Defining Infocommunications and Related Terms

Gyula Sallai

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Magyar tudósok krt. 2, H-1117 Budapest, Hungary
e-mail: sallai@tmit.bme.hu

Abstract: The convergence of telecommunications, informatics (IT) and media based on the common digital technology affected these sectors in different ways and at different times, resulting in different approaches and terms. The paper presents infocommunications as an evolving expansion of telecommunications with information processing and content management functions. Information and communication(s) technology (ICT) is considered an extended synonym for IT to emphasis the integration of the unified (tele)communications. Contents involved in the convergence process are expanded by cognitive features; and the term cognitive infocommunications was established. Subtle distinctions between these and other terms that have emerged are clarified and compared, as well as a consistent terminology is proposed.

Keywords: infocommunications; information and communication technologies; ICT; information society technologies; electronic communications; media informatics; cognitive infocommunications; digital convergence; TIM sector; Digital Ecosystem

1 Introduction

People refer to the mosaic words infocommunications, info-communications and the acronyms ICT, IST, TIM and other terms to identify a sector born as a result of digital convergence, a convergence process triggered by the huge scale development of digital technology. The role of the convergent sector is relevant in the realization of the information/knowledge/networked society. This sector, more precisely the ICT sector, represents about 5% of the EU economy, but it generates 25% of total business expenditure in Research and Development (R&D), and investments in this sector account for 50% of all European productivity growth [7]. The uniform digital technology base has given rise not only to the effectiveness of the economies of scale and the efficient increase of the complexity of the products and services, but has also provided additional synergic opportunities for the combination of the functions. Therefore the sector is not only

an integration of previously independent sectors, but it is dynamically expanded by new products and services, and thereby its scope is continuously changing and expanding. Different terms are formed for the same entity from different aspects; the same term is often used in many different contexts. When we use a term, it is a shortcut that allows us to refer to an entity without having to repeatedly describe it or precisely define it. The selected term in itself does not define content; it is created institutionally, or emerges from usage within a specific context which is based on previous meaning or meanings and in turn influences its interpretation. These issues in the respect of the information society versus the knowledge society are clearly presented in [3], while some aspects in relation to the abovementioned terms are discussed in [2, 13, 14, 30]. Institutions and organizations usually use these terms with their general meaning without precise definition, but rather with describing characteristic elements of their activity [11, 17, 20]. The detailed definition, based on an international standard classification of activities, was considered in order to obtain sector indicators [26].

The information technology and telecommunications sectors were first affected by digital convergence process, which was manifested in the unification of their technologies, in the integration of their markets and in the harmonization of their regulation. The convergence process affected the two sectors in different ways; different approaches, models and terms were formulated on both sides, and the same term is often used in different contexts. Then the electronic media and content producing sectors also entered the convergence process, which resulted in the birth of the first real convergent sector. The term differences were inherited. The process of the convergence has been further extended; synergic combinations with the cognitive science and other content and application areas have been realized. For the deployment of the synergies the market structure of the concerned sectors is reconsidered, and the regulation of the converging areas is harmonized.

In the body of the paper, first we present the digital convergence process and terms from the telecommunications approach, and then from the aspect of the information technology. We try to clarify the subtle distinctions and build up a consistent terminology using a series of layer models and a colour mixing scheme.

2 Telecommunications, Electronic Communications, Infocommunications

Digital technology has radically reshaped telecommunications. From the point of view of telecommunications, four main overlapping phases of the digital convergence process have been identified [30]. To present the evolutionary phases, we use a simple value chain model which represents the consequent value-generating functions from information sources to the usage at the destination (Figure 1).

