-Mechanical Treatment on the Phase Composition and Resistance to Plastic Deformation of Chromium-Nickel Steel

L. V. Zaitseva
Centre for Technical Inspections Techdiagaz Affiliated Company
Ukrtransgaz National Joint-Stock Company
Naftogaz of Ukraine
56 Volynska st., 03151 Kyiv, Ukraine
E-mail: LZaitseva@tdg.kiev.ua

Boris I. Koval'chuk

National Technical University of Ukraine "Kyiv Polytechnical Institute"
37 Peremogy Prospect, UA-03056, Kyiv-56, Ukraine
E-mail: mef@users.ntu-kpi.kiev.ua

Abstract: The experimental investigation of the influence of time-dependent, thermal and mechanical factors on the kinetics of martensite transformation has been performed in the austenitic chromium-nickel steels. Additionally, the effects of martensite transformation on the mechanical properties are studied and the nature of strengthening is analyzed during low-temperature deformation in the linear and two-dimensional stress states. The equation is evaluated which describes the kinetics of martensitic transformation in the stress strain states. A variant of the deformation theory of plasticity for metastable materials under proportional loading is considered.

Keywords: metastable austenitic steels; plastic deformation; tensile; torsion; compression; biaxial tension; low temperatures; phase $(\gamma \to \alpha)$ *-transformations; plasticity theory*

1 Introduction

It is well known [1-3] that processes of phase transformations during plastic deformation in metastable austenitic steels depend essentially on temperature and force loads. The variation of either temperature or force load or their combined change can regulate the intensity of phase transformations and, hence, control the physical and mechanical properties of steels.

This paper presents the experimental results of the comprehensive structural and mechanical studies of 18-10 chromium-nickel austenitic steels. The influence of chemical composition, the mode of refrigeration and storage at low temperatures, the levels of prior plastic deformation on phase compositions and the mechanical properties were analyzed in the steels.

2 Materials, Treatment and Testing

Investigations were carried out using solid (∅ 6 mm) and thin-walled tubular (external diameter 26 mm, wall thickness 0.5 mm) samples. The samples were annealed in vacuum at 1350 K.

The chemical composition of the chromium-nickel steel Cr18Ni10Ti, according to the requirements provided by the GOST, is the following in wt. %: 0,07-0,12 C; 17,0-19,0 Cr; 9,0-11,0 Ni; 1,03 Mn; 0,67-0,41 Ti; 0,39-043 Si; 0,11 Mo; 0,06 V; 0,04 Cu; balance Fe.

The samples were cooled in liquid nitrogen or its vapors.

The phase compositions of steel after the different stages of deformation were determined by X-ray diffraction (XRD) analysis using a DRON-2.0 unit after unloading and reheating to room temperature.

3 Results and Discussion

The intensity of phase transformations under mechanical loading depends significantly on the chemical composition of the steel.

The analysis of the mechanical properties and kinetics of phase transformations of 18-10 type steels under uniaxial tension has demonstrated that a variation of chemical composition within the specified standard limits influences the stability of the structure and, respectively, the steels' mechanical behaviors at low temperatures. A decrease in the test temperature to 77 K results in various changes of strain hardening of the materials. The more unstable steels show their high strain hardening behaviors.

For the purpose of studying the influence of cooling on martensite formation, the samples were tested in the non-deformed and cold deformed (at room temperature) states [4]. One batch of samples was subjected to different numbers of the thermo-cycle: cooling to 77 K, holding for 2-3 min and reheating to 293 K. The second batch was kept in liquid nitrogen for a long time (up to 250 hours). The results of XRD have shown that the volume fraction of martensite f_M increases with an increase in both the number of cycles and the holding time in the liquid nitrogen.

However, the dominant factor contributing to the formation of martensite in material is cooling compared to holding in the refrigerant.

Preliminary deformation by tension in 20% and the long-time recovery of samples at room temperature (for 2-3 months) initiates the different amounts of martensite formation during their subsequent thermal cycling. For instance, the pre-deformed and non-deformed samples have the volume fractions of martensite about 13 and 9%, respectively, after 7 cycles.

The holding of materials under stress at low temperatures results in the formation of an additional amount of martensite. In this case, under holding for 2 hours in the elastic range, the amount of martensite is negligibly small (f_M ~4%), and under holding in the plastic condition, the amount of martensite is high (f_M ~11%).

The major factor affecting the phase composition of metastable austenitic steels is low-temperature plastic deformation. However, the stress state under which deformation is performed is important.

Studies on the influence of stress states on the kinetics of phase transformations during tensile, torsion and compression tests were performed at 77 K. Biaxial tensile tests of thin-walled tubular samples were carried out at the different ratios of principal stresses at temperature 173 K. It was established that the intensity formation of α -phase under tension tests is essentially higher than during torsion and compression experiments.

A study of the structural changes in the material under biaxial tension has indicated that the amount of martensite decreases at the same value of deformation compared to the uniaxial tests, though the resistance of the steel expressed by stress intensity σ_i increases essentially. This behavior is connected with the formation of a stronger martensite under biaxial tension in comparison to uniaxial.

The analysis of the obtained results has shown that the volume content of martensite f_M formed under deformation of austenitic steels depends on the stress deviator (Lode-Nadai parameter $\mu_\sigma = (2\sigma_2 - \sigma_1 - \sigma_3)/(\sigma_1 - \sigma_3)$) and on the rigidity of a stress state, which is characterized by the parameter $k_\sigma = \sigma_0 / \sigma_i$, where $\sigma_0 = (1/3)(\sigma_1 + \sigma_2 + \sigma_3)$ is the mean stress. With an increase in the stress-state rigidity and a decrease in the parameter μ_σ, the process of martensite formation becomes more intensive.

The relation $f_l\ (\varepsilon_i, \mu_\sigma, k_\sigma)$ is complex in nature. Based on the analysis of the surfaces $f_M(\mu_\sigma, k_\sigma)$ which correspond to the different constant values ε_i, a function describing the kinetics of martensite formation under loading was evaluated along arbitrary ray trajectories in three-dimensional stress space:

$$f_M(\varepsilon_i, \mu_\sigma, k_\sigma) = f_0(\varepsilon_i) - c_0(\varepsilon_i)\mu_\sigma + g(\varepsilon_i)(1 + \mu_\sigma)thk_\sigma \qquad (1)$$

The functions $f_0(\varepsilon_i)$, $c_0(\varepsilon_i)$ and $g(\varepsilon_i)$ are determined from tensile, compression, and biaxial tension or torsion tests, respectively.

The surfaces $f_M(\mu_\sigma, k_\sigma)$, constructed according to Eq. (l) for some ε_i=const, are shown in Fig. 1. As is seen in Fig. 1, Eq. (1) adequately describes the effect of a stress state on martensitic transformations in 18-10 steels during plastic deformation under combined stress conditions.

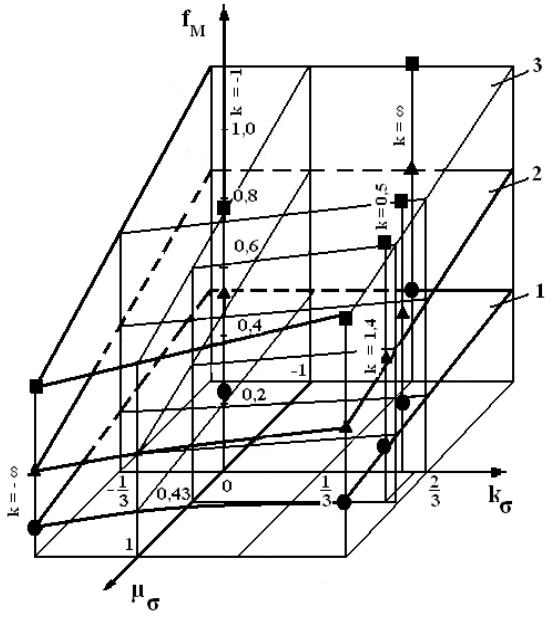


Figure 1
Martensite content f_M vs μ_σ, and κ_σ for fixed strain intensities: $\varepsilon_i = 5\%$ (1, •), $\varepsilon_i = 10\%$ (2, ▲) and $\varepsilon_i = 15\%$ (3, ■). (Surfaces 1-3 are constructed by Eq. (1); points correspond to experimental data.)

Investigations of the loading prehistory on the kinetics of martensite transformations in austenitic steel under repeated plastic deformation, at different preliminary and repeated loadings, temperatures and stress states were performed in [5].

The comprehensive information about the influence of the preliminary thermo-mechanical treatments on plastic deformation and creep behaviours of metals is available in ref. [6-9].

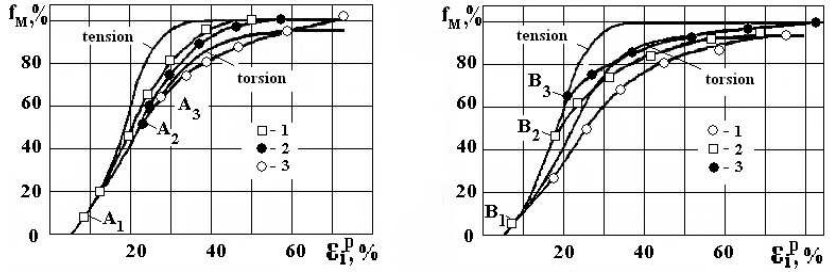
The mechanical tests were performed by six (I—VI) programs according to which the prior and subsequent loadings were carried out by tension and torsion in the different combinations at the temperatures 77 K and 293 K (Table 1). The amounts of plastic pre-strain were reached about 17-30%.

Table 1
Mechanical Test Programs

Program	Prior loading			Repeated loading	
	Form	ε_{ipr}^{p} , %	T_1,K	Form	T_2,K
I	Torsion	8,7 22,9 28,2	77	Tension	77
II	Tension	6,7 17,5 20,1	77	Torsion	77
III	Tension	11,3 18,3 23,9	293	Tension	77
IV	Torsion	19,4 30,4 50,9	293	Tension	77
V	Tension	9,5 17,7	77	Tension	293
VI	Torsion	10,7 7,5	77	Tension	293

The steel investigated was comparatively stable, since its deformation at room temperature did not initiate $(\gamma \to \alpha)$-transformation up to the failure.

The analysis of the test results under tension after torsion and vice versa (programs I and II) at the same temperature 77 K of preliminary and repeated loadings indicates that a change of a stress state at a low temperature slows down the martensite formation in the steel (Fig. 2).



a *b*

Figure 2
Kinetics of martensitic transformation in chromium-nickel steel at 77 K:
a) program I,1) ε_{ipr}^{p} = 8,7%(A_1); 2) ε_{ipr}^{p} = 22,9%(A_2); 3) ε_{ipr}^{p} = 28,2% (A_3);
b) program II, 1) ε_{ipr}^{p} = 6,7%(B_1); 2) ε_{ipr}^{p} = 17,5%(B_2);) ε_{ipr}^{p} = 20,1% (B_3)

Apparently this result can be explained by the specific nature of martensite transformation. Since $(\gamma \to \alpha)$-transformation is accomplished by shear during plastic deformation, the preferred orientation of a high density of martensite plates in one direction forms due to the slip lines in one direction in the structure. Similar structures have also been observed with deformation of specimens of steel Cr18Ni9 in the temperature range 100-150 K. The development of transverse slip in these grains is difficult because the transverse slip must intersect a slip band of dense bundles of shear lines and martensite plates which exhibit lower ductility and higher yield stresses compared with austenite. Consequently, the formation of a network of directed plates of martensite during prior deformation limits slip in retained austenite under other forms of stressed states. As a result, there is inhibition of martensite transformation at the beginning of deformation, which localizes intensive plastic deformation along slip planes.

The tests at the different stressed states and temperatures (Programs III and IV) indicate that prior loading by both tension and torsion to strain of about 30% at room temperature has little effect on the kinetics of martensite transformation during subsequent low-temperature tension.

At the same time, the prior low-temperature deformation by tension and torsion (Programs V and VI) initiates martensitic transformation during subsequent loading in tension under room temperature conditions (Fig. 3).

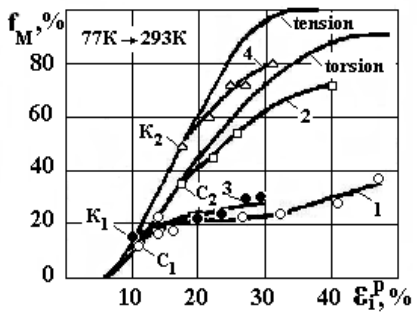


Figure 3
Effect of prior low-temperature deformation for steel by programs
V: 1) ε_{ipr}^{p} =10,7% (C_1), 2) ε_{ipr}^{p} =17,5% (C_2), and VI: 3) ε_{ipr}^{p} =9,5% (K_1), 4) ε_{ipr}^{p} =17,7%(K_2)

The prior deformation at the low-temperature condition (77 K) initiates the formation of α - phase with repeated "warm" (293 K) loading. The tension of 10% at room temperature after prior low-temperature tension to ε_{ipr}^{p} =9.5% increases the volume content of martensite to 9%, but after tension up to ε_{ipr}^{p} =17.7% the amount of martensite is about 2.6%.

Similar results were confirmed by tests of the thin-walled tubular samples under biaxial tension. A preliminary axial tension up to 10% was carried out at

temperatures of 77, 123, and 293 K. Repeated loadings were performed at the different ratios of principal stresses and temperatures 123 K and 293 K. In these experiences, strain curves of the steel are recorded in longitudinal and tangential directions that allow measuring the character of the strain hardening and the evolution of the yield surface of steel in the process of plastic deformation.

The results have demonstrated that during plastic deformation of metastable austenitic steel at different temperatures, the direction of maximum hardening in space of stresses may be different from the direction of the preloading. This effect is especially pronounced when the temperatures of preliminary and repeated loading are different and the material is in a metastable state at one of the indicated temperatures. This leads to the complicated transformation of the yield surface that cannot be described by the well-known models of hardening that are usually used in flow theory.

The analysis of the mechanical behavior of metastable steel under plane stress conditions at proportional loading allows for the checking of the basic hypotheses of the strain theory of plasticity. In particular, it was found that the strain curves $\sigma_i = f(\varepsilon_i)$ of steels, obtained during tests of tubular specimens, are non invariant to the stress state (Fig. 4) under plastic deformation. A significant residual change of the steel volume was observed. However, the condition of a similarity of stress and strain deviators is implemented.

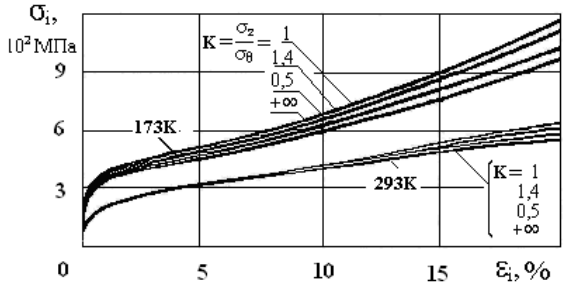


Figure 4

Deformation curves for Cr18Ni10Ti steels at different temperatures under biaxial tension

Based on the formulation of new principles that take into account the influence of the phase transformations on mechanical properties, the constitutive equations of the strain theory of plasticity were deduced. It was applied to metastable materials with deformational phase $(\gamma \to \alpha)$ - transformations in the following form [10, 11]

$$\varepsilon_{ij} = \frac{3}{2}\frac{\varepsilon_i}{\sigma_i(\varepsilon_i, \mu_\sigma, k_\sigma)}(\sigma_{ij} - \delta_{ij}\sigma_0) + \delta_{ij}\left[\frac{\sigma_0}{3K_0} + \varepsilon_f\ (\varepsilon_i, \mu_\sigma, k_\sigma)\right], \tag{2}$$

where K_0 is the volumetric modulus of elasticity, ε_f is the mean residual strain, and δ_{ij} is the Kronecker deltas.

The results of comprehensive mechanical and structural studies of the 18-10 steels made it possible to specify the functions of hardening $\sigma_i(\varepsilon_i, \mu_\sigma, k_\sigma)$ and the residual change of the volume $\varepsilon_i(\varepsilon_i, \mu_\sigma, k_\sigma)$, and they are presented as:

$$\sigma_i(\varepsilon_i, \mu_\sigma, k_\sigma) = \sigma_p(\varepsilon_i)\sqrt{\frac{\mu_\sigma{}^2 + 3}{(\mu_\sigma + 1)^2 - 2\chi(\varepsilon_i)(\mu_\sigma + 1) + 4}} \times$$
$$\times\left[1 + \psi(\varepsilon_i, \mu_\sigma, k_\sigma)\right] - \sigma_a(\varepsilon_i)\psi(\varepsilon_i, \mu_\sigma, k_\sigma), \quad (3)$$

where

$$\psi(\varepsilon_i, \mu_\sigma, k_\sigma) = \frac{\eta(\varepsilon_i)(1 + \mu_\sigma)}{\varepsilon_f^0(\varepsilon_i) - \lambda(\varepsilon_i)\mu_\sigma} \operatorname{th} k_\sigma \quad (4)$$

$$\varepsilon_f(\varepsilon_i, \mu_\sigma, k_\sigma) = \varepsilon_f^0(\varepsilon_i) - \lambda(\varepsilon_i)\mu_\sigma + \eta(\varepsilon_i)(1 + \mu_\sigma) \operatorname{th} k_\sigma \quad (5)$$

The functions $\sigma_p(\varepsilon_i)$, $\sigma_a(\varepsilon_i)$, $\chi(\varepsilon_i)$, $\eta(\varepsilon_i)$, $\varepsilon_f^0(\varepsilon_i)$, $\lambda(\varepsilon_i)$ are calculated from the results of the three base experiments, for example, uniaxial tension, uniaxial compression and biaxial tension at σ_2 / σ_1=0,5 or torsion.

The calculation results obtained by Eqs. (2), (3), (4) and (5) are in a good agreement with experimental data for proportional loading of 18-10 steels in a metastable state.

Conclusions

The thermal cycling of the metastable stainless steels initiates phase transformations leading to some increase in their strength. The external load and the thermal stresses contribute to these processes.

The type of stress state has a significant effect on the kinetics of martensitic transformations and the stress-hardening of steels. There is no unique dependence between the amount of martensite formed and the resistance of steel, since the strength of martensite depends on the type of stress state.

Preliminary plastic deformation of steels at room temperature has no significant influence on their mechanical properties at low temperatures. At the same time, a preliminary low-temperature deformation initiates phase transitions and improves the mechanical properties of steels at room temperature. The discrepancy between the forms of stress states in the preliminary and subsequent deformation slows down the phase transformations in steel Cr18Ni10Ti.

The mathematical model describing the kinetics of martensitic transformations in steels during plastic deformation under complex stress states as well as the plastic deformation of unstable structure under proportional loading are considered.

References

[1] M. Smagaa, F. Walther, A, D. Eiflera.: Deformation-Induced Martensitic Transformation in Metastable Austenitic Steels, *Materials Science and Engineering: A* Vol. 483-484, pp. 394-397, 15 June 2008

[2] Noriyuki Tsuchida, Kenzo Fukaura, Yo Tomota, Atsushi Moriai, Hiroshi Suzuki: Tensile Deformation Behaviors of Metastable Austenitic Stainless Steels Studied by Neutron Diffraction, *J. Materials Science Foru,* Vol. 652, pp. 233-237, 2010

[3] Tetsu Narutani: Effect of Deformation-Induced Martensitic Transformation on the Plastic Behavior of Metastable Austenitic Stainless Steel, *Materials Transactions*, JIM, Vol. 30, No. 1, pp. 33-45, 1989

[4] Lebedev A. A., Koval'chuk B. I., Zaitseva L. V.: Effect of Thermomechanical Conditions on the Kinetics of Phase Transformations in Austenitic Metastable Steel, *J. Strength of Materials*, Vol. **27**, No. **3**: pp. 124-128, 1995

[5] Koval'chuk B. I., Zaitseva L. V., Lebedev A. A., Kosarchuk V. V.: Effect of Loading Prehistory on Phase Transformations in Chromium-Nickel Steel with Plastic Deformation, *J. Strength of Materials*, Vol. **10:** pp. 1101-1105, 1991

[6] Rusynko A. Influence of Preliminary Mechanical and Thermal Treatment on the Steady-State Creep of Metals, *J. Materials Science* **40**: 223-231, 2004

[7] Ruszinko, E. The Influence of Preliminary Mechanical-thermal Treatment on the Plastic and Creep Deformation of Turbine Disks, *J. Meccanica* **44**: 13-25, 2009

[8] Rusinko, A., Rusinko, K. Synthetic Theory of Irreversible Deformation in the Context of Fundamental Bases of Plasticity, *Int. J. Mech. Mater.* **41**: 106-120, 2009

[9] Rusinko A. Non-Classical Problems of Irreversible Deformation in Terms of the Synthetic Theory, Acta Polytechnica Hungarica Vol. 7, No. 3, pp. 25-62, 2010

[10] Koval'chuk B. I., Zaitseva L. V.: Derivation of the Stress-Strain Equations for Metastable Materials with Martensite Transformations under Proportional Loading, *J. Strength of Materials*, Vol. **31**, No **3**, pp. 232-240, 1999

[11] Lebedev A. A., Koval'chuk B. I., Zaitseva L. V.: Deformation of Structure-Unstable Medias under Combined Stress Conditions and his Mathematical Design, *Vestnik of the KPI Nat.Techn.Univ*., ser. Engineering, VIPOL Publ., **52**: pp. 78-88, 2008

Towards a New Approach in Available Bandwidth Measures on Mobile Ad Hoc Networks

Redouane Belbachir, Zoulikha M. Mekkakia, Ali Kies

Department of Data Processing, University of Sciences and Technology of Oran
USTO-MB, BP 1505 El M'Naouar, Oran, Algeria
belbachir_red@yahoo.fr, mekkakia@univ-usto.dz, kies_ali@yahoo.fr

Abstract: Given the development of multimedia applications with intensive consumption of network resources by the traffic generated, and the emergent use of mobile ad hoc networks, the guarantee of the quality of service (QoS) has become essential. The available bandwidth is the crucial resource, particularly in mobile ad hoc networks, since it is very limited; this is why it must be optimally managed and accurately estimated. Several projects have been conducted to provide new techniques for measuring the available bandwidth in mobile ad hoc networks. But mobility, which is the basic criterion of this type of network, is neglected and isn't processed by any measurement technique proposed. In this paper we disclose the importance of the taking into account this criterion and we will unveil a novel measurement technique ABE_MM (Available Bandwidth Estimation with Mobility Management) which is an extension of other robust existing measurement methods and which treats mobility in a specific manner.

Keywords: ad hoc networks; mobility; estimation; bandwidth; ABE_MM

1 Introduction

In Wireless networks, the IEEE 802.11 norm is most frequently used. It provides ad hoc configuration that operates with fundamental mechanism to access on medium, which is called *Distributed Coordination Function* DCF mode [1]. It allows a more flexible communication model than others networks since the user is not limited to a fixed physical location. Mobile ad hoc networks allow autonomy and independence of any fixed infrastructures or coordinating points. Considering topology changes due to the mobility of hosts, these last must self organize to transfer data packets or any information with mobility and wireless physical characteristics management.

Observing the use of multi-hop communication because of limiting the communication range and the traffic diversity transiting the mobile ad hoc networks, the quality of service (QoS) became the subject in art. However, the term QoS is so vague that it has been extensively studied and QoS solutions are increasingly being proposed. The QoS solutions take different concepts to ensure the meeting of criteria required by applications (like bandwidth, delay transmission and others).

Most QoS solutions proposed are interested in particular resources in the network. These solutions are usually embedded with measurement methods of these resources which are the object of their interest. Available bandwidth is a fundamental one of these resources, especially in the 802.11 ad hoc networks, since it is limited and shared by all neighboring nodes; its optimal management is indispensable.

In this paper we focus on the optimal management of bandwidth in 802.11-base ad hoc networks to provide better QoS. This requires prior knowledge of the availability of this resource, where an accurate estimate of available bandwidth is essential. The estimation of available bandwidth is such a delicate operation because of the dynamic topologies of ad hoc networks, where links between nodes may disparate at any time which has the effect of very frequent disconnection paths.

The available bandwidth estimation techniques in mobile ad hoc networks have experienced real progress in accuracy terms. After a thorough study of these techniques and QoS solutions in ad hoc networks, we noticed that the mobility problem is still persistent. So through this paper we present a new approach, ABE_MM, which is an extension of the ABE [2] technique (which made the evidence in terms of accuracy, which is why we chose it) to manage mobility specifically in bandwidth measurement. The particularity of this approach is that it can be combined with any other measurement technique of bandwidth on dynamic topologies. A collaboration is used in this paper between the MAC, physical and network layers to respond to our approach. Knowing different methods could be used to respond to our approach.

This paper is organized as follows: In Section 2, Existing Work on Bandwidth in Mobile Ad hoc Networks, we provide an overview on techniques to measure bandwidth and look especially at the technical ABE. In Section 3, we will give motivation points for our approach to managing mobility in bandwidth measurement. In Section 4, we explain the concept of our approach, the protocol extensions and the changes necessary to achieve the ABE_MM technique, and finally in Section 5, in order to evaluate our approach, a comparison is made between ABE_MM, ABE techniques and AODV through a network simulator.

2 Existing Work on Bandwidth in Mobile Ad hoc Networks

There have been significant changes in the types of traffic types transiting networks, led by the development of different applications. The request for the transmission of multimedia applications in wireless networks has increased significantly. Consequently, the quality of service management has created strong interest. Note that the issues raised in the specific context of wireless networks are different and more complex than those encountered in wired networks.

The majority of QoS solutions are loaded with techniques for measuring the delay, jitter, available bandwidth or other criteria. Consequently, computing the available bandwidth with as much precision as possible must also consider the existing traffic and topology. The available bandwidth of one link (between two neighbor nodes) can be defined as the maximum throughput that can be transmitted without influence on any existing flow in the network. Contrary to this is capacity, which is the maximum throughput that flow can achieve in one idle link on an idle network. The most bandwidth reservation solutions designed to 802.11 based ad hoc networks use passives estimations for available bandwidth. Passives estimations are based on monitoring of the networks activities which corresponds to the radio channel activities (transmissions, receptions, and idle periods) or data transmission quantities. One of these passive estimation techniques used is the QoS-aware protocol, proposed by Lei Chen in [3].

The QoS-aware routing protocol is dedicated to the admission control flow depending on throughput requested by applications and available bandwidth on the network to meet the requirements of QoS.

In QoS-aware, there are two methods used to estimates the residual available bandwidth. In the first method, each node estimates its residual bandwidth by monitoring the free times on radio channel every Δ time interval. Monitoring is done at the MAC layer, which detects that the channel is free if three conditions are met:

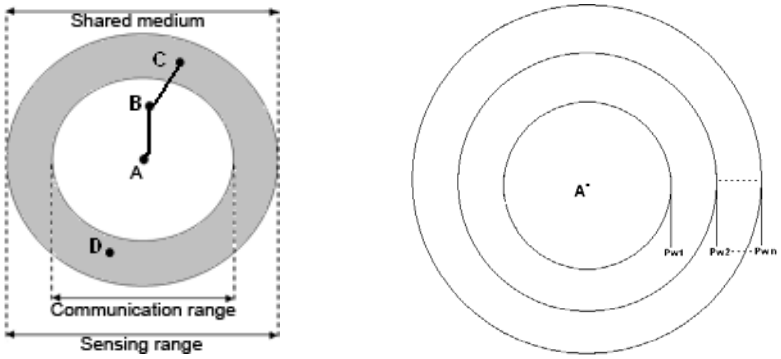
- The NAV's value is less than the current time;
- The receive state is idle;
- The send state is idle.

Then the residual available bandwidth of a node is computed from multiplying the rate of free times on Δ interval by the capacity.

Observing that the bandwidth is shared among neighboring nodes, in the second method of available bandwidth measurement in QoS-aware routing protocol, each node broadcasts its current bandwidth consumption value on "Hello" standard message to its one-hop neighbor, and each node estimates its residual available bandwidth from its consumption and the information received from neighbors nodes.

Always keeping the same idea, to transmit the value of bandwidth through "Hello" standard message, RENESSE and his team use in QoS-AODV [4] the BWER (Bandwidth Efficiency Ratio) measurement method, which broadcasts to the one-hop neighborhood by "Hello" messages the ratio value between the number of transmitted and received packets. And the available bandwidth is computed as the minimum of available bandwidth over a closed single-hop neighborhood.

It is known that in wireless networks the power of a signal emitted by a node is weakened along the path mainly according to the traveled distance (particularly on free space). The power of this signal at reception must exceed a certain threshold in order for it to be decoded. The communication zone is the area in which the receiver nodes can decode the signal. But the sensing zone is where the signal strength is below the threshold, and where the nodes detect the signals (then medium activity) and cannot decode it. This means that the medium (Figure 1(a), the medium (then the bandwidth) is also shared between the node A and D) is not shared only on the communication area but also on the sensing area; and solutions like BWER, of which the QoS-aware protocol is one, are not able to take into consideration this phenomenon in bandwidth measurement (because the node A cannot exchange information with node C directly, as depicted on Figure 1(a)).



(a) Sensing and Communication ranges of node A

(b) the communication areas changing according to the signals strength

Figure 1
Sensing and Communication ranges of node A

In order to solve the problem of estimating available bandwidth with taking into consideration the communication and sensing areas, firstly the authors of [5] present a modelling in a theoretical study of the interferences on the carrier sensing zones, and then in [6], they propose the BRuIT protocol which approximates this zone by using the neighborhood, where each node broadcasts to all its neighbors (on communication zone) the bandwidth rate which it uses and also information from all its one-hop neighbors, propagating this information at a two-hop distance. (The information are exchanged by using the Hello packets).

But the solution proposed with BRuIt fails in some scenarios, as depicted in Figure 1(a), where the node D, which is inside the carrier sensing of node A, can't be reached. The authors of CACP-power [7] propose using a physical layer to increase the transmission power signals when neighboring nodes exchange the information of bandwidth in order to reach the nodes which are inside sensing range and outside the communication rage (Increasing the transmission power of signals will increase the communication range Figure 1(b)).

Note that all these solutions take the hypothesis that the available bandwidth on the end-to-end path is defined as a minimal residual available bandwidth among the hosts in that path, as in [10] (The available bandwidth of one link is defined as the minimum between the two residual available bandwidths of the two nodes forming this link). Then such techniques are interested only by a residual available bandwidth of nodes, and constitute a class that is named "node by node" measurement techniques. The other class includes techniques oriented toward the estimation of available bandwidth especially at the links, as in ABE [2] and QOLSR [8]; this class is named "link by link" measurement techniques.

With the goal of moving towards the standardization of available bandwidth measurement techniques, Sarr and his team present in [2] the ABE technique. ABE treats QoS-based bandwidth in ad hoc networks through admission control of new flows in order to avoid any degradation of the existing traffic on the network, and so it requires the most accurate estimate possible. In [2], considering that the random changing of the topology on mobile ad hoc networks is the source of typical scenarios, the authors showed in first time that the hypothesis of "node by node" techniques is not always true because there are some scenarios where the available bandwidth of one link is less than the minimum residual available bandwidth of the two nodes forming this link. These scenarios are very common because of the random synchronization of the idle times between the sender and receiver nodes, for example in Figure 1(a), when the node (A) is in position to send data to the node (B), but this latter at this time is occupied; therefore, despite the availability of bandwidth at the two nodes, the traffic cannot be forwarded because of the non-synchronization of availabilities between the transmitter and receiver.

To address this problem, the authors have used probabilistic methods to define the available bandwidth on the link. For example on the (s,r) link, each node (s and r) monitors the medium at the MAC layer, as has been described with the first method of estimation in QoS-aware. Every Δ time (set to 1 second) the ratios of the idle times of each of the two nodes are computed, and the available bandwidth of this link is calculated as follows[1].

$$b_{(s,r)} = (t_s \cdot t_r).C \quad (1)$$

[1] The proof of formula (1) and collision probability calculating, are given in [2].

- $b_{(s,r)}$ is the available bandwidth of the link (s,r).
- t_s (resp. t_r) is the ratios of idle times on MAC layer of node s (resp. r).
- C is the capacity of the medium.

Collision rates are high in the mobile ad hoc networks; ABE has taken into consideration this problem by adding the probability of nonexistence of collision (1-*P*), where *P* is the probability of collision[1]. IEEE 802.11 defines the *backoff* system in order to reduce collision phenomenon. ABE also took into consideration the proportion of bandwidth "*K*" consumed by the *backoff* system, through the average of *backoff* time. So the formula (1) became:

$$ABE_{(s,r)} = (1-K)\cdot(1-P)\cdot b_{(s,r)} \quad (2)$$

- $ABE_{(s,r)}$ is the finale available bandwidth of the link (s,r).

ABE is integrated into the AODV [9] protocol; the throughput desired by the applications will be added in the RREQ packet. During the broadcast of the RREQ from a source node, the admission control will be executed by each node receiving this packet using the equation (2). This mechanism will allow for the movement of traffic in the network without being perturbed by others.

We note that all these techniques use an exchange of "Hello" standard messages periodically (for ABE, it's every Δ seconds) for mobility management. Because the protocol design of ABE is an extension of the AODV protocol, the detection of a link expiry is made according to the technique used in AODV. ABE also uses this exchange of "Hello" packets to exchange the bandwidth information between neighboring nodes; i.e. the Node *r* will need the t_s value to run the equation (2).

3 Motivations

Passive techniques are best suited for measuring available bandwidth in IEEE 802.11-based Ad Hoc networks. But reliability in accuracy terms is still not reached. And for reaching this reliability, we must consider all the essential criteria of the environment studied. This is not the case for the techniques presented previously. ABE proved the evidence against other technical in accuracy by taking into account a random synchronization mechanism. But the main element, the mobility that characterizes particularly the ad hoc mobile networks, is not treated in a specific manner whatsoever by ABE or any other measurement technique. Increasing the rate of packet "Hello" exchange is the solution most used.