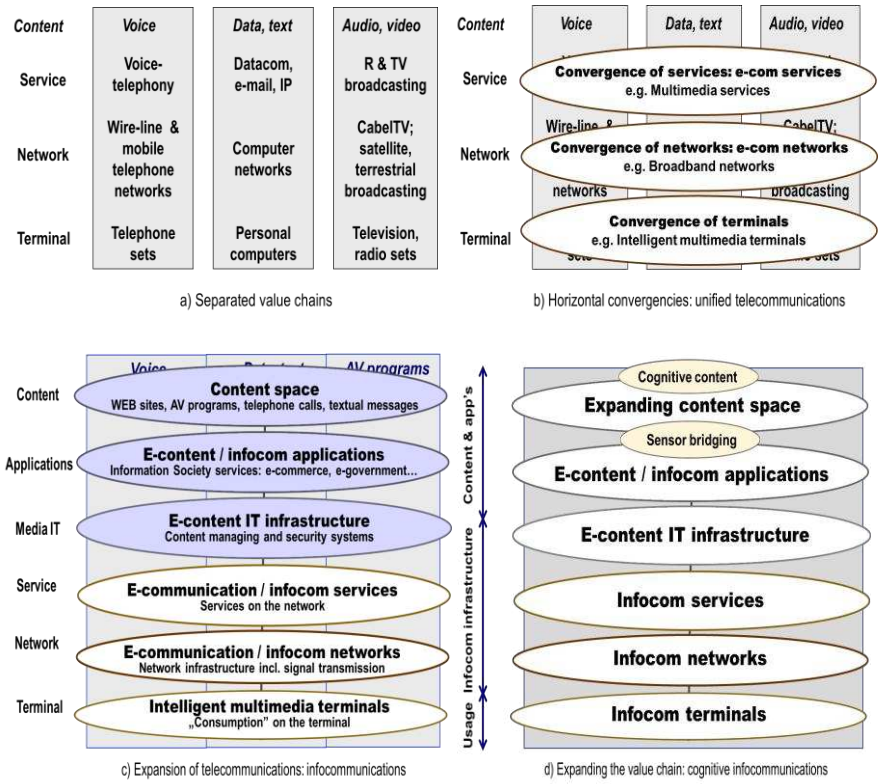


Figure 1
Phases of digital convergence

1) *Separated internal digitization.* Traditionally, the various contents have been associated with separated networks, services and user terminals (Figure 1a), and their markets and regulation have been separately managed. Voice has been managed by telephony (voice communications), data and text by data communications, audio-visual programs by radio & television broadcasting and distribution (media communications). These separated sectors had their own specific technology; however the use of digital technology for the various functions has been intensively introduced. The digitization of telephony started with the use of digital transmission, followed by the introduction of digital control and switching, which led to an integrated digital telephone network [19, 27]. Modern mobile telephony is already fully based on digital technology. Digital solutions have also penetrated into media communications.

2) *Unification of telecommunications: electronic communications*. Any kind of digitized information content can be transmitted through various digital networks and therefore the integrated realization of these networks is

reasonable [19, 25, 32]. A broadband IP-based network is equally able to transmit voice, data, text, audio-visual programs, multimedia etc. The combination of voice, data and audio-visual services offers new IP-based multimedia service opportunities. At the users, various integrative terminals appeared. The value chains of the voice, data and media communications have been merged; horizontal convergence and some integration of the services, networks and terminals can be identified, and a single value chain with horizontal layers can be shaped (Figure 1b) [16, 24, 27, 28]. A unified telecommunication sector has emerged, which is formally called electronic communications. In deploying these horizontal convergences, uniform regulation was introduced for electronic communications in the European Union [10].