Let us consider that there is a link (s,r) in the 802.11 ad hoc network with a high level of mobility. This high mobility leads to instability of the (s, r) link because of a distance changing between the two nodes s and r. Suppose there is no traffic on the network and the distance between the two nodes is larger than the distance

of communication. At the time T the distance between the two nodes (s and r) is enough for a communication (each of them is in the communication area of the other).

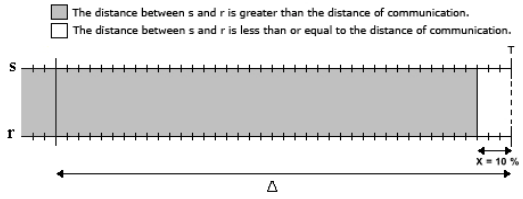


Figure 2
Link status during a Δ measurement period

Figure 2 shows the (s,r) link state during the measurement period Δ. We remark that this link exists only on X=10% of Δ measurement (X value represents the percentage of the existence of the link on the Δ period). At the T instant, the ABE technique can be executed because of exchanging possibility of available residual bandwidth on "Hello" packets.

We intend to compute the available bandwidth on the link (s,r) (the scenario depicted on Figure 2) with the ABE technique at time T with several X values, by using the NS-2 simulator with 2 Mb/s of medium capacity, which corresponds to 1.6 Mb/s application layer achievable throughput. The ABE estimations results are showed in graph of Figure 3.

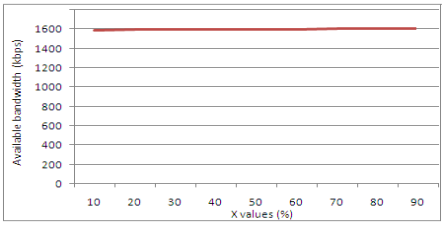


Figure 3
Available bandwidth estimation results with ABE Technique

Following the estimation results of the technical ABE (where Δ = 1), we understand that 1.6 Mbps of bandwidth was available during the period [(T-1)...T], regardless the existence rate (X) of the link (s, r) during this period. So up to 1.6 Mb of traffic could be transmitted during that period even if the link (s, r) exists only for X=10% of the period; but this is false, because 10% of Δ is 0.1 second and 1.6 Mb of traffic cannot be transmitted in 0.1 seconds with 1.6 Mbps of capacity. So this error on estimation is due to bad mobility management, and we will see that this error can affect the traffic circulation in mobile ad hoc networks with a penalization of certain applications, particularly in the admission control of QoS flows.

4 The Mobility in Available Bandwidth Measurement

Based on the physical quality of a link in wireless networks and considering how the IEEE 802.11 protocol operates, we can point out a few characteristics that we consider in our mobility management:

- Two nodes are separated by a distance *d*. A dynamic topology means that the distance *d* is variable.
- Two nodes can communicate if they are separated at most by a distance C_d (communication area).
- A node sends the signals with SSS strength, and these signals are received with RSS strength.

In this section, we will examine mobility by using these three points, and we describe how we can effectively take these phenomena into account.

Firstly, to solve the accuracy problem in available bandwidth estimation in the scenario of the previous section (Figure 2), we add the mobility criterion "*M*" in formula (2). Now, we consider this mobility criterion the rate of the link existence during the last measurement period (the value of X in Figure 2). Hence, we obtain the new formula with the criterion of mobility:

$$ABE\_MM_{(s,r)} = (1-K)\cdot(1-P)\cdot b_{(s,r)}\cdot M \quad (3)$$

where $ABE\_MM_{(s,r)}$ is the available bandwidth that we use in our measurement approach ABE_MM, but the value of "*M*" will be different, and in what follows we will see the new value of "*M*" and why.

4.1 Link Expiration in Bandwidth Measurement

Figure 4 shows a link (s, r) for two consecutive measurement periods (Δ), with two ((a) and (b)) identical activities scenarios (the same bandwidth consumption), but in different states: stable (a) and unstable (b), which we can observe at t_j instant on (b).

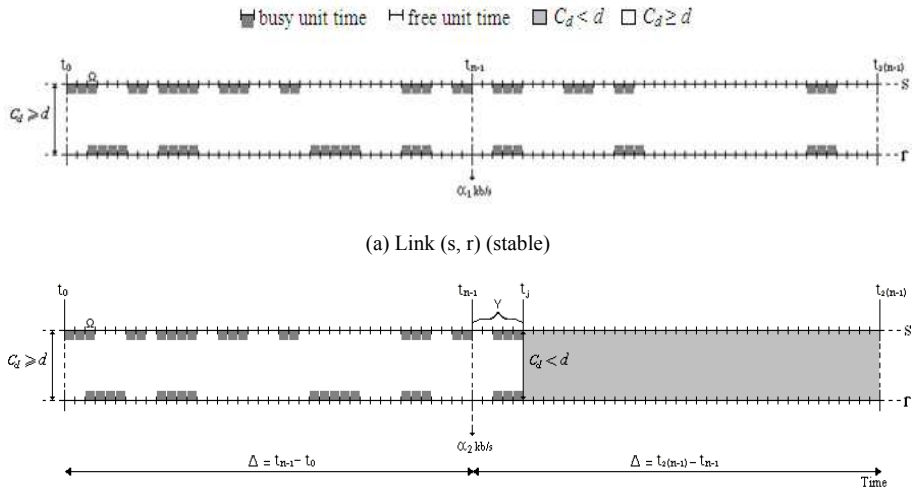
We say that a link is stable if the distance (*d*) between the two nodes of this link is less than or equal to the distance of communication C_d at any time on Δ. We note:

$\forall\, t_k / k \in [0...n],\ d \leq C_d$ (where t_k is the time unit of calculus in Δ period, and $n \in$

N*); this is the link which we find in wireless networks with stable topologies. Note that the stable links are the only scenarios discussed in the measurement formulas of bandwidth measurement.

We say that a link is dynamic or unstable, if there are some moments when the distance d between the two nodes of this link becomes larger than communication distance C_d, and we note: $\exists\, t_k / k \in [0..n],\ d > C_d$.

The absence of mobility criterion in the classical "link by link" measurement techniques such as ABE has negative consequences, particularly in the admission control process. It is clear that the available bandwidth α_1 (Figure 4 (a) at t_{n-1}) calculated by the ABE technique or even the formula (3) (M value as it is) will have the same value as α_2 (Figure 4(b) at t_{n-1}). So if ($\alpha_1=\alpha_2$), regardless the throughput admitted to reservation at t_{n-1} of Figure 4 (a) whichever α_1, also can be admitted, whichever α_2 in second scenario (Figure 4 (b) at t_{n-1}), while the (s, r) link in this second scenario will be available just in (t_j-t_{n-1}) time (during (Y) fragment). Hence the available bandwidth to be used at t_{n-1} depends also on the probability of future existence of the link; therefore, for the accurate value of α_2, we must take into account (Y) fragment, and the mobility criterion "M" is newly defined as the probability of the link existence in the next Δ measurement period time.



(a) Link (s, r) (stable)

(b) Link (s, r) (instable)

Figure 4
Expiration link on available bandwidth estimation

$$M = (t_j - t_{n-1})/\Delta \tag{4}$$

- t_j is the moment where the link disparate ($d > C_d$).

Now the problem that arises is to know at the instant t_{n-1} or before the value of t_j. In this paper we propose the calculation through the prediction of the link expiration, in order to predict the t_j value.

4.2 Predicting Link Expiration

Several studies have been conducted in the field of the prediction of nodes movement. In [11] the authors present a summary of mobility prediction methods in mobile ad hoc networks and in [12] a fairly comprehensive method is proposed which could be used in our approach in order to calculate the t_j value.

To predict the link expiration time, we based on the location of nodes as in [13], but for simplicity reasons, we use a simple method that uses mainly a physical layer. Here we propose a scheme which depends of the mobility speed of nodes, by using the received signal strength (RSS). In our prediction method, we assume a free-space propagation model [14], where the RSS solely depends on its distance to the transmitter and the *SSS*. We assume that all nodes transmit their signals with the same strength *SSS*.

The intended receiving node measures the signal strength received, which holds the following relationship for free-space propagation model.

$$RSS = SSS\ (\lambda / 4\ \pi\ d\ )^2\ G_T G_R \tag{5}$$

where λ is wavelength of carrier, G_T and G_R are the unity gain of the transmitting and receiving antennas, respectively. The effects of noise and fading are not considered. The receiver node can calculate d from *RSS,* as illustrated in Figure 5, crossing d value calculated on Phy to the MAC layer.

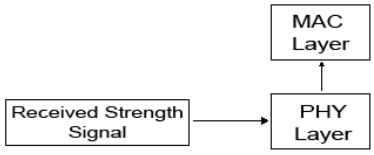


Figure 5
Crossing the distance value from physical layer to MAC layer

Receiving two signals with two different powers in free space involves two different distances ($RSS_1 \neq RSS_2 \rightarrow d_1 \neq d_2$).

If a node detects a distance d_1 at time t_1 from its neighbor and at t_2 detects a distance d_2 different to d_1 from the same neighbor, and d_2 is larger than d_1 (in linear motion) the node concludes that it is moving away from its neighbor with a speed of: $SP = (d_2 - d_1)/(t_2 - t_1)$. Assuming that each node can know at what speed it is moving away from its neighbor. So each node can know at what point the distance from its neighbor reaches the distance of communication C_d, and the expiration of the link with its neighbor. So to calculate the mobility criterion "*M*" of the formula (4), the t_j value is no longer unknown and is calculated by:

$$t_j = (C_d - d)/SP \tag{5}$$

- where d is the last distance the moving neighbor node (in *SP* formula, d is d_2).

4.3 The Protocol Design

For the protocol version of ABE_MM, it is the same version that is used in the ABE technique [2] aiming to examine the mobility criterion in the formula (3). Certainly, there are some changes that has been found necessary. So there are some steps that we have deleted and other new steps added, and it gives the following:

In the admission control, an extension of the RREQ packet to insert a new field where the source indicates the throughput requested (described in previous sections). But now, each mobile which receives a RREQ, performs an admission control by comparing the throughput requirement carried in the RREQ packet by the estimated available bandwidth with the equation (3) on the link where it received the RREQ. If this request is admitted, the node adds its own address to the route and forwards the RREQ; otherwise it discards it. When the destination receives a first RREQ, following the AODV protocol after a control admission, it sends a unicast route reply (RREP) to the initiator of the request along the reverse path. The resources are then reserved and the new QoS flow can be sent.

We have added a new step of sending the route error (RERR) packets by nodes which discover that available bandwidth is no longer sufficient on the reserved link, as in Figure 4 (b) at t_{n-1}. (Then the RERR packets are sent because of available bandwidth failure). This step is caused particularly by the mobility phenomena and the changing of the "*M*" value, where the senders of these RERR packets are the nodes which have predicted the breaking of the link. And another RERR will be sent after the failure of the link by nodes which detect it but the available bandwidth is sufficient.

Note: In ABE_MM, the prediction of link expiration does not reject the QoS flow automatically, but it can be admitted if will be satisfied in the duration of existence.

5 Performances Evaluation

5.1 Evaluation Method

To evaluate the effectiveness of the proposed mobility management approach in bandwidth measurement, we conducted a comparative experiment using Network Simulator 2 (NS2.34) and the IEEE 802.11 implementation provided with the simulator. The scheme used was (CSMA/CA) without (RTS/CTS) Mechanism. Five constant bit rate (CBR) flows (Flow1, Flow2, Flow3, Flow4, Flow5) are generated with 1000 bytes of data packet size. A free space propagation model is

adopted in our experiment, with the radio propagation range length for each node being 250 meters (d) and the channel capacity being 2Mbits/sec.

Our simulation models a network of 5 sources and 5 destinations (for each flow) in 20 mobile nodes that move to a common and constant velocity equal to 12 m/s. The movement of nodes is random, following the model of Manhattan [15] without buildings (without obstacles, to ensure a free space environment) on an area of 1000 m X 1000 m. The street length on this Manhattan model is 100m where nodes move through. At each arrival in a corner, a node can stay where it is, continue its movement in the same direction, or change it. For this simulation we used the prediction model that is explained in the previous section, where nodes predict with accuracy when they move away each from other on a straight line (other prediction models more efficient, such as in [12], can be used). The efficiency of our approach is evaluated through comparison between ABE_MM, ABE and AODV without any QoS guarantee with the following criteria:

- The throughput on destination nodes of each flow obtained along the simulations.
- Average loss ratio: The average of the sent data packets but not received on destinations over the total number of data packets sent.
- Data packet delivery: The total number of data packets delivered to each destination along simulations.

5.2 Simulation Results

After several simulations of random scenarios, we chose one of them, and the results are as follows:

5.2.1 Throughput

Following the simulation logs (movement of nodes, routing flow and control packets) we noted the following:

To show the importance of the QoS garantees, we also provided results obtained with the AODV routing protocol, where no admission control is applied. Figure 6 shows the throughput on destination nodes of the five flows when AODV is used. Since no admission control is performed in AODV and no QoS garantee, each new requesting flow which reaches the desired destination is admitted automatically; therefore all five flows are admitted despite the lack of available of bandwidth, and the network is congested as new flows are added, resulting in a dramatically decreasing in throughput of the flows.

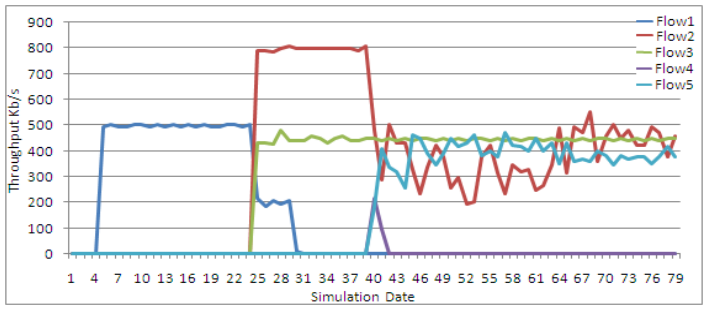


Figure 6
Throughput of each flow using AODV

Figure 7 shows the throughput of the five flows along simulation time when the ABE is enabled for path reservation. We remark that the number of admitted flows are limited compared to the AODV results, and there is no throughput degradation of the accepted flows. But in the absence of a subsequent monitoring of the bandwidth evolution depending on the mobility, the flow (1) continues to consume bandwidth while the path is in failure, penalizing other flows (2) and (3), given the slow process of detecting the link failure. We also note that the absence of a mobility criterion in the ABE measurement formula has allowed flow (4) a reservation path (A false admission control) which contains a link being missing, which caused delays in the admission of the flow(2) and (3).

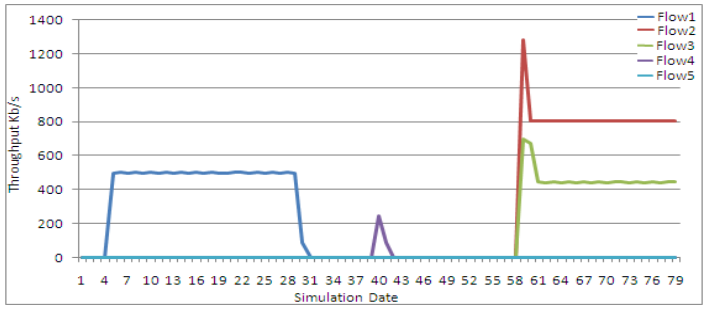


Figure 7
Throughput of each flow using ABE

Figure 8 shows the throughput of the five flows when the path reservation is activated with ABE_MM. We observe that the consumption of bandwidth on the network is much more optimal compared to ABE. Through the equitable utilization of this resource flow (1) is stopped because of the mobility that has reduced the available bandwidth with which it was admitted, allowing a non-belated admission of flows (2) and (3). Also, eliminating unemployment times in the network caused by a flow (4) with ABE technique (the interconnection zone of flow (2) and (4) paths for [31 ... 58] seconds).

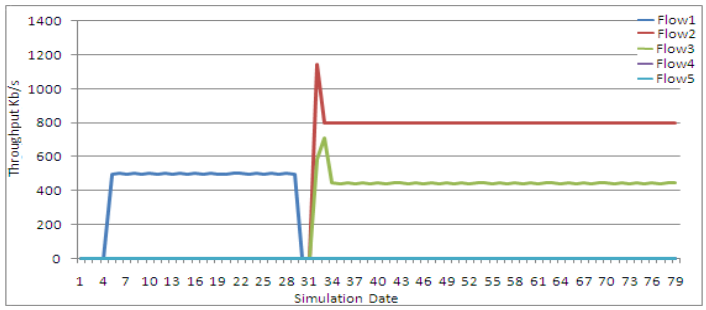


Figure 8
Throughput of each flow using ABE_MM

5.2.2 Average Loss Ratio

The average loss ratio along the simulation is shown in the diagram in Figure 9. We note that the absence of the QoS guaranteed in AODV protocol has not only caused catastrophic degradation on throughput (Figure 6), but also significant loss rates for all flows, particularly flow (4) where almost all data sent are lost. When applying the reservation beforehand of bandwidth with ABE, we notice an improvement in the results compared to AODV, but the absence of the mobility management in the ABE estimations has caused a false admission control of flow (4) resulting in a loss total data. As well, the absence of mobility criterion has caused significant losses in flow 1, because of late detection by the source of the path failure. ABE_MM, with the criterion of mobility, has prevented the loss of data flow by stopping the source of flow 1 at the right time through the predictive detection of path failure. In addition, the admission control of flow 4 has been corrected with respect to ABE, which has helped prevent the loss of data of this flow.

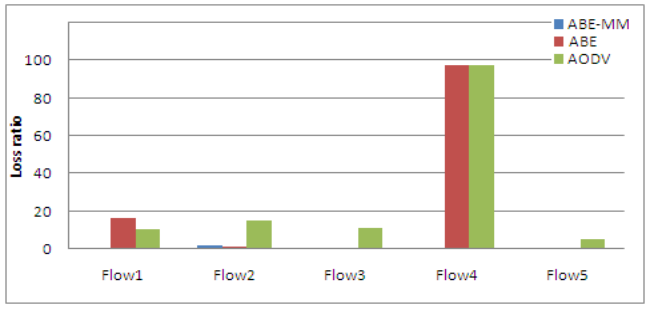


Figure 9
Average loss ratio diagram using ABE_MM, ABE and AODV

5.2.3 Data Packet Delivery

The total data packet delivery to each destination is shown in the diagram of Figure 10. Given that the goal of AODV is to deliver the greatest possible amount of data without taking into consideration any QoS criterion, this is why we find some delivery on large number of flows (quantitatively and not qualitatively). Enabling ABE has reduced the number of deliveries seen the bandwidth guarantees, but this reduction was significant where delivery of flow (2) and flow (3) were low, given their late admissions. With ABE_MM we notice the mobility management has improved the delivery of flows (2) and (3) with admissions at the right time.

Note: The loss ratio and the number of data packets delivery of flow (1) by using ABE confirms that the source of flow (1) continued the transmission after the 31 seconds (Figure 7).

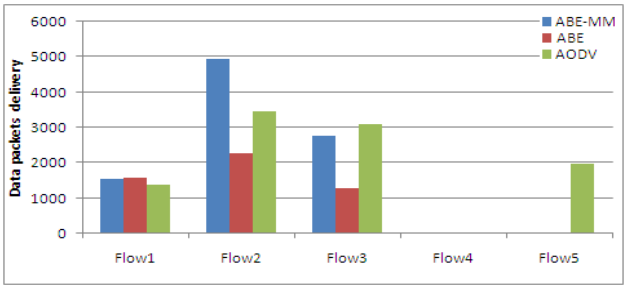


Figure 10
Data packets delivery diagram using ABE_MM, ABE and AODV

Conclusion

In this paper, we present the importance of taking into consideration the mobility phenomenon in available bandwidth measurement, especially during the path reservations. Our solution is based on the distances changing between neighboring nodes which are linked together. ABE_MM was the result of the extension of the ABE technique with our approach.

The results obtained from a comparison between ABE and ABE_MM are satisfactory in terms of the consumption optimality of bandwidth in the network. We have noticed an improvement of flow circulation where the density of traffic has increased over the network while decreasing loss rates.

References

[1] IEEE Computer Society LAN MAN Standards Committee, Wireless LAN Medium Access Protocol (MAC) and Physical Layer (PHY) Specification, IEEE Std 802.11-1997. The Institute of Electrical and Electronics Engineers, New York, 1997

[2] C. Sarr, C. Chaudet, G. Chelius and I. G. Lassous. Bandwidth Estimation for IEEE 802.11-based Ad Hoc Networks, IEEE Transactions on Mobile Computing, Vol. 7, Num. 10, 2008

[3] L. Chen and W. Heinzelman. QoS-aware Routing Based on Bandwidth Estimation for Mobile Ad Hoc Networks. IEEE Journal on Selected Areas of Communication, 3, 2005

[4] R. Renesse, M. Ghassemian, V. Friderikos, A. Hamid Aghvami. QoS Enabled Routing in Mobile Ad Hoc Networks. In IEE3G, 2004

[5] K. Bertet, C. Chaudet, I. G. Lassous, and L. Viennot. Impact of Interferences on Bandwidth Reservation for Ad Hoc Networks: a First Theoretical Study. In *accepted to IEEE Globecom'01*, San Antonio, Texas, USA, November 2001

[6] C. Chaudet, I. G. Lassous. BRuIT - Bandwidth Reservation under InTerferences Influence. In In Proceedings of European Wireless 2002 (EW2002), Florence, Italy, Feb 2002

[7] Y. Yang and R. Kravets. Contention Aware Admission Control for Ad Hoc Networks. IEEE Transactions on Mobile Computing, 4:363-377, 2005

[8] H. Badis and K. Al Agha, QOLSR, QoS Routing for Ad Hoc Wireless Networks Using OLSR, European Transactions on Telecommunications, Vol. 15, No. 4, 2005

[9] C. E. Perkins and E. M. Royer. The Ad hoc On-Demand Distance Vector Protocol. In C. E. Perkins, editor, Ad hoc Networking, pp. 173-219, Addison-Wesley, 2000

[10] H. Zhao, E. Garcia-Palacios, J. Wei, Y. Xi. Accurate Available Bandwidth Estimation in IEEE 802.11-based Ad Hoc Networks. Computer Communications 32 (2009) 1050-1057

[11] D. Gavalas, C. Konstantopoulos, G. Pantziou, Mobility Prediction in Mobile Ad Hoc Networks, Next Generation Mobile Networks and Ubiquitous Computing, Samuel Pierre (Ed.) (Ecole Polytechnique de Montreal, Canada), pp. 226-240, 2010

[12] S. Lee, W. Su, M. Gerla. Ad Hoc Wireless Multicast with Mobility Prediction. Computer Communications and Networks, Proceedings, 4-9, 1999

[13] Y-B. Ko and N. H. Vaidya, Location-aided Routing (LAR) in Mobile Ad-Hoc Networks, IEEE/ACM Wireless Networks, Volume 6, Issue 4, Pages: 307-321, July 2000

[14] T. S. Rappaport, Wireless Communications: Principles and Practice, Prentice Hall, Upper Saddle River, NJ, Oct. 1995

[15] N. Meghanathan1 and S. Gorla, On the Probability of K-Connectivity in Wireless Ad Hoc Networks under Different Mobility Models, International Journal on Applications of Graph Theory in Wireless Ad Hoc Networks and Sensor Networks (GRAPH-HOC), Vol. 2, No. 3, September 2010

Some Categorial Aspects of the Dorroh Extensions

Dorina Fechete

University of Oradea, Department of Mathematics and Informatics
Oradea, Romania
e-mail: dfechete@uoradea.ro

Abstract: Given two associative rings R and D, we say that D is a Dorroh extension of the ring R, if R is a subring of D and D = R ⊕ M for some ideal M ⊆ D. In this paper, we present some categorial aspects of the Dorroh extensions and we describe the group of units of this ring.

Keywords: bimodule; category; functor; adjoint functors; exact sequence of groups; (group) semidirect product

1 Introduction

If R is a commutative ring and M is an R-module then the direct sum $R \oplus M$ (with R and M regarded as abelian groups), with the product defined by $(a,x)\cdot(b,y) = (ab, bx+ay)$ is a commutative ring. This ring is called the idealization of R by M (or the trivial extension of M) and is denoted by $R \ltimes M$. While we do not know who first constructed an example using idealization, the idea of using idealization to extend results concerning ideals to modules is due to Nagata [12]. Nagata in the famous book, Local rings [12], presented a principle, called the principle of idealization. By this principle, modules become ideals.

We note that this ring can be introduced more generally, namely for a ring R and an (R, R) - bimodule M, considering the product $(a,x)\cdot(b,y) = (ab, xb+ay)$.

The purpose of idealization is to embed M into a commutative ring A so that the structure of M as R-module is essentially the same as an A-module, that is, as on ideal of A (called ringification). There are two main ways to do this: the idealization $R \ltimes M$ and the symmetric algebra $S_R(M)$ (see e.g. [1]). Both constructions give functors from the category of R-modules to the category of R-algebras.

Another construction which provides a number of interesting examples and counterexamples in algebra is the triangular ring. If R and S are two rings and M is an $(R,S)-$ bimodule, the set of (formal) matrices

$$\begin{pmatrix} R & M \\ 0 & S \end{pmatrix} = \left\{ \begin{pmatrix} r & x \\ 0 & s \end{pmatrix} : r \in R, s \in S, x \in M \right\}$$

with the component-wise addition and the (formal) matrix multiplication,

$$\begin{pmatrix} r & x \\ 0 & s \end{pmatrix} \cdot \begin{pmatrix} r' & x' \\ 0 & s' \end{pmatrix} = \begin{pmatrix} rr' & rx' + xs' \\ 0 & ss' \end{pmatrix}$$

becomes a ring, called triangular ring (see [10]). If R and S are unitary, then $\begin{pmatrix} R & M \\ 0 & S \end{pmatrix}$ has the unit $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. If we identify R, S and M as subgroups of $\begin{pmatrix} R & M \\ 0 & S \end{pmatrix}$, we can regard $\begin{pmatrix} R & M \\ 0 & S \end{pmatrix}$ as the (abelian groups) direct sum, $R \oplus M \oplus S$. Also, R and S are left, respectively right ideals, and M, $R \oplus M$, $M \oplus S$ are two sided ideals of the ring $\begin{pmatrix} R & M \\ 0 & S \end{pmatrix}$, with $M^2 = 0$, $(R \oplus M \oplus S)/(R \oplus M) \cong S$ and $(R \oplus M \oplus S)/(M \oplus S) \cong R$. Finally, $R \oplus S$ is a subring of $\begin{pmatrix} R & M \\ 0 & S \end{pmatrix}$.

If R and S are two rings and M is an (R,S)-bimodule, then M is a $(R \times S, R \times S)$-bimodule under the scalar multiplications defined by $(r,s)x = rx$ and $x(r,s) = xs$. The triangular ring $\begin{pmatrix} R & M \\ 0 & S \end{pmatrix}$ is isomorphic with the trivial extension $(R \times S) \ltimes M$ and conversely, if R is a ring and M is an (R,R)-bimodule, then the trivial extension $R \ltimes M$ is isomorphic with the subring $\left\{ \begin{pmatrix} a & x \\ 0 & a \end{pmatrix} : a \in R, x \in M \right\}$ of the triangular ring $\begin{pmatrix} R & M \\ 0 & R \end{pmatrix}$.

Thus, the above construction can be considered the third realization of the idealization.

The idealization construction can be generalized to what is called a semi-trivial extension. Let R be a ring and M a (R,R)-bimodule. Assume that $\varphi = [-,-] : M \otimes_A M \to R$ is an (R,R)-bilinear map such that $[x,y]z$

$= x[y,z]$ for any $x, y, z \in M$. Then we can define a multiplication on the abelian group $R \oplus M$ by $(a,x)\cdot(b,y) = (ab + [x,y], xb + ay)$ which makes $R \oplus M$ a ring called the semi-trivial extension of R by M and φ, and denoted by $R \ltimes_{\varphi} M$.

M. D'Anna and M. Fontana in [2] and [3] introduced another general construction, called the amalgamated duplication of a ring R along an R-module M and denoted by $R \bowtie M$. If R is a commutative ring with identity, $T(R)$ is the total ring of fractions and M an R-submodule of $T(R)$ such that $M \cdot M \subseteq M$, then $R \bowtie M$ is the subring $\{(a, a+x) : a \in R,\ x \in M\}$ of the ring $R \times T(R)$ (endowed with the usual componentwise operations).

More generally, given two rings R and M such that M is an (R,R)-bimodule for which the actions of R are compatible with the multiplication in M, i.e.

$$(ax)y = a(xy),\ (xy)a = x(ya),\ (xa)y = x(ay)$$

for every $a \in R$ and $x, y \in M$, we can define the multiplication

$$(a,x)\cdot(b,y) = (ab, xb + ay + xy)$$

to obtain a ring structure on the direct sum $R \oplus M$. This ring is called the Dorroh extension (it is also called an ideal extension) of R by M, and we will denote it by $R \bowtie M$. If the ring R has the unit 1, the ring $R \bowtie M$ has the unit $(1,0)$. Dorroh [5] first used this construction, with $R = \mathbb{Z}$, (the ring of integers), as a means of embedding a (nonunital) ring M without identity into a ring with identity.

In this paper, in Section 3, we give the universal property of the Dorroh-extensions that allows to construct the covariant functor $\mathbf{D} : \mathfrak{D} \to \mathfrak{Rng}$, where $\mathfrak{D}$ is the category of the Dorroh-pairs and the Dorroh-pair homomorphisms. We prove that the functor $\mathbf{D}$ has a right adjoint and this functor commute with the direct products and inverse limits. Also we establish a correspondence between the Dorroh extensions and some semigroup graded rings.

L. Salce in [13] proves that the group of units of the amalgamated duplication of the ring R along the R-module M is isomorphic with the direct product of the groups $\mathbf{U}(R)$ and M°. In Section 4 we prove that in the case of the Dorroh extensions, the group of units $\mathbf{U}(R \bowtie M)$ is isomorphic with the semidirect product of the groups $\mathbf{U}(R)$ and M°.

2 Some Basic Concepts

Recall that if S is semigroup, the ring R is called S**-graded** if there is a family $\{R_s : s \in S\}$ of additive subgroups of R such that $R = \bigoplus_{s\in S} R_s$ and $R_s R_t \subseteq R_{st}$ for all $s,t \in S$. For a subset $T \subseteq S$ consider $R_T = \bigoplus_{t\in T} R_t$. If T is a subsemigroup of S then R_T is a subring of R. If T is a left (right, two-sided) ideal of R then R_T is a left (right, two-sided) ideal of R.

The semidirect product of two groups is also a well-known construction in group theory.

Definition. Given the groups H and N, a group homomorphism $\varphi : H \to \mathrm{Aut}\, K$, if we define on the Cartesian product, the multiplication

$$(h_1,k_1)(h_2,k_2) = (h_1h_2, k_1\cdot\varphi(h_1)(k_2)),$$

we obtain a group, called the semidirect product of the groups H and N with respect to φ. This group is denoted by $H \times_\varphi N$.

Theorem. Let G be a group. If G contain a subgroup H and a normal subgroup N such that $H \cap K = \{1\}$ and $G = K \cdot H$, then the correspondence $(h,k) \mapsto kh$ establishes an isomorphism between the semidirect product $H \times_\varphi N$ of the groups H and N with respect to $\varphi : H \to \mathrm{Aut}\, K$, defined by $\varphi(h)(k) = hkh^{-1}$ and the group G.

Definition. A short exact sequence of groups is a sequence of groups and group homomorphisms

$$1 \longrightarrow N \xrightarrow{\alpha} G \xrightarrow{\beta} H \longrightarrow 1$$

where α is injective, β is surjective and $\mathrm{Im}\,\alpha = \ker\beta$. We say that the above sequence is split if there exists a group homomorphism $s : H \to G$ such that $\beta \circ s = \mathrm{id}_H$.

Theorem. Let G, H, and N be groups. Then G is isomorphic to a semidirect product of H and N if and only if there exists a split exact sequence

$$1 \longrightarrow N \xrightarrow{\alpha} G \xrightarrow{\beta} H \longrightarrow 1$$

3 The Dorroh Extension

To simplify the presentation, we give the following definition:

Definition 1. *A pair* (R,M) *of* (*associative*) *rings*, *is called a Dorroh-pair if* M *is also an* (R,R)*-bimodule and for all* $a \in R$ *and* $x, y \in M$, *are satisfied the following compatibility conditions:*

$(ax)\,y = a(xy),\ (xy)\,a = x(ya),\ (xa)\,y = x(ay).$

We denote further with $\mathcal{D}$, the class of all Dorroh-pairs.

If $(R,M) \in \mathcal{D},$ on the module direct sum $R \oplus M$ we introduce the multiplication

$(a,x)\cdot(b,y) = (ab, xb + ay + xy).$

$(R \oplus M, +, \cdot)$ is a ring and it is denoted by $R \bowtie M$ and it is called the Dorroh extension (or ideal extension (see [8], [11])). Moreover, $R \bowtie M$ is a (R,R)-bimodule under the scalar multiplications defined by

$\alpha(a,x) = (\alpha a, \alpha x), \qquad (a,x)\alpha = (a\alpha, x\alpha)$

and $(R, R \bowtie M)$ is also a Dorroh-pair.

If R has the unit 1, then $(1,0)$ is a unit of the ring $R \bowtie M$. Dorroh first used this construction (see [5]), with $R = \mathbb{Z}$, as a means of embedding a ring without identity into a ring with identity.

Remark 2. *If* M *is a zero ring, the Dorroh extension* $R \bowtie M$ *coincides with the trivial extension* $R \ltimes M$.

Example 3. *If* R *is a ring, then* (R,M) *is a Dorroh-pair for every ideal M of the ring R. Another example of a Dorroh-pair is* $(R, \mathcal{M}_{n\times n}(R))$.