3) *Expansion of telecommunications: infocommunications.* Telecommunications combined with some information processing and content handling functions on digital technology base are called infocommunications, or in short form, infocom(s) or infocomm(s). The term first emerged in the beginning of eighties at scientific conferences, then was gradually adopted in the 1990s by the players of telecommunications industry, both the manufacturers and the service providers, to clearly express their participation in the convergence process, and it was regularly used by the International Telecommunication Union (ITU) [18]. Electronic communications provide the bearing digital infrastructure for the digitalized content services and applications, whereby the digital convergence process has been naturally expanded to all information and media technology functions. The same digital message form is used in the computer industry for passing messages within and between the computers, together with the growing operational use of computers within telecommunications, resulting in a synthesis between the telecommunications sector and the computer-based information sector (info-telecom/info-com convergence). The IP-based solutions in both the computer and telecommunications industries integrated the isolated areas and generated an integrated structure for processing, storing, accessing and distributing information [33]. The electronic media and content production and management have also been involved into the convergence process (media convergence) by using the opportunities delivered by IP-based telecommunications and computer communications, which is demonstrated by the immense spread of the Internet provided by telecommunications and Internet service providers [8, 13]. Electronic content services and applications based on web technologies and delivered by electronic communication networks and services have emerged, e.g. e-business, e-commerce, e-health, e-learning, e-government, smart home, office and cities, and intelligent transportation and energy systems. In general they can be called e-content or infocommunication applications; the terms information society's services, content services and e-services are also used [6, 15, 16, 28]. The value chain of infocommunications shown in Figure 1c contains three additional layers:

- the layer of content space, symbolising the jointly-managed information sources and the customer's payable demands;
- the layer of e-content or infocommunication applications, including from simple content services to the wide variety of secure and multi-content Internet services;
- the layer of the common IT infrastructure for applications (media IT, middleware layer), including common message handling, content management functions (e.g. directory assistance, editing, indexing), browsers, portals, search engines, security systems etc.

The layers of IT infrastructure, e-communication services and networks together can be considered the infocommunication infrastructure [30]. Similar layer models of infocommunications embracing the functions provided by the Internet technology have been shown and discussed to allocate the players of infocommunications to layers [13, 22, 23].

4) *Expanding content space: cognitive infocommunications.* Traditionally the sensory information managed has been limited to sight and hearing, but the content space can be expanded to all senses, including touch, smell or any other modality, in general human emotions and feeling, as well as gestures in 3D space. The sensory information experienced is transferred to the destination and transformed to an appropriate sensory modality in a way that the user can process it effectively. Cognitive infocommunications (CogInfoCom) defined in [2] combines infocommunications and cognitive science and expands the content space with cognitive and sensory contents [4, 5]. Thereby in the value chain the content layer is expanded and the applications layer involves the bridging of sensory information to a more applicable one, if necessary (Figure 1d) [30].

Recently, the term infocommunications as expanded telecommunications in the abovementioned meaning has generally been used by telecommunications manufacturers, service providers and regulatory authorities, in scientific papers and university curriculums, and in the name of scientific and professional conferences and journals (e.g. IEEE Infocom, Infocommunications Journal). The term is also used in politics in a wider sense as a shorter form of information and communications technology.

3 ICT, Information and Communication(s) Technologies

Information and Communications Technology, usually abbreviated as ICT, has been in use from the second half of the nineties [31] and is used as an extended synonym for information technology (IT) to emphasis the role of unified

(tele)communications, the integration of telecommunications with computers, as along with the necessary software, middleware, storage and audio-visual systems that enable users to create, access, store, transmit and manipulate information. In other words, ICT consists of IT as well as telecommunications, broadcast media, all types of audio and video processing and transmission, and network based control and monitoring functions [9, 12, 20].

At present, the term ICT is generally used and usually refers to *the integration of information and telecommunication technology sectors involving their convergence with the media technology sector* based on common digital technology. ICT includes all types of telecommunication and broadcasting systems and services (wireline, wireless, mobile, satellite), computer hardware, software, networks and services, content producing and managing multimedia systems, Internet technologies, services and applications, machine-to-machine applications, etc. The term *Information Society Technologies* (IST) was generally used to the integration of telecommunications, IT and media sectors in the *EU's research, technological development and demonstration framework programmes* (FP5 and FP6) between 1998 and 2007 [1]. However, since 2007 in FP7 and Horizon 2020 the term ICT has been used. The EU's Horizon 2020 programme (2014-2020) will preferably support the ICT research and innovation, in particularly the development of [7]:

- next generation computing, advanced computing systems and technologies;
- infrastructures, technologies and services for the Future Internet;
- content technologies and information management, including ICT for digital content and creativity;
- advanced interfaces, robotics and smart spaces;
- a new generation of components and systems including nano-electronics and photonics technologies and embedded systems.