Since the applications

$i_R : R \to R \bowtie M, \quad a \mapsto (a,0)$

$i_M : M \to R \bowtie M, \quad x \mapsto (0,x)$

are injective and both rings homomorphisms and (R,R) linear maps, we can identify further the element $a \in R$ with $(a,0) \in R \bowtie M$ and $x \in M$ with $(0,x) \in R \bowtie M$. Also, the application

$\pi_R : R \bowtie M \to R, \quad (a,x) \mapsto a$

is a surjective ring homomorphism which is also (R,R) linear. Consequently, R is a subring of $R \rtimes\!\ltimes M$, M is an ideal of the ring $R \rtimes\!\ltimes M$, and the factor ring $(R \rtimes\!\ltimes M)/M$ is isomorphic with R.

Remark 4. Given two associative rings R and D, we can say that D is a Dorroh extension of the ring R, if R is a subring of D and $D = R \oplus M$ for some ideal $M \subseteq D$.

If $(A,R),(A,M),(R,M) \in \mathcal{D}$, then M is an $(A \rtimes\!\ltimes R, A \rtimes\!\ltimes R)$-bimodule with the scalar multiplication

$$(\alpha,a)x = \alpha x + ax \quad \text{and} \quad x(\alpha,a) = x\alpha + xa,$$

respectively, $R \rtimes\!\ltimes M$ is an (A,A)-bimodule with the scalar multiplication

$$\alpha(a,x) = (\alpha a, \alpha x) \quad \text{and} \quad (a,x)\alpha = (a\alpha, x\alpha).$$

Obviously, $(A \rtimes\!\ltimes R, M),(A, R \rtimes\!\ltimes M) \in \mathcal{D}$ and since,

$$((\alpha,a),x)+((\beta,b),y) = ((\alpha+\beta, a+b), x+y),$$

$$((\alpha,a),x)\cdot((\beta,b),y) = ((\alpha\beta, \alpha b + a\beta + ab), \alpha y + ay + x\beta + xb + xy),$$

respectively,

$$(\alpha,(a,x))\cdot(\beta,(b,y)) = (\alpha+\beta,(a+b,x+y)),$$

$$(\alpha,(a,x))\cdot(\beta,(b,y)) = (\alpha\beta,(\alpha b + a\beta + ab, \alpha y + ay + x\beta + xb + xy)),$$

the rings $(A \rtimes\!\ltimes R) \rtimes\!\ltimes M$ and $A \rtimes\!\ltimes (R \rtimes\!\ltimes M)$ are isomorphic, and the isomorphism of these rings is given by the correspondence $((\alpha,a),x) \mapsto (\alpha,(a,x))$. Due to this isomorphism, further we can write simply $A \rtimes\!\ltimes R \rtimes\!\ltimes M$.

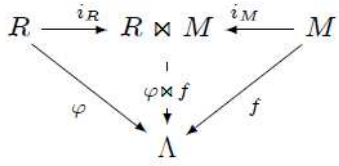
Example 5. *If $R_1,\ldots,R_n$ are rings such that (R_i,R_j) are Dorroh-pairs whenever $i \le j$, we can consider the ring $R = R_1 \rtimes\!\ltimes R_2 \rtimes\!\ltimes \ldots \rtimes\!\ltimes R_n$. Since for any $i,j \in I_n$, $R_i R_j \subseteq R_{\max(i,j)}$, we can consider the ring R as a I_n-graded ring, where I_n is the monoid $\{1,\ldots,n\}$ with the operation defined by $i \vee j = \max(i,j)$. Conversely, if a ring R is I_n-graded and $R = \bigoplus_{i \in I_n} R_i$, since $R_i R_j \subseteq R_{i \vee j}$ for all $i,j \in I_n$, the subgroups $R_1,\ldots,R_n$ are subrings of R and R_j is a (R_i,R_i)-bimodule whenever $i \le j$, the rings R and $R_1 \rtimes\!\ltimes R_2 \rtimes\!\ltimes \ldots \rtimes\!\ltimes R_n$ are isomorphic.*

Definition 6. *By a homomorphism between the Dorroh-pairs* (R,M) *and* (R',M') *we mean a pair* (φ, f), *where* $\varphi: R \to R'$ *and* $f: M \to M'$ *are ring homomorphisms for which, for all* $\alpha \in R$ *and* $x \in M$ *we have that*

$f(\alpha \cdot x) = \varphi(\alpha) \cdot f(x)$ and $f(x \cdot \alpha) = f(x) \cdot \varphi(\alpha)$.

The Dorroh extension verifies the following universal property:

Theorem 7. *If* (R,M) *is a Dorroh-pair, then for any ring* Λ *and any Dorroh-pairs homomorphism* $(\varphi, f): (R,M) \to (\Lambda, \Lambda)$, *there exists a unique ring homomorphism* $\varphi \bowtie f: R \bowtie M \to \Lambda$ *such that*



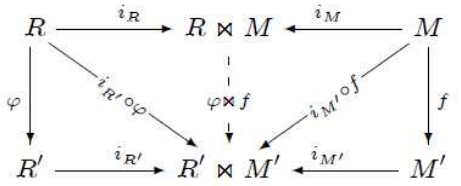
$(\varphi \bowtie f) \circ i_M = f$ and $(\varphi \bowtie f) \circ i_R = \varphi$.

Proof. It is routine to verify that the application $\varphi \bowtie f$, defined by

$(\varphi \bowtie f)(a,x) = \varphi(a) + f(x)$

is the required ring homomorphism.

Corollary 8. *If* (R,M) *and* (R',M') *are two Dorroh-pairs, and* $(\varphi, f): (R,M) \to (R',M')$ *is a Dorroh-pairs homomorphism, then there exists a unique ring homomorphism* $\varphi \bowtie f: R \bowtie M \to R' \bowtie M'$ *such that*



$(\varphi \bowtie f) \circ i_R = i_{R'} \circ \varphi$ and $(\varphi \bowtie f) \circ i_M = i_{M'} \circ f$.

Proof. Apply Theorem 7, considering $\Lambda = R' \bowtie M'$ and the homomorphisms pair $(i_{R'} \circ \varphi, i_{M'} \circ f)$.

Consider now the category $\mathfrak{D}$ whose objects are the class $\mathcal{D}$ of the Dorroh-pairs and the homomorphisms between two objects are the Dorroh-pairs homomorphisms and the category $\mathfrak{Rng}$ of the associative rings.

By Corollary 8, we can consider the covariant functor $\mathbf{D}:\mathfrak{D}\to\mathfrak{Rng}$, defined as follows: if (R,M) is a Dorroh-pair, then $\mathbf{D}(R,M)=R\rtimes M$, and if $(\varphi,f):(R,M)\to(R',M')$ is a Dorroh-pair homomorphism, then $\mathbf{D}(\varphi,f)=\varphi\rtimes f$.

Consider also the functor $\mathbf{B}:\mathfrak{Rng}\to\mathfrak{D}$, defined as follows: if A is a ring, then $\mathbf{B}(A)=(A,A)$ and if $h:A\to B$ is a ring homomorphism, $\mathbf{B}(h)=(h,h)$.

Theorem 9. *The functor* $\mathbf{D}$ *is left adjoint of* $\mathbf{B}$.

Proof. If $(R,M)\in Ob\mathfrak{D}$ and $\Lambda\in Ob\mathfrak{Rng}$, define the function

$$\phi_{(R,M),\Lambda}:Hom_{\mathfrak{Rng}}(R\rtimes M,\Lambda)\to Hom_{\mathfrak{D}}\big((R,M),(\Lambda,\Lambda)\big)$$

by $\Phi\mapsto\left(\Phi|_R,\Phi|_M\right)$, which is evidently a bijection.

Since, for any Dorroh-pairs homomorphism $(\varphi,f):(R,M)\to(R',M')$ and for any ring homomorphisms $\beta:\Lambda\to\Lambda'$ and $\Psi:R'\rtimes M'\to\Lambda$ we have that

$$\begin{aligned}(\beta,\beta)\circ(\Psi|_{R'},\Psi|_{M'})\circ(\varphi,f) &= \big((\beta\circ\Psi|_{R'}\circ\varphi),(\beta\circ\Psi|_{M'}\circ f)\big)\\ &= \big((\beta\circ\Psi\circ i_{R'}\circ\varphi),(\beta\circ\Psi\circ i_{M'}\circ f)\big)\\ &= \big((\beta\circ\Psi\circ(\varphi\rtimes f)\circ i_R),(\beta\circ\Psi\circ(\varphi\rtimes f)\circ i_M)\big)\\ &= \big((\beta\circ\Psi\circ(\varphi\rtimes f))|_R,\ (\beta\circ\Psi\circ(\varphi\rtimes f))|_M\big)\end{aligned}$$

the diagram

$$\begin{array}{ccc}\mathrm{Hom}_{\mathfrak{Rng}}(R'\rtimes M',\Lambda) & \xrightarrow{\phi_{(R',M'),\Lambda}} & \mathrm{Hom}_{\mathfrak{D}}((R',M'),(\Lambda,\Lambda))\\ \downarrow{\scriptstyle \mathrm{Hom}_{\mathfrak{Rng}}(\varphi\rtimes f,\beta)} & & \downarrow{\scriptstyle \mathrm{Hom}_{\mathfrak{D}}((\varphi,f),(\beta,\beta))}\\ \mathrm{Hom}_{\mathfrak{Rng}}(R\rtimes M,\Lambda') & \xrightarrow{\phi_{(R,M),\Lambda'}} & \mathrm{Hom}_{\mathfrak{D}}((R,M),(\Lambda',\Lambda'))\end{array}$$

is commutative and the result follow.

Proposition 10. *Consider* $\{(R_i, M_i) : i \in I\}$ *a family of Dorroh-pairs and the ring direct products* $\prod_{i\in I} R_i$ *and* $\prod_{i\in I} M_i$ *(with the canonical projections* p_i *and* π_i*, respectively, the canonical embeddings* q_i *and* σ_i*)*

Then $\left(\prod_{i\in I} R_i, \prod_{i\in I} M_i\right)$ *is also a Dorroh-pair, for all* $i \in I$, (p_i, π_i) *and* (q_i, σ_i) *are Dorroh-pairs homomorphisms and*

$$\left(\prod_{i\in I} R_i\right) \bowtie \left(\prod_{i\in I} M_i\right) \cong \prod_{i\in I} (R_i \bowtie M_i).$$

Proof. Since for all $i \in I$, (R_i, M_i) are Dorroh-pairs, $\prod_{i\in I} M_i$ is a $\left(\prod_{i\in I} R_i, \prod_{i\in I} R_i\right)$-bimodule with the componentwise scalar multiplications and evidently, the compatibility conditions are satisfied. Thus $\left(\prod_{i\in I} R_i, \prod_{i\in I} M_i\right)$ is a Dorroh-pair.

If $a = (a_i)_{i\in I} \in \prod_{i\in I} R_i$ and $x = (x_i)_{i\in I} \in \prod_{i\in I} M_i$, then for all $j \in I$,

$$\pi_j(a \cdot x) = a_j \cdot x_j = p_j(a) \cdot \pi_j(a) \quad \text{and} \quad \pi_j(x \cdot a) = x_j \cdot a_j = \pi_j(a) \cdot p_j(a)$$

respectively, if $i \in I$, $a_i \in R_i$ and $x_i \in M_i$, then

$$\sigma_i(a_i \cdot x_i) = q_i(a) \cdot \sigma_i(a) \quad \text{and} \quad \sigma_i(x_i \cdot a_i) = \sigma_i(a) \cdot q_i(a)$$

and so (p_i, π_i) and (q_i, σ_i) are Dorroh-pairs homomorphisms.

Proposition 11. *Let* I *be a directed set and* $\left\{(R_i, M_i)_{i\in I}; (\varphi_{ij}, f_{ij})_{i,j\in I}\right\}$ *an inverse system of Dorroh-pairs. Then* $\left\{(R_i \bowtie M_i)_{i\in I}, (\varphi_{ij} \bowtie f_{ij})_{i,j\in I}\right\}$ *is an inverse system of rings and*

$$\lim_{\leftarrow}(R_i \bowtie M_i) \cong \left(\lim_{\leftarrow} R_i\right) \bowtie \left(\lim_{\leftarrow} M_i\right).$$

Proof. Consider the elements $i, j \in I$ such that $i \le j$. By Corolary 8, the Dorroh-pairs homomorphism $(\varphi_{ij}, f_{ij}) : (R_j, M_j) \to (R_i, M_i)$ can be extended to the ring homomorphism $\varphi_{ij} \bowtie f_{ij} : R_j \bowtie M_j \to R_i \bowtie M_i$ which is defined by

$$(\varphi_{ij} \bowtie f_{ij})(a_j, x_j) = (\varphi_{ij}(a_j), f_{ij}(x_j)), \text{ for all } (a_j, x_j) \in R_j \bowtie M_j.$$

Obviously, $\left\{\left(R_i \rtimes\!\ltimes M_i\right)_{i\in I}, \left(\varphi_{ij} \rtimes\!\ltimes f_{ij}\right)_{i,j\in I}\right\}$ is an inverse system of rings.

Consider now $s,t \in I$ such that $s \le t$ and $\left(a_i, x_i\right)_{i\in I} \in \lim\limits_{\leftarrow}\left(R_i \rtimes\!\ltimes M_i\right)$. Since

$$\left(a_s, x_s\right) = \left(\varphi_{st} \rtimes\!\ltimes f_{st}\right)\left(a_t, x_t\right) = \left(\varphi_{st}\left(a_t\right), f_{st}\left(x_t\right)\right)$$

we obtain that $\left(a_i\right)_{i\in I} \in \lim\limits_{\leftarrow} R_i$, $\left(x_i\right)_{i\in I} \in \lim\limits_{\leftarrow} M_i$ and the correspondence

$$\left(a_i, x_i\right)_{i\in I} \mapsto \left(\left(a_i\right)_{i\in I}, \left(x_i\right)_{i\in I}\right)$$

establishes an isomorphism between $\lim\limits_{\leftarrow}\left(R_i \rtimes\!\ltimes M_i\right)$ and $\left(\lim\limits_{\leftarrow} R_i\right) \rtimes\!\ltimes \left(\lim\limits_{\leftarrow} M_i\right)$.

4 The Group of Units of the Ring R⋊⋉M

If A is a ring with identity, denote by $\mathbf{U}(A)$ the group of units of this ring.

Let (R,M) a Dorroh-pair where R is a ring with identity and consider the Dorroh extension $R \rtimes\!\ltimes M$. In this section we will describe the group of units of the ring $R \rtimes\!\ltimes M$. Firstly, observe that if $(a,x) \in \mathbf{U}\left(R \rtimes\!\ltimes M\right)$, then $a \in \mathbf{U}(R)$.

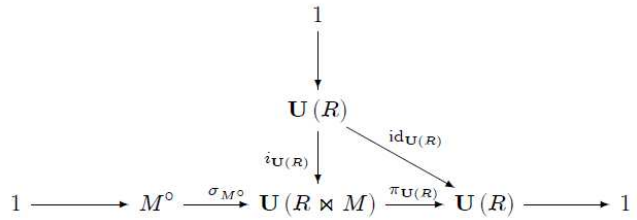
The set of all elements of M forms a monoid under the circle composition on M, $x \circ y = x + y + xy$, 0 being the neutral element. The group of units of this monoid we will denoted by M°.

Theorem 12. *The group of units* $\mathbf{U}\left(R \rtimes\!\ltimes M\right)$ *of the Dorroh extension* $R \rtimes\!\ltimes M$ *is isomorphic with a semidirect product of the groups* $\mathbf{U}(R)$ *and* M°.

Proof. Consider the function

$$\sigma_{M^{\circ}} : M^{\circ} \to \mathbf{U}\left(R \rtimes\!\ltimes M\right), \quad x \mapsto (1, x),$$

which is an injective group homomorphism. Consider also the group homomorphisms $i_{\mathbf{U}(R)} : \mathbf{U}(R) \to \mathbf{U}\left(R \rtimes\!\ltimes M\right)$ and $\pi_{\mathbf{U}(R)} : \mathbf{U}\left(R \rtimes\!\ltimes M\right) \to \mathbf{U}(R)$ induced by the ring homomorphisms $i_R : R \to R \rtimes\!\ltimes M$ and $\pi_R : R \rtimes\!\ltimes M \to R$, respectively. Since the following sequences



are exacts and $\pi_{\mathbf{U}(R)} \circ i_{\mathbf{U}(R)} = id_{\mathbf{U}(R)}$, the group of units $\mathbf{U}(R \bowtie M)$ of the Dorroh extension $R \bowtie M$ is isomorphic with the semidirect product of the groups $\mathbf{U}(R)$ and M°, $\mathbf{U}(R) \times_{\delta} M^{\circ}$. The homomorphism $\delta : \mathbf{U}(R) \to \operatorname{Aut} M^{\circ}$, is defined by $a \mapsto \delta_a$ where $\delta_a : M^{\circ} \to M^{\circ}$, $x \mapsto axa^{-1}$ and the multiplication of the semidirect product $\mathbf{U}(R) \times_{\delta} M^{\circ}$, is defined by

$$(a,x)\cdot(b,y) = \left(ab, x \circ \left(aya^{-1}\right)\right) = \left(ab, x + aya^{-1} + xaya^{-1}\right).$$

The isomorphism between the groups $\mathbf{U}(R) \times_{\delta} M^{\circ}$ and $\mathbf{U}(R \bowtie M)$ is given by $(a,x) \mapsto (a, xa)$.

Remark 13. *If M is a ring with identity, the correspondence $x \mapsto x - 1$ establishes an isomorphism between the groups $\mathbf{U}(M)$ and M°, and therefore the group $\mathbf{U}(R \bowtie M)$ is isomorphic with a semidirect product of the groups $\mathbf{U}(R)$ and $\mathbf{U}(M)$.*

Corollary 14. *The group of units $\mathbf{U}(R \ltimes M)$ of the trivial extension $R \ltimes M$ is isomorphic with a semidirect product of the group $\mathbf{U}(R)$ with the additive group of the ring M.*

Conclusions

The Dorroh extension is a useful construction in abstract algebra being an interesting source of examples in the ring theory.

References

[1] D. D. Anderson, M. Winders, *Idealization of a Module*, Journal of Commutative Algebra, Vol. 1, No. 1 (2009) 3-56

[2] M. D'Anna, *A Construction of Gorenstein Rings*, J. Algebra 306 (2006) 507-519

[3] M. D'Anna, M. Fontana, *An Amalgamated Duplication of a Ring Along an Ideal: the Basic Properties*, J. Algebra Appl. 6 (2007) 443-459

[4] G. A. Cannon, K. M. Neuerburg, *Ideals in Dorroh Extensions of Rings*, Missouri Journal of Mathematical Sciences, 20 (3) (2008) 165-168

[5] J. L. Dorroh, *Concerning Adjunctions to Algebras*, Bull. Amer. Math. Soc. 38 (1932) 85-88

[6] I. Fechete, D. Fechete, A. M. Bica, *Semidirect Products and Near Rings*, Analele Univ. Oradea, Fascicola Matematica, Tom XIV (2007) 211-219

[7] R. Fossum, *Commutative Extensions by Canonical Modules are Gorenstein Rings*, Proc. Am. Math. Soc. 40 (1973) 395-400

[8] T. J. Dorsey, Z. Mesyan, *On Minimal Extensions of Rings*, Comm. Algebra 37 (2009) 3463-3486

[9] J. Huckaba, *Commutative Rings with Zero Divisors*, M. Dekker, New York, 1988

[10] T. Y. Lam, *A First Course in Noncommutative Rings*, Second Edition, Springer-Verlag, 2001

[11] Z. Mesyan, *The Ideals of an Ideal Extension*, J. Algebra Appl. 9 (2010) 407-431

[12] M. Nagata, *Local Rings*, Interscience, New York, 1962

[13] L. Salce, *Transfinite Self-Idealization and Commutative Rings of Triangular Matrices*, preprint, 2008

Definite Integral and the Gibbs Paradox

TianZhi Shi

College of Physics, Electronics and Electrical Engineering, HuaiYin Normal University, HuaiAn, JiangSu, China, 223300
e-mail: stzz2@126.com; stzz2@yahoo.com.cn

Abstract: The Gibbs paradox does not exist at all. Entropy is an additive quantity but not an extensive quantity. Neither entropy nor the increment of entropy is an extensive quantity. When T denotes thermodynamic temperature, lnT does not make sense. The correctness of dimension is a necessary condition for proper formulas. The definite integral is preferable to the indefinite integral; it should be used in science instead of the latter. The Gibbs correction factor is not only cumbersome but also erroneous; it causes some contradictions in statistical mechanics. Entropy is not a quantity that can be measured, but the increment of entropy can be measured. So what the experiment can check is not entropy but the increment of entropy.

Keywords: Gibbs paradox; entropy; extensive; additive; definite integral

1 Introduction

In many textbooks on thermodynamics and statistical physics, there are some erroneous formulas. The causes are: 1. The dimension is not correct; 2. When the definite integral should be used on all terms of a differential equation, some terms are integrated definitely but others not; 3. The basic notion is wrong; 4. The argumentation is illogical. This article solves these problems exemplarily.

2 Does lnT Make Sense?

Symbol explanation : "≡"denotes identity; $\frac{\partial f(x,y)}{\partial x} \equiv f_x \equiv \left(\frac{\partial f}{\partial x}\right)_y \equiv f_{x|y}$.

As is well known, the argument of lnx is a pure number, either a real number or a complex number. So when T denotes a pure number, lnT makes sense; when T denotes thermodynamic temperature, lnT does not make sense, nor does lnV when V denotes thermodynamic volume. If a quantity x is not a pure number, neither lnx nor e^x makes sense. Only when we specify the unit of quantity x so that x denotes a pure number can lnx or e^x make sense.

2.1 There Are Two Erroneous Formulas in Literature [1, p. 53]

$$S = nC_{V,m}\ln T + nR\ln V + S_0 \quad <1.15.4>$$

$$S_0 = n(S_{m0} - R\ln n) \quad <1.15.5>$$

Because $\ln T$, $\ln V$, $\ln n$ do not make sense, the above formulas are wrong. Let C_V be the heat capacity of ideal gas at constant volume, N the number of molecules of the gas, and let k denote Boltzmann's constant. $dU = TdS - pdV$, $pV = NkT$, so the correct formulas are as follows:

$$dS = \frac{C_V}{T}dT + \frac{p}{T}dV = \frac{C_V}{T}dT + \frac{Nk}{V}dV \quad (1)$$

$$\int_{S_0}^{S} dS = \int_{T_0}^{T}\frac{C_V}{T}dT + \int_{V_0}^{V}\frac{p}{T}dV = \int_{T_0}^{T}\frac{C_V}{T}dT + \int_{V_0}^{V}\frac{Nk}{V}dV \quad (2)$$

$$\therefore S - S_0 = \int_{T_0}^{T}\frac{C_V}{T}dT + \int_{V_0}^{V}\frac{Nk}{V}dV \quad (3)$$

For monatomic ideal gas : $C_V = \frac{3}{2}Nk$

$$\therefore S - S_0 = \frac{3}{2}Nk\ln\frac{T}{T_0} + Nk\ln\frac{V}{V_0} = Nk\ln[\frac{V}{V_0}(\frac{T}{T_0})^{\frac{3}{2}}] = Nk\ln[\frac{p_0}{p}(\frac{T}{T_0})^{\frac{5}{2}}] \quad (4)$$

2.2 There Are The Following Two Formulas in Literature [1, p. 81]

$$dS = \frac{C_V}{T}dT + p_{T|V}dV \quad <2.4.4>$$

$$S = \int\frac{C_V}{T}dT + p_{T|V}dV + S_0 \quad <2.4.5>$$

Formula <2.4.5> is wrong because when equation <2.4.4> is integrated, some terms use the definite integral but others not. The correct method is that all terms should use the definite integral as follows:

$$\int_{S_0}^{S} dS = \int_{T_0}^{T}\frac{C_V}{T}dT + \int_{V_0}^{V} p_{T|V}dV \quad (5)$$

$$S = S_0 + \int_{T_0}^{T}\frac{C_V}{T}dT + \int_{V_0}^{V} p_{T|V}dV \quad (6)$$

The merit of the definite integral is clear and explicit. The indefinite integral has an arbitrary constant; it causes misunderstanding and erroneous argumentation easily, so it should only be used as a training method in calculus. In order to avoid ambiguity, the definite integral should be used in science instead of the indefinite integral.

2.3 There Are The Following Two Formulas in Literature [1, p. 346]

$$S = k\ln\Omega = Nk\ln[\frac{V}{Nh^3}(\frac{4\pi mE}{3N})^{\frac{3}{2}}] + \frac{5}{2}Nk + k(\ln\frac{3N}{2} + \ln\frac{\Delta E}{E})$$

$$S = Nk\ln[\frac{V}{Nh^3}(\frac{4\pi mE}{3N})^{\frac{3}{2}}] + \frac{5}{2}Nk \quad <9.3.26>$$

Apparently formula <9.3.26> is wrong because when $\Delta E \to 0+$, $\ln\frac{\Delta E}{E} \to -\infty$, we cannot get it from the former formula. This kind of error is illogical argumentation. It is an exemplary case of omitting some terms unreasonably. It shows that an ideal isolated system which does not exchange mass and energy with its surroundings does not exist. Furthermore, the micro-canonical ensemble should not be used in the case in question.

3 Is Entropy an Extensive Quantity?

Many textbooks [2-4] classify thermodynamic quantities as two classes: extensive or intensive. The following is a typical definition [3]:

The quantities we use to describe the macroscopic behavior of a system in equilibrium are called properties, the observable characteristics of a system. Other names are thermodynamic variables or thermodynamic coordinates. An extremely important concept is that of a state variable, a property whose differential is exact.

Properties are extensive or intensive. An extensive is proportional to the mass. An example is the volume V; if the mass is doubled, the volume is doubled (assuming that the density remains constant). An intensive property is independent of the mass. Temperature T is an intensive property; its value is not affected by a change of mass. Pressure p and density ρ are further examples of intensive properties.

In other words, extensive quantities are ones that are proportional to the number of molecules and are additive; intensive quantities are ones that are got by an extensive quantity divided by another extensive quantity and are non-additive. But this kind of classification is not complete. The complete classification is as follows:

$$\text{scalar}\begin{cases}\text{additive scalar}\begin{cases}\text{extensive and additive (mass, volume,}\\ \text{number of particles etc.)}\\ \text{non-extensive but additive (time, entropy,}\\ \text{probability etc.)}\end{cases}\\ \text{non-additive scalar}\begin{cases}\text{intensive and non-additive (temperature,}\\ \text{density etc.)}\\ \text{nonintensive and nonadditive (speed,}\\ \text{rate of increase etc.)}\end{cases}\end{cases}$$

From Boltzmann entropy $S = k\ln\Omega = k\ln(N!\prod_{i=1}^{l}\frac{g_i^{n_i}}{n_i!})$ and Gibbs entropy [4] $S = -k\sum_j P_j \ln P_j$,we can know that entropy is not an extensive quantity.

A system is an entity which consists of many particles and has a definite volume and temperature. The equation for the entropy of a monatomic gas is:

$$S = Nk\left\{\frac{3}{2} + \ln\left[\frac{V}{h^3}(2\pi mkT)^{\frac{3}{2}}\right]\right\} \quad (7)$$

It follows that entropy is a systematic quantity, a measure of disorder of the system.

Let Ω be the number of microstates for a system in equilibrium. If the system is composed of two independent subsystems, then $\Omega = \Omega_1\Omega_2$, $S = k\ln\Omega = k\ln(\Omega_1\Omega_2) = k\ln\Omega_1 + k\ln\Omega_2 = S_1 + S_2$, so entropy is an additive quantity. For a system in equilibrium, all the extensives are accumulations of the corresponding quantity of a particle and are proportional to the number of particles. A particle has mass but not entropy, so mass is extensive but entropy is not. If entropy S were extensive, then $\frac{S}{N}$ would be the entropy of a particle, which (like the temperature of a particle) is meaningless.

Some people confuse additive quantity with extensive quantity, others confuse

entropy with the increment of entropy, which brings about some erroneous notions in thermodynamics and statistical mechanics. Correcting these wrong notions is one of the main reasons why I have written this article.

4 Is the Gibbs Correction Factor Right?

There is the following problem in literature [5]:

There are two ideal monatomic gases in an adiabatic container with volume V. The number of molecules are N_1 and N_2 respectively. The temperature of the system in equilibrium is T. Find the equation of state, internal energy and entropy of the system, starting with canonical ensemble.

Solution: Let m_1 and m_2 be the molecular mass of the two gases respectively, the energy of the system is $E = \sum_{i=1}^{3N_1} \frac{{p_{1i}}^2}{2m_1} + \sum_{j=1}^{3N_2} \frac{{p_{1j}}^2}{2m_2}$ (8).

$$\text{Partion function is}: Z = \frac{1}{N_1!N_2!h^{3N_1}h^{3N_2}} \int e^{-\beta E} d\Gamma_A d\Gamma_B$$

$$= \frac{V^{N_1}}{N_1!h^{3N_1}} (\frac{2\pi m_1}{\beta})^{\frac{3N_1}{2}} \frac{V^{N_2}}{N_2!h^{3N_2}} (\frac{2\pi m_2}{\beta})^{\frac{3N_2}{2}} = Z_1 Z_2 \quad (9)$$

The pressure of the mixed ideal gases is:

$$p = \frac{1}{\beta}(\ln Z)_V = \frac{N_1 kT}{V} + \frac{N_2 kT}{V} = \frac{(N_1 + N_2)kT}{V} \quad (10)$$

The equation of state is: $pV = (N_1 + N_2)kT$ (11)

The pressure of the mixed ideal gases is equal to the sum of the pressures generated by all the constituent gases; this is Dalton's law of partial pressures. When the molecules are of the same kind, formula (10) still holds, which means that the pressure of the ideal gas is equal to the sum of the pressures generated by all the constituent subsystems.

$$U = -(\ln Z)_\beta = \frac{3}{2}(N_1 + N_2)kT \quad (12)$$

The internal energy of the mixed ideal gases is equal to the sum of the internal energies of all the constituent gases. When the molecules are of the same kind, formula (12) still holds, which means that the internal energy of the ideal gas is equal to the sum of the internal energies of all the constituent subsystems.

$$S = k[\ln Z - \beta(\ln Z)_\beta] = N_1 k \ln[\frac{V}{N_1 h^3}(2\pi m_1 kT)^{\frac{3}{2}}] + \frac{5}{2}N_1 k$$

$$+N_2 k \ln[\frac{V}{N_2 h^3}(2\pi m_2 kT)^{\frac{3}{2}}] + \frac{5}{2}N_2 k \quad (13)$$

The entropy of the mixed ideal gases is equal to the sum of the entropies of all the constituent gases. When the molecules are of the same kind, it follows from formula (13):

$$S = N_1 k \ln[\frac{V}{N_1 h^3}(2\pi mkT)^{\frac{3}{2}}] + \frac{5}{2}N_1 k + N_2 k \ln[\frac{V}{N_2 h^3}(2\pi mkT)^{\frac{3}{2}}] + \frac{5}{2}N_2 k \quad (14)$$

According to the Sackur-Tetrode formula, the entropy of monatomic ideal gas is:

$$S' = Nk \ln[\frac{V}{Nh^3}(2\pi mkT)^{\frac{3}{2}}] + \frac{5}{2}Nk$$

$$= (N_1 + N_2)k \ln[\frac{V}{Nh^3}(2\pi mkT)^{\frac{3}{2}}] + \frac{5}{2}(N_1 + N_2)k \quad (15)$$

$$S - S' = (N \ln N - N_1 \ln N_1 - N_2 \ln N_2)k = Nk(\frac{N_1}{N}\ln\frac{N}{N_1} + \frac{N_2}{N}\ln\frac{N}{N_2}) > 0$$

Apparently $S \neq S'$, which means that the entropy of the ideal gas is not equal to the sum of the entropies of all the constituent subsystems. This is contradictory to the additive quality of entropy. The cause that makes the error is introducing the Gibbs correction factor $\frac{1}{N!}$ to the partition function of canonical ensemble. The introduction of the Gibbs correction factor $\frac{1}{N!}$ to the partition function is not only cumbersome, in that it does not change the probability distribution of the system in question; but it is also erroneous: in that it is based on the wrong notion that entropy is an extensive quantity, it leads to some contradictions in statistical mechanics [6-9]. It may be justified in the grand canonical ensemble as an ad hoc prescription [8]: *the statistical weight of a state (l) must be weighted by* $\frac{1}{N_l!}$.

The correct calculation for the entropy is by using formula (7):

$$S_1 = N_1 k\left\{\frac{3}{2} + \ln\left[\frac{V}{h^3}\left(2\pi m_1 kT\right)^{\frac{3}{2}}\right]\right\}; S_2 = N_2 k\left\{\frac{3}{2} + \ln\left[\frac{V}{h^3}\left(2\pi m_2 kT\right)^{\frac{3}{2}}\right]\right\};$$

$$S = S_1 + S_2 = N_1 k \ln[\frac{V}{h^3}(2\pi m_1 kT)^{\frac{3}{2}}] + \frac{3}{2}Nk + N_2 k \ln[\frac{V}{h^3}(2\pi m_2 kT)^{\frac{3}{2}}] (13').$$

The entropy of the mixed ideal gases is equal to the sum of the entropies of all the constituent gases. When the molecules are of the same kind, formula $(13')$ still holds, which means that the entropy of the ideal gas is equal to the sum of the entropies of all the constituent subsystems.

5 Does the Gibbs Paradox Exist?

There are two different, classical, ideal monatomic gases initially in the two compartments (volumes V_1 and $V_2, V_1 + V_2 = V$) of a rigid adiabatic container, separated by a diathermal diaphragm. They have the same pressure and temperature, the number of molecules of each compartment are N_1 and N_2 respectively, $N_1 + N_2 = N$. If the diaphragm is removed, calculate the increment of entropy of the system.