On the usage of the term ICT, some additional remarks should be mentioned:

- Originally, only "*information and communications technology*" (with communications in the plural) was considered correct since ICT refers to communications (in the sense of a technology of sending and receiving information), not communication (the act of sending or receiving information by speaking, writing, phoning, emailing, etc.). Nevertheless, recently, the single form "*information and communication technology*" is becoming increasingly common and is now used in about half the books, and it is also used by the ITU [20]. In order to express these dual forms, we use "communication(s)". Sometimes the acronym ICT stands for a wider interpretation: "information, communication and technology", which includes "information" and "communication" themselves as well as "information and communication technology" [14].

- The *International Telecommunication Union* (ITU) and the *European Telecommunications Standards Institute* (ETSI) – according to their own definition – deal with not only telecommunications, but also ICT issues; however, their relevant activity, their recommendations and standards resp. are focused on network-centric issues [11, 20]. Some characteristic citations are from their web-sites: "ITU is committed to connecting all the world's people. …We allocate global radio spectrum and satellite orbits, develop the technical standards that ensure networks and technologies seamlessly interconnect, and strive to improve access to ICTs to underserved communities worldwide." "ETSI produces globally-applicable standards for ICT, including fixed, mobile, radio, converged, broadcast and internet technologies. …ETSI is the recognized regional standards body dealing with telecommunications, broadcasting and other electronic communications networks and services".
- The member countries of the *Organisation for Economic Cooperation and Development* (OECD) agreed in 1998 to define the ICT sector as a combination of manufacturing and services industries that capture, transmit and display data and information electronically. In 2011, the OECD set a number of standards for measuring and comparing the information society across countries [26], including definitions and classifications of ICT as well as Content & Media products as Information Economy products (goods and services). Accordingly, ICT products must primarily be intended to fulfill or enable the function of information processing and communication by electronic means, including transmission and display; contents published in mass communication media such as printed, audio-visual and online contents and related services are not considered ICT products.
- The terms info-com(s) and info-communications (with a hyphen) are used to express the integration of the IT and (tele)communication sectors [23, 33], or simply to interpret the abbreviation ICT.

Over the past several years, the terms *Digital World* (DigiWorld) and *Digital Ecosystem* have emerged to embrace all those sectors that are already or on the verge of being based on digital technologies. As well, *the abbreviations TIM,* as the Telecom IT/Internet Media sector *or TIME,* as the Telecom IT/Internet Media & Entertainment/Edutainment sector, are used to express the full integration of these sectors and to enhance the significance of content respectively [17, 21, 34]. The Digital Ecosystem is defined by the World Economic Forum as the space formed by the convergence of the media, telecommunications and IT sectors, and consists of users, companies, government and civil society, as well as the infrastructure that enables digital interactions [34].

4 Comparison of the Terms

The convergence of telecommunication, information and media technologies using common digital technology has resulted in an integrated sector that achieves the functions of content and information management and communication by electronic means, including processing, handling, transmission and display. The integrated sector has aptly been called *IST (Information Society Technology),* or currently the *TIM (Telecom, IT and Media)* sector. The term Digital Ecosystem has a broader meaning; *Digital Technology sector* seems to be appropriate. Sometimes the ICT sector also has this meaning in its widest interpretation, but it cannot be preferred. Henceforth, we use the term TIM.

The *Information and Communication(s) Technology* (ICT) sector's products fulfill the function of information processing and communication by electronic means, including transmission and display. Recently, the usage of ICT for digital content management has also been included in the term.

Infocommunications (Infocom) is the natural expansion of telecommunications with information processing and content handling functions including all types of electronic communications (fixed and mobile telephony, datacommunications, mediacommunications, broadcasting, etc.) on a digital technology base, mainly through Internet technology.

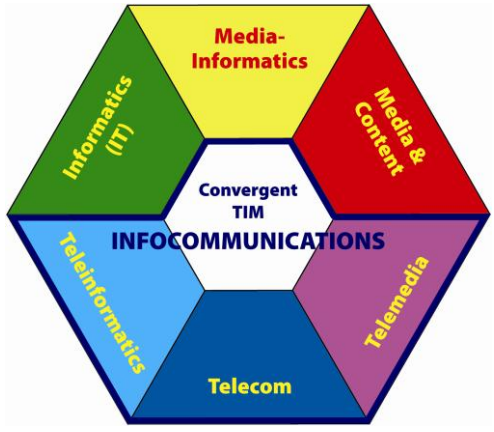


Figure 2
Digital convergence prism: positioning Infocommunications

The relationship and position of the terms is presented by a *digital convergence prism* (Figure 2) [29], which shows the three components (T, I, M) and their pairs and the triple combination (convergent TIM triplet) according to the rule of additive colour mixing. Assuming that telecommunications (Telecom) is blue, informatics (IT) is green and Media & Content is red, then teleinformatics/telematics is cyan, telemedia/networked media is magenta, media informatics is yellow, and the convergent TIM is white. In such a way, the

integrated TIM sector corresponds to the prism as a whole, the ICT sector (in wider, but not the widest sense) to the whole minus the red area (Media & Content), and the Infocom sector relates to telecommunications and neighbouring three areas (blue, cyan, magenta and white) [29]. That means that, for example, media informatics is a part of ICT but not part of Infocom.

Conclusion

The widespread deployment of mobile communications and the Internet in the 1990s accelerated the convergence process based on common digital technology. Telecommunications and the Internet formed more and more an integrated system for processing, storing, accessing and distributing information. Telecommunications was unified and significantly expanded and was referred to as Infocommunications. Later on, the term Information and Communication(s) Technology (ICT) was also used as an extended synonym for information technology (IT) to emphasis the integration of IT and (tele)communications. Recently, the convergence of telecommunication, information and media technologies has deployed, frequently keeping the term ICT; but the terms TIM or Digital Technology sector seem to be more pertinent. In the future, the convergence process will widen further, mainly via the expansion of the managed content space. The relative position of the different terms is shown by layer models and an additive colour mixing scheme to build up a consistent terminology.

References

[1] Arend, M.: SEAMATE: Socio-Economic Analysis and Macro-modeling of Adapting to Information Technology in Europe. Cambridge Econometrics, Econcept AG, Information Society Technologies (IST-2000-31104), June 2002

[2] Baranyi P., Csapó A.: Definition and Synergies of Cognitive Infocommunications. Acta Polytechnica Hungarica, ISSN 1785-8860, Vol. 9, No. 1, 2012, pp. 67-83

[3] Burch, S.: The information society / The knowledge society. Chapter in Word Matters. C & F Editions, November 2005

[4] CogInfoCom 2010 (1st International Conference on Cognitive Infocommunications), 29 Nov. - 1 Dec. 2010, Tokyo, Japan

[5] CogInfoCom 2011 (2nd International Conference on Cognitive Infocommunications), 7-9 July 2011, Budapest, Hungary, E-ISBN: 978-963-8111-78-4, Print-ISBN: 978-1-4577-1806-9

[6] Commission of European Communities: Green Paper on the Convergence of the Telecommunications, Media and Information Technology Sectors, and Implications for Regulation. Towards an Information Society Approach. 3 Dec. 1997, COM (1997) 623

[7] Commission of European Communities: Information / Communications Technologies (ICT) in Horizon 2020, Brussels, November 2011 http://ec.europa.eu/research/horizon2020/index_en.cfm

[8] Dominigue, J. at al. (ed): The Future Internet - Future Internet Assembly 2011: Achievements and Technological Promises, 17-19 May 2011, Budapest, ISBN 978-3-642-20898-0, 2011, Springer, Heidelberg (Lectures Notes in Computer Science 6656)

[9] EITO: European Information Technology Observatory, Yearbooks from 2001 to 2009, ISSN 097-4862

[10] EU legislation: Regulatory framework for electronic communications. http://europa.eu/legislation_summaries/information_society/legislative_framework/l24216a_en.htm

[11] European Telecommunications Standards Institute: About ETSI, 2011 http://www.etsi.org/website/aboutetsi/aboutetsi.aspx

[12] FOLDOC: Information and Communication Technology

http://foldoc.org/Information+and+Communication+Technology

[13] Fransman, M.: Mapping the Evolving Telecom Industry: The Uses and Shortcomings of the Layer Model. Telecommunication Policy, Vol. 26, 2002

[14] Giles, J.: What is ICT? Michalsons, June 2009 http://www.michalsons.co.za/what-is-ict/2525

[15] Henten, A.: Convergence, Synergies and Media Power. ITU Policy and Regulatory Summit, Geneva, 1999. http://www.itu.int

[16] Henten A., Samarajiva R., Melody W. H.: Designing Next Generation Telecom Regulation: ICT Convergence or Multisector Utility? Lirne.net. Report on the WDR Dialogue Theme, 2003, www.regulateonline.org

[17] IDATE: DigiWorld Yearbook 2009, ISBN: 978-2-84822-143-4

[18] International Telecommunication Union (ITU): World Communications - Going global with a networked society. Editor: G. L. Franco Novara, Italy, 1991

[19] International Telecommunication Union: Convergence and Regulation, Volume of Trends in Telecommunication Reform, 1999, Geneva

[20] International Telecommunication Union: Measuring the Information Society: The ICT Development Index. 2009, p. 108, ISBN 92-61-12831-9

[21] Keyrus group: Telecom, IT & Media. http://www.keyrus.be/keyrus/industries-/telecom,-it-&-media/telecom,-it-&-media/id/67722

[22] Krafft, J.: Vertical Structure of the Industry and Competition: an Analysis of the Evolution of the Info-Communications Industry, Telecommunication Policy, Vol. 27, pp. 625-649, 2003

[23] Krafft, J.: Profiting in the Info-Coms Industry in the Age of Broadband: Lessons and New Considerations. Technological Forecasting & Social Change, Vol. 77, pp. 265-278, 2010

[24] Melody, W. H.: Telecom Reform: Progress and Prospects. Telecommunications Policy, Vol. 23, No. 1, pp. 7-34, 1999

[25] Organization for Economic Cooperation and Development (OECD): Telecommunications and Broadcasting: Convergence or Collision? 1992, Paris

[26] OECD: Guide to Measuring the Information Society 2011, p. 206, ISBN 978-92-64-09598-4, August 2011

[27] Saito, T.: An Evolving Scenario of Communication Network towards B-ISDN, in V. B. Iversen (ed.): Integrated Broadband Communication Networks and Services. North-Holland, 1994

[28] Sallai Gy.: Converging Information, Communication and Media Technologies. Chapter in the book entitled Assessing Societal Implications of Converging Technological Development. pp. 25-43, Ed.: G. Banse, A. Grunwald, I. Hronszky, G. Nelson. Sigma, Berlin, 2007

[29] Sallai Gy., Abos I., Kósa Zs., Magyar G.: Dimensions of Infocommunication Convergence. Híradástechnika (in Hungarian), Vol. 64, Special issue, pp. 17-22, 2009

[30] Sallai, Gy: The Cradle of the Cognitive Infocommunications. Acta Polytechnica Hungarica, ISSN 1785-8860, Vol. 9, No. 1, 2012, pp. 171-181

[31] Stevenson, D.: Information and Communications Technology in UK Schools, an Independent Inquiry. The Independent ICT in Schools Commission. London, UK, 1997

[32] Telecommunications Policy: Special Issue on “Competition and Convergence”, Vol. 18, No. 8, 1994

[33] Valtonen, T. P.: Governmental Visions for Future Info-Communication. TUCS Technical Report, No. 425, Turku Centre for Computer Sciences. May 2001

[34] World Economic Forum: Digital Ecosystem - Convergence between IT, Telecoms, Media and Entertainment: Scenarios to 2015. World Scenario Series, 2007

http://www3.weforum.org/docs/WEF_DigitalEcosystem_Scenario2015_ExecutiveSummary_2010.pdf

Wireless Sensor Network (WSN) Control for Indoor Temperature Monitoring

Yi-Jen Mon[1*], Chih-Min Lin[2], Imre J. Rudas[3]

[1*] Department of Computer Science and Information Engineering, Taoyuan Innovation Institute of Technology, Chung-Li, Taoyuan, 320, Taiwan, e-mail: monbuy@tiit.edu.tw

[2] Department of Electrical Engineering, Yuan Ze University, Chung-Li, Taoyuan, 320, Taiwan, e-mail: cml@saturn.yzu.edu.tw

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

Abstract: In this paper, a wireless sensor network (WSN) is constructed to carry out certain applications. This WSN is composed of a sensor, monitor, controller, etc. It has the benefits of low cost and low power consumption. A WSN can be used in many applications in a range of different control technologies, such as temperature monitoring. ZigBee is used to test the performance of the WSN. The experimental results reveal that the design requirement can be achieved; they also demonstrate that the WSN control methodology allows good performance of data transfer using a liquid crystal display (LCD) and motor control.

Keywords: wireless sensor network; WSN; ZigBee; temperature monitoring

1 Introduction

The purpose of this paper is to construct a wireless sensor network (WSN), comprising a sensor, monitor, controller unit, etc. Taking advantage of low cost and low power consumption, a variety of perceived control networks can be concatenated into a sensor network and can achieve a variety of control techniques. In the past, applications have been developed in many areas, such as home security, environmental monitoring, home/building automation, indoor location identification, etc. These achievements make human life more comfortable and convenient [1].

The 89C51 chip produced by Atmel Corporation is a single-chip processor comprising a CPU, memory, I/O and other useful integrated peripheral interfaces. It is also known as a micro-processor or micro-controller unit (MCU) [2]. This type of

MCU is widely used in industry, such as in home consumer electrical applications and in industrial control products. The MCU was developed in response to the need for small, cheap and low power systems [3-5].

The use of assembly language involves a certain degree of complexity and difficulty when it is used to implement a number of features; meanwhile, it is also difficult to use in a cross-platform system, and the written code is difficult to understand. Using C language instead has the benefits of easy understanding and maintenance of programs. In this paper, the Keil u-Vision2 software platform [6] is used to compile the developed high-level C language. It will then automatically generate machine code which is easier and simpler to burn into the MCU's program memory using the Simple type-A PGMSx IC WRITER [7, 8].

The universal asynchronous receiver/transmitter (UART) includes a start bit, 8-bit data bits, parity bits, and stop bits. When the UART has received data or characters, the execution of serial to parallel conversions will be completed. Then, the UART will put the serial bits into a serial buffer (SBUF) to do parallel transmission, a process called parallel to serial conversion. The MCU can read all of the data transmitted by the UART by using a PC Super Terminal to set and display the communication results. The experimental results show that all simple ASCII code can be successfully sent and received [6-9].

The wireless sensor network (WSN) is envisaged to monitor the environment for many years. A challenge is to reduce the WSN's energy consumption so as to extend its lifetime [10]. The ZigBee Alliance is an association of companies working together to develop standards (and products) for reliable, cost-effective, and low-power wireless networking. The ZigBee technology will probably be embedded in a wide range of products and applications across consumer, commercial, industrial and government markets worldwide [11]. ZigBee builds upon the IEEE 802.15.4 standard, which defines the physical and MAC layers for low cost, low rate personal area networks. ZigBee defines the network layer specifications for star, tree and peer-to-peer network topologies and provides a framework for application programming in the application layer. Route discovery in ZigBee is based on the well-known Ad Hoc On Demand Distance Vector routing algorithm (AODV). When a node needs a route to a certain destination, it broadcasts a route request (RREQ) message that propagates through the network until it reaches the destination. As it travels in the network, a RREQ message accumulates (in one of its fields) a forward cost value that is the sum of the costs of all the links it has traversed. The cost of a link can be set to a constant value or be dynamically calculated based on a link quality estimation provided by the IEEE