Solution 1(thermodynamic method): Let m_1 and m_2 be the molecular mass of the two gases respectively.

Initial state: $\text{system } N_1 \to V_1, T_0, P_0; \text{system } N_2 \to V_2, T_0, P_0$

Final state: $\text{system } N_1 \to V, T, P_1; \text{system } N_2 \to V, T, P_2$.

Because there is no energy, heat introducing and chemical reaction, the temperature of the system remains constant, i.e. $T = T_0$. Using Dalton's law of partial pressure, we know:

$$p_1 = \frac{N_1 kT}{V}; p_2 = \frac{N_2 kT}{V}$$

$$\therefore p = p_1 + p_2 = \frac{(N_1 + N_2)kT}{V} = \frac{NkT}{V} = p_0 = \frac{N_1 kT}{V_1} = \frac{N_2 kT}{V_2}$$

From formula (4), we have:

$$\Delta S_1 = S_1 - S_{10} = N_1 k \ln[\frac{V}{V_1}(\frac{T}{T_0})^{\frac{3}{2}}] = N_1 k \ln(\frac{V}{V_1}) = N_1 k \ln(\frac{N}{N_1})$$

$$\Delta S_2 = S_2 - S_{20} = N_2 k \ln[\frac{V}{V_2}(\frac{T}{T_0})^{\frac{3}{2}}] = N_2 k \ln(\frac{V}{V_2}) = N_2 k \ln(\frac{N}{N_2})$$

So the increment of entropy after mixing is:

$$\Delta S = \Delta S_1 + \Delta S_2 = N_1 k \ln(\frac{V}{V_1}) + N_2 k \ln(\frac{V}{V_2}) = N_1 k \ln(\frac{N}{N_1}) + N_2 k \ln(\frac{N}{N_2})$$

$$> 0 \quad (16)$$

Apparently the increment of entropy after mixing is the same whether the two gases are different or not, because the two gases will occupy the total volume finally and homogeneously. Meanwhile, from formula (16) we can see that the increment of entropy is not an extensive quantity.

Solution 2(statistical mechanics method): From formula (7) we have:

$$S_{10} = N_1 k \left\{ \frac{3}{2} + \ln\left[\frac{V_1}{h^3} \left(2\pi m_1 kT\right)^{\frac{3}{2}} \right] \right\}; S_1 = N_1 k \left\{ \frac{3}{2} + \ln\left[\frac{V}{h^3} \left(2\pi m_1 kT\right)^{\frac{3}{2}} \right] \right\}$$

$$\therefore \Delta S_1 = S_1 - S_{10} = N_1 k \ln(\frac{V}{V_1}). \text{ Similarly, } \Delta S_2 = S_2 - S_{20} = N_2 k \ln(\frac{V}{V_2}).$$

$\therefore \Delta S = \Delta S_1 + \Delta S_2 = N_1 k \ln(\frac{V}{V_1}) + N_2 k \ln(\frac{V}{V_2})$. The result is the same as solution 1.

Some people think that if the two gases are identical, there can be no change in entropy when the diaphragm is removed ($\Delta S = 0$). This is the so-called Gibbs paradox.

Firstly, neither entropy nor the increase of entropy is an extensive quantity. Based on the wrong notion that entropy is an extensive quantity, some people argue that $\Delta S = 0$.

Secondly, after the diaphragm is removed, each gas moves into the other, and a new equilibrium state is obtained in which both gases occupy the total volume. It is clear that the process of mixing is irreversible: once the mixing is done, the two gases, whether they are different or not, will not return spontaneously to their initial compartments. The final state is more disordered than the initial one. According to the second law of thermodynamics, $dS \geq \frac{dQ}{T} = \frac{dU + pdV}{T}$. For adiabatic process, $dQ = 0$. For irreversible adiabatic process, $dS > 0$.

Thirdly, if we re-install the diaphragm, we will not regain the initial state. The number of molecules in V_1 may not be exactly N_1, let alone they are the same initial N_1 molecules. The reason why mixing takes place is that molecules move randomly and forever.

Fourthly, the above two solutions give the same result whether the gases are different or not. For the mixing of two identical gases, literature [3] gives the third solution. Before the diaphragm is removed, there are N_1, N_2 molecules in V_1, V_2 respectively. This is only one method of arranging the N molecules with N_1 in one container and N_2 in the other, but arranging the N molecules with N_1 in one container and N_2 in the other has $\frac{N!}{N_1!N_2!}$ methods, so the increment of entropy after the diaphragm is removed is: $\Delta S = k\ln\Omega - k\ln\Omega_0 = k\ln\frac{\Omega}{\Omega_0} = k\ln\frac{N!}{N_1!N_2!}$.

Using Stirling's approximation, $\ln N! = N\ln N - N$, we obtain:

$$\Delta S = k\ln\frac{N!}{N_1!N_2!} = N_1 k\ln(\frac{N}{N_1}) + N_2 k\ln(\frac{N}{N_2}) = N_1 k\ln(\frac{V}{V_1}) + N_2 k\ln(\frac{V}{V_2})$$

Lastly, classical particles are distinguishable by their positions in the phase lattice and may be traced by their trajectory. They obey the Maxwell-Boltzmann distribution [3]. The total number of microstates corresponding to an allowable configuration is $\Omega_{MB} = N!\sum_{i=1}^{l}\frac{g_i^{n_i}}{n_i!}$. Quantum particles, both bosons and fermions, are indistinguishable. The total number of microstates obeying the Bose-Einstein distribution is $\Omega_{BE} = \sum_{i=1}^{l}\frac{(g_i+n_i-1)!}{n_i!(g_i-1)!}$; while those obeying the Fermi-Dirac distribution is $\Omega_{FD} = \sum_{i=1}^{l}\frac{g_i!}{n_i!(g_i-n_i)!}$. Apparently $\Omega_{MB} > \Omega_{BE} > \frac{\Omega_{MB}}{N!} > \Omega_{FD}$. Because $S = k\ln\Omega$,

$S_{MB} > S_{BE}$; $S_{MB} > S_{FD}$.

A particle cannot be distinguishable and indistinguishable simultaneously, so it cannot obey the Maxwell-Boltzmann distribution and quantum distribution (either the Bose-Einstein distribution or the Fermi-Dirac distribution) simultaneously.

Therefore the classic Maxwell-Boltzmann distribution is not the limit case of quantum distributions. They have their respective ranges of applicability. Different particles obey different distribution. Under standard temperature and pressure [8], the ideal gases obey the Maxwell-Boltzmann distribution. They are distinguishable and within the classical regime.

Conclusions

In all, the Gibbs paradox does not exist at all. Entropy is additive but not extensive. Neither entropy nor the increment of entropy is extensive. Some people confuse entropy with the increment of entropy, others confuse additive quality with extensive quality. Based on a wrong notion that entropy is extensive, the Gibbs correction factor $\frac{1}{N!}$ is added to the partition function of canonical ensemble. This is not only cumbersome but also erroneous. Entropy is not a quantity that can be measured, but the increment of entropy can be measured. So what the experiment can check is not entropy but the increment of entropy.

References

[1] Wang, Z. C. Thermodynamics and Statistical Physics (3rd edition) [M]. Beijing: Higher Education Press, 2003:17-18, 53-55, 81-83, 276-278, 341-347 (in Chinese)

[2] [German] Greiner, W.; Neise, L.; Stocker, H. Translated into Chinese by Zhong, Y. X. Thermodynamics and Statistical Mechanics [M]. Beijing: Beijing University Press, 2001:4-5

[3] Carter, A. H. Classical and Statistical Thermodynamics [M]. Beijing: Tsinghua University Press, 2007:5-6, 216-273

[4] Cowan, B. Topics in Statistical Mechanics [M]. Shanghai: Fudan University Press, 2006:3-4, 18-20, 67-69

[5] Wang, Z. C. A Guidance Book for Thermodynamics and Statistical Physics (3rd edition) [M]. Beijing: Higher Education Press, 2004:224-225 (in Chinese)

[6] Su, R. K. Statistical Physics (2nd edition) [M]. Beijing: Higher Education Press, 2004:1-3, 72-74, 144-147 (in Chinese)

[7] Schwabl, F. Statistical Mechanics (2nd edition) [M]. Beijing: Science Press, 2008:115-117

[8] Bellac, M. L.; Mortessagne, F.; Batrouni, G. G. Equilibrium and Non-equilibrium Statistical Thermodynanics [M]. Beijing: World Publishing Corporation Beijing Company, 2007:74-77, 98-101

[9] Plischke, M.; Bergersen, B. Equilibrium Statistical Physics (2nd edition) [M]. Shanghai: Fudan University Press, 2006:32-35, 43-44

Multipattern Road Traffic Crashes and Injuries: A Case Study of Xi'an City

Yong-gang Wang, Shen-sen Huang

School of Highway, Chang'an University
Middle Section of South 2 Ring Rd., Xi'an 710064, P. R. China
e-mail: sdqdwyg@163.com, hss.008@163.com

Wen-sen Xiang

Shanghai Municipal Engineering Design Institute (Group) Co., LTD
901 Zhongshan North Second Rd., Shanghai 200096, P. R. China
e-mail: xiangwensen0115@163.com

Yu-long Pei

School of Transportation Science and Engineering, Harbin Institute of Technology
73 Huanghe Rd., Harbin 150090, P. R. China
e-mail: yulongp@263.net

Abstract: Many studies focused on the development of crash analysis approaches have resulted in aggregate practices and experiences to quantify the safety effects of human, geometric, traffic and environmental factors on the expected number of deaths, injuries, and/or property damage crashes at specific locations. Traffic crashes on roads are a major cause of road crashes in the metropolitan area of Xi'an. In an attempt to identify causes and consequences, reported traffic crashes for six years in Xi'an were analyzed using a sample of 2038 reports. The main types of information from such reports were extracted, coded, and statistically analyzed. Important results were obtained from frequency analyses as well as multiple contributory factors related to traffic crashes, including crash severity, time and location of occurrence, geometry of the road, AADT and v/c. This paper presents the results of such analyses and provides some recommendations to improve traffic safety and further studies to analyze potential crash locations.

Keywords: traffic accidents; crash features; contributory factor; crash type; v/c

1 Introduction

Worldwide deaths, disabilities, and injuries from road accidents, a major concern all over the world, have reached epidemic proportions. In 2002, more than 1.18 million people died in road crashes, and approximately fifteen million are injured annually (Source: World report on road traffic injury prevention: summary, 2004). In Europe, the sharp increase in accidents related to urban traffic costs more than two billion dollars during recent years. In Britain, around 3,500 people are killed each year and around 33,000 people are injured in road accidents. There were nearly 6,420,000 auto accidents reported to the police in the United States in 2005, and 2.9 million people were injured and 42,636 people killed, in which talking on a cell phone caused nearly 25% of accidents (Source: http://www.griefspeaks.com/id114.html). In New York City, specifically, the review of 2010 traffic accident database found: 1) 269 traffic deaths of all types and a 34 percent increase in auto fatalities; 2) 18 and 152 fatalities from bicycle accidents and pedestrian accidents, respectively (Source: http://www.lawfitz.com/new-york-city-traffic-accident-deaths-rise-slightly-in-2010). Nowadays, China has also witnessed a substantial rise in the number of traffic crashes, injuries and fatalities, especially since 1998 [1, 2].

In response to these issues, various studies have recently examined aspects of motorcycle safety, in combination with available influencing variables and causes information. Typically, researchers employ statistical techniques (i.e., Poisson, negative binomial and regression models, etc.) in these types of studies [3]. Over the past years, numerous investigations have looked at the age and gender of drivers as risk factors, and younger and older drivers are more prone to be involved in a serious accident than median-aged ones [4]. Road geometric type, lightness, weather condition and other environmental factors (e.g., the time of the day, traffic volume, etc.) also play an important role when analyzing the aspects of accident crashes and injuries [5~8].

To provide a broad overview, numerous statistical researches have been conducted in order to understand the relations between influence factors and crash features via types of models [9, 10]. Another class of studies utilize time series techniques for identifying the change of accident counts and accident rates at a given time, analyzing the effects of potential factors on accident occurrence, and developing a statistical model for forecasting future trends of crashes and injuries such as in references [11~13], and so on. Actually, each crash is a unique event that is caused or influenced by combinations of variables that may not even be observable [14]. Furthermore, reaching such conclusions that a set of variables can be identified as the causes of traffic crash is almost impossible.

This study follows an earlier work on the analysis of traffic accidents at intersections [15] and aims to identify the multiple crash features in Xi'an city. Several potential factors including district, human, vehicle, time, traffic volume,

environmental and site factors are considered. Statistical analyses of multiple vehicle traffic accidents are conducted using the crash data in Xi'an city over the time of 2004~2009. The corresponding safety improvement measures are suggested so as to enhance the safety performance of road traffic and decrease the probability of crash occurrence in China's metropolitan regions.

2 Data

Xi'an is the capital of the Shaanxi province and one of the oldest cities in China, with more than 3,100 years of history experiencing Zhou, Qin, Han, Sui, and Tang. Since the 1990s, as part of the economic revival of interior China especially for the central and northwest regions, Xi'an has re-emerged as one of the most populous metropolitan areas in inland China, with more than 7 million inhabitants, which has also alarmingly increased the annual traffic crashes and injuries.

The research relies mainly on accident reports archived from General Department of Transportation and its different branches located in different traffic policy teams in Xi'an city. All 2361 crashes within 6 districts, namely Xincheng, Beilin, Yanta, Lianhu, Baqiao and Weiyang, from the year 2004 to 2009 were collected randomly without any specifications or criteria. In this paper, only motor vehicle crashes are considered.

For each sample, a copy of the accident report was obtained to learn more details about the crash, such as the time, location, cause, type of crash, weather, drivers and vehicles. Unfortunately, missing or incomplete data was a common problem in the crash reports (e.g., the condition of road surface, control pattern of intersection, etc.). Thus it is difficult to know exactly how some crashes occurred simply from the accident report [16]. On the other hand, hand-writing is commonly used to record and draw the crash description on spot, and it is sometimes clear enough to identify the crash details.

Therefore, 323 records accounting for 13.68% of the original collection were removed from the crash database due to missing or incomplete information, and this left 2038 samples used for the future research, in which 12.41% proved fatal, 28.16% resulted in serious injuries and the rest in slight injuries.

3 Crash Features

Analyses were performed on the data collected from the General Administration of Transportation in the form of accident reports. The analyses included frequency, cross-classification tabulation, and construction of crash spot maps.

3.1 Temporal-Spatial Patterns

Xi'an is at the top of list of cities with alarmingly high traffic crash fatality rates in China. A summary of crash type statistics for the year 2004~2009 for each district is presented in Table 1. More than 85% of crashes were vehicle collisions/crashes. About 53.8% of the total road traffic accidents occurred in the Yanta district and Beilin district.

Table 1
Number of crashes and injuries by districts

Districts	Crash Type					Injuries			Deaths
	Collision	Running over	Overturning	Drugs	Total	Minor	Major	Total	
Yanta	587	24	32	19	662	70	36	105	38
Beilin	509	12	23	4	548	54	15	68	36
Lianhu	274	15	24	13	326	79	42	121	45
Weiyang	92	21	18	3	134	61	30	91	57
Xincheng	171	14	31	23	239	69	26	95	60
Baqiao	101	11	9	7	128	75	19	94	17
Total	1734	97	137	69	2038	407	167	574	253

Around 318 fatalities occurred in the year 2009 with a decreasing trend. However, the overwhelmingly high proportion of younger drivers involved in traffic crashes has become a serious new concern. Over 60% of the fatalities for 2009 were in the age group of 18-40, though Xi'an had implemented new and stricter penalties under the new national traffic laws in May 2008 in order to minimize traffic crashes. The statistics show that the main cause of crashes is careless driving (e.g., mobile phone use while driving, fatigue, obsessive behavior, etc.) and speeding, which represent about 43% and 26% of the total crashes, respectively. The second main cause of crashes is drunk driving representing about 14%. The rest of the reasons together represent about 17% of total crashes due to facilities, weather and other effects. Therefore, all but a few crashes can be blamed on careless driving, and other factors also play important roles.

Statistics also demonstrate that the safety situation in Xi'an was still serious as shown by the numbers of reported crashes during the years 2004~2009, which are presented in Table 2. The number of crashes increased steadily during the years 2004~2006, and following this period began to decrease. Collisions represented the larger proportion of the total number of crashes. However, the crash rate was still at a high rate. The average number of crashes per day increased from 0.91 in 2004 to 1.06 in 2006 and then decreased to 0.85 in 2009. Crashes involving personal injuries fluctuated between a daily average of 0.15 to 0.4 and, fatal crashes which included at least one death occurred at a daily average of about 0.13, which meant perhaps one life was lost weekly because of road crashes.

Table 2
Statistics of traffic crashes over the year 2004-2009

Year	2004	2005	2006	2007	2008	2009
Total number of crashes	332	346	387	345	317	311
Collision crashes	270	284	324	291	262	259
Crash of overrun/overturning/drugs/others	62	61	64	54	55	52
Total injuries	113	73	102	146	86	54
Minor injuries	76	48	68	107	60	38
Major injuries	37	25	34	39	26	16
Fatal crashes	41	41	37	39	46	49
Average no. of crashes per day	0.91	0.95	1.06	0.95	0.87	0.85
Average no. of injuries per day	0.31	0.20	0.28	0.40	0.24	0.15
Average no. of fatalities per day	0.11	0.11	0.10	0.11	0.13	0.13

3.2 Frequency in Age and Weather

Frequency analysis was performed for the following variables: time, day, weather, road surface condition, crash type, crash location, number of lanes, road type, car movement, crash type (general crash, rear end side collision), and severity.

Table 3 presents the injuries and deaths involved in crashes. Clearly, the number of male persons related to crashes was about three times of that of females and about a half of the injuries and deaths were drivers, which also confirmed the importance of educating the safe driving program. Moreover, the percentage of motorcyclists was up to 15.79%, an alarming number among all the injuries and deaths. It was also noticed that 70.77% of injuries and deaths involved persons aging 25~60, especially those between 31~40, a group with higher driving risk. However, we found that the crash involvement rate of younger and older drivers was significant higher than that of medium aged ones.

Table 3
Injury and death categories involved in crashes

Persons	Male	Female	Driver	Passenger	Pedestrian	Motorcycler	Bicycler
Percent/%	74.86	25.14	46.17	13.43	13.45	15.79	11.16
Age	< 16	16-24	25-30	31-40	41-50	51-60	> 61
Percent/%	9.34	11.54	15.93	18.89	19.24	16.71	8.35

Table 4 shows the effects of weather and brightness on crash occurrence. Obviously, the weather has a significant effect on crash occurrence and more than two third of observations occurred on bad weather days including cloudy, rainy, foggy, snowy, and heavily windy ones. Brightness also affects the frequency of crashes through affecting driving behavior of drivers, and we can see that as high as 17.94% of crashes during darkness, with street lamps.

Table 4
Crash distribution by weather and bright effect

Weather	Sunny	Cloudy	Rainy	Foggy	Snowy	Heavy windy
Percent/%	31.73	13.79	17.38	11.17	17.48	8.45
Brightness	Dawn	Daylight	Twilight	Dark with street lamp	Dark without street lamp	Entrance and exit to urban tunnel
Percent/%	1.33	63.74	6.15	17.94	9.97	0.87

3.3 Location Specification

Crash distributions according to road type are shown in Table 5. It can be noticed that 33.46% of crashes were in arterial roads. Cumulative percentage shows that 45.09% of crashes occurred on expressways and arterials. This is because of high speeds and high traffic volumes on these two types of roads, and more frequent entrances and exits on such type of roads also contributed to the occurrence of crashes. The second highest incidence rate was in the sub-arterials category, with 19.72% of crashes occurring on them, due partially to the lack of effective safety facilities. Significantly, unsignalized intersection is a typical black spot prone to traffic crashes.

Table 5
Crash distribution by road type and crash location

Road Type	**Percent/%**	**Position**	**Percent/%**
Expressway	11.63	Roundabout	2.49
Arterial	33.46	Bridge	0.65
Sub-arterial	19.72	Tunnel	0.87
Branch and minor roads	10.12	Unsignalized intersection	12.43
Residential Street	2.36	Signalized intersection	6.27

Table 6 shows the cross tabulation of crash types and districts. It can be noticed from the results that car collision which had the highest percentage of crash occurrence had its highest percentage of 21.8% in Beilin district. This may be attributed to the high traffic volume and low speeds in these districts. The other types are noticed to have their highest percentage of occurrence in Weiyang and Baqiao. Unfortunately, truck-involved crashes made up a large percentage in the crash reports, and this is particularly true for outlying or peripheral areas [17]. Roadside barriers, light poles, even trees and other stationary objects sometimes are prone to cause serious secondary collisions and severe injuries.

Table 6
Crash percentage in types on districts

Crash type	Yanta	Beilin	Lianhu	Weiyang	Xincheng	Baqiao
Car collision	18.8	21.8	19	12.5	14.5	13.4
Collision into stationary objects	12.6	12.6	12.1	23.5	19.8	19.4
Collision into pedestrians	14.7	19.5	17.1	13.9	21.7	13.1
Overturning	18.9	15.1	22.5	13.3	11.9	18.3
Falling down	17.3	11.5	16.1	22.9	9.5	22.7
Others	17.7	19.5	13.2	13.9	22.6	13.1

Stationary objects includes light pole, trees, wall and street furniture, etc.

3.4 Potential Causes

As a result of reviewing the crash reports, crashes caused by sudden stop were observed statistically as the most frequent types, responsible for 32.3% of the total records. These usually occurred in traffic congestions where vehicles were moving slowly and traffic was mainly in stop-and-go situations, and drivers usually did not keep enough space between themselves and other vehicles [18]. It also happened when a driver at a relatively high speed was surprised with congestion ahead.

Actually, traffic safety performance function (SPF), has a statistical relationship with congestion on urban freeways and observed safety, measured in the number of crashes over a unit of given time [crashes per kilometer per year (CPKPY)] and it varies with traffic exposure [14, 19], measured in annual average daily traffic (AADT). Six years of data were used to analyze such effects on selected multi-type urban road segments. Figure 1 presents a pattern of total observed CPKPYs as the AADT increases for expressway and arterial segments. It should be pointed out that such a function for branch and minor streets was neglected due to a lack of crash records.

In Figure 1, traffic density at 34,000 AADT is a critical point and can be viewed as a critical density (point A), beyond which notably higher crash rates are observed with AADT changes, and the portion of the left of this critical density can thus be considered as a sub-critical zone, as AADT increasing from 8,000 to 34,000 induces the increase of traffic density by 3.5 times [from 5.7pcu per kilometer per lane to 25.8 pcu/(km·ln)], while traveling speed remains almost the same (62 km/h to 55 km/h). The portion to the right of point B can be viewed as a super-critical zone and the portion between sub-critical and super-critical densities can be viewed as a transitional zone, featuring an increase in CPKPY from 21 to 52, compared with 8-21 from point C to point A. Further examination of SPF reflects that passing an AADT of 34,000, the number of crashes increases at a much faster rate with an increase in AADT.

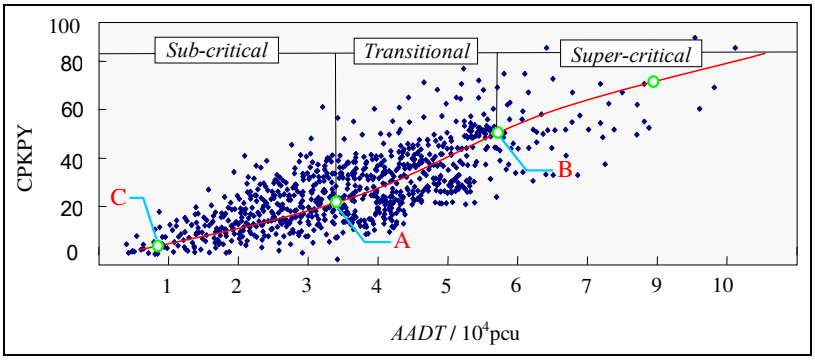


Figure 1

Statistical changes in segment crashes with AADT

A: critical density, with v = 55 km/h and ρ = 25.8 pcu/(km·ln); B: super-critical density, with v < 45km/h and ρ > 52.8 pcu/(km·ln); C: v = 62 km/h and ρ = 5.7 pcu/(km·ln)

In actuality, the crash rate changes with traffic volume, and SPF reflects how these changes take place. Lower rates within equal SPF mean higher safety rather than higher rates. However, numerous research reports have shown that predictive models that use traffic volume as the only explanatory variable may not adequately characterize the crash process on freeway segments [19, 20]. Functional forms that incorporate density and *v/c* ratio offer a richer description of crashes occurring on these facilities locating in a rural or urban environment [21].

Figure 2 shows the relation between crash rates measured in crashes per 100 million vehicle kilometers travelled and *v/c* ratio. Obviously, the operation ranging 0.5~0.6 under LOS C meets the lowest level of crash rate and we yield the lowest 45 crashes with respective to moderate *v/c* (0.54). Such a V regression mode also shows that free traffic flow usually brings serious rear-end or other types of crashes due to drivers' low vigilance of danger. Furthermore, more congested flow under LOS D or worse improves the likelihood of minor crashes.

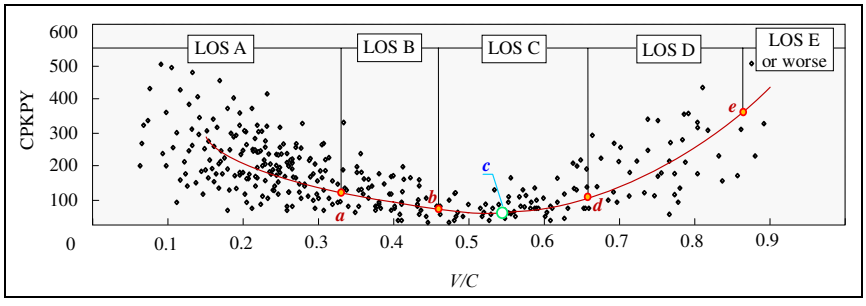


Figure 2

Statistical relation of crash rate and *v/c*

a. (0.32, 120); *b.* (0.46, 75); *c.* (0.54, 45); *d.* (0.70, 105); *e.* (0.87, 370)

Nevertheless, separate predictive models for single- and multi-vehicle crashes should be developed rather than one common model for all crash types [22]. Different variables (i.e., type of roads, barriers, etc.), for example, all have varying effects on the occurrence of crashes and statistical rates. Individual analysis may help find potential causes and effective measures to deal with the particular locations or long term safety improvement programs.

Conclusions

The presented study is designed to be a first step towards analyzing traffic crashes in Xi'an city. Crash features related to time, position, type of roads and type of crashes, etc., and exposure effects involving AADT and *v/c* analyses provide many important types of information that are required by engineers and decision-makers to improve safety for the driving public in this western metropolis of China. Such analyses must be performed regularly (annually) for long-term traffic safety environment consideration [15].

During the review process of crash reports, some recommendations were suggested. Road geometry in Xi'an should be checked on the national level and improvements should be made at many locations that are not up to standards, for example improving sight distances at intersections and roundabouts; checking the appropriateness of (and redesign if necessary) the acceleration and deceleration lanes to and from arterials and sub-arterials [23], etc. Moreover, it is also recommended to use traffic signals or warning signs before entering the intersections and to provide enough pedestrian crossings and overpasses, especially at heavy traffic intersections and commercial regions [24, 25]. Sufficient traffic safety education is also necessary not only for drivers but also for voluntary participants (i.e., pedestrians, bicyclists, et al). For scientific research demand, crash messages need to be more accurately identified in order to construct more accurate crash database using available GPS technology [26, 27].

Acknowledgement

This research is supported by the China Postdoctoral Science Foundation (No. 2011M501434). Here the helps from our industry partner, Xi'an General Traffic Policy Teams, the editors of Acta Polytechnica Hungarica, anonymous reviewers and authors of cited papers are gratefully acknowledged. We would like to thank, in particular, Prof. K. M. Chen from Chang'an University, for suggesting this research problem and the valuable advice he has given.

References

[1] D. D. Clarke, P. Ward, C. Bartle, et al, "Killer Crashes: Fatal Road Traffic Crashes in the UK", Accident Analysis and Prevention, 42(2), 2010, pp. 764-770

[2] H. L. Huang, H. C. Chin, "Disaggregate Propensity Study on Red Light Running Crashes Using Quasi-induced Exposure Method", Journal of Transportation Engineering, 135(3), 2009, pp. 104-111

[3] S. Mitra, "Spatial Autocorrelation and Bayesian Spatial Statistical Method for Analyzing Intersections Prone to Injury Crashes", Transportation Research Record, 2136, 2009, pp. 92-100

[4] B. Falk, H. Montgomery, "Developing Traffic Safety Interventions from Conceptions of Risks and Accidents", Transportation Research Part F: Traffic Psychology and Behaviour, 10(5), 2007, pp. 414-427

[5] M. A. Quddus, C. Wang, S. G. Ison, "Road Traffic Congestion and Crash Severity: Econometric Analysis Using Ordered Response Models", Journal of Transportation Engineering, 136(5), 2010, pp. 424-435

[6] A. Gan, K. Y. Liu, L. D. Shen, et al, "Prototype Information System for Estimating Average Vehicle Occupancies from Traffic Accident Records", Transportation Research Record, 2049, 2008, pp. 29-37

[7] C. C. Sun, V. Chilukuri, "Dynamic Incident Progression Curve for Classifying Secondary Traffic Crashes", Journal of Transportation Engineering, 136(12), 2010, pp. 1153-1158

[8] C. Y. Chan, B. S. Huang, X. D. Yan, et al, "Investigating Effects of Asphalt Pavement Conditions on Traffic Accidents in Tennessee based on the Pavement Management System (PMS)", Journal of Advanced Transportation, 44(3), 2010, pp. 150-161

[9] X. S. Wang, M. Abdel-Aty, P. A. Brady, "Crash Estimation at Signalized Intersections Significant Factors and Temporal Effect", Transportation Research Record, 1953, 2006, pp. 10-20

[10] D. Szoke, J. Logo, D. B. Merczel, "Optimal Suspension Settings for Ride Comfort of Road Vehicles", Periodica Polytechnica - Civil Engineering, 54(2), 2010, pp. 73-78

[11] X. Ye, R. M. Pendyala, S. P. Washington, et al, "A Simultaneous Equations Model of Crash Frequency by Collision Type for Rural Intersections", Safety Science, 47(3), 2009, pp. 443-452

[12] M. Abdel-Aty, X. S. Wang, J. B. Santos, "Identifying Intersection-related Traffic Crashes for Accurate Safety Representation", ITE Journal-Institute of Transportation Engineers, 79(12), 2009, pp. 38-44

[13] A. Bener, "Emerging Trend in Motorisation and the Epidemic of Road Traffic Crashes in an Economically Growing Country", International Journal of Crashworthiness, 14(2), 2009, pp. 183-188

[14] Y. G. Wang, K. M. Chen, Y. L. Pei, et al, "Integrating Before and After Crash Features into Measuring the Effectiveness of Intersection Safety Improvement Project in Harbin", Transport, 26(1), 2011, pp. 112-121

[15] H. Zhu, K. K. Dixon, S. Washington, et al, "Predicting Single-Vehicle Fatal Crashes for Two-Lane Rural Highways in Southeastern United States", Transportation Research Record, 2147, 2010, pp. 88-96

[16] X. P. Yan, M. Ma, Ming, H. L. Huang, et al, "Motor Vehicle-Bicycle Crashes in Beijing: Irregular Maneuvers, Crash Patterns, and Injury Severity", Accident Analysis and Prevention, 43(5), 2011, pp. 1751-1758

[17] K. Ratkeviciute, "Model for the Substantiation of Road Safety Improvement Measures on the Roads of Lithuania", The Baltic Journal of Road and Bridge Engineering, 5(2), 2010, pp. 116-123

[18] I. Fi, J. Galuska, "Recommendations for New Capacity Values on Freeways", Periodic Polytechnica - Civil Engineering, 54(2), 2010, pp. 127-136

[19] J. Kononov, B. Bailey, B. K. Allery, "Relationships between Safety and Both Congestion and Number of Lanes on Urban Freeways", Transportation Research Record, 2083, 2009, pp. 26-39

[20] D. P. G. de Wet, "WIM Calibration and Data Quality Management", Journal of the South African Institution of Civil Engineering, 52(2), 2010, pp. 70-76

[21] L. D. Zhong, X. D. Sun, Y. S. Chen, et al, "Research on the Relationship between V/C and Crash Rate on Freeway", Journal of Beijing University of Technology, 33(1), 2007, pp. 37-40

[22] S. R. Geedipally, D. Lord, "Investigating the Effect of Modeling Single-Vehicle and Multi-Vehicle Crashes Separately on Confidence Intervals of Poisson-Gamma Models", Accident Analysis and Prevention, 42(4), 2010, pp. 1273-1282

[23] A. Pande, M. Abdel-Aty, A. Das, "A Classification Tree-based Modeling Approach for Segment-related Crashes on Multilane Highways", Journal of Safety Research, 41(5), 2010, pp. 391-397

[24] O. Prentkovskis, E. Sokolovskij, V. Bartulis, "Investigating Traffic Accidents: a Collision of Two Motor Vehicles", Transport, 25(2), 2010, pp. 105-115

[25] M. Rajsman, G. D. Lisicin, "Influence of Legislation on Road Traffic Safety", Promet – Traffic & Transportation, 22(1), 2010, pp. 75-82

[26] M. C. Sanjs, C. Petar, "Skewness and Kurtosis in Function of Selection of Network Traffic Distribution", Acta Polytechnica Hungarica, 7(2), 2010, pp. 95-106

[27] S. S. Durduran, "A Decision Making System to Automatic Recognize of Traffic Accidents on the Basis of a GIS Platform", Expert Systems with Applications, 37(12), 2010, pp. 7729-7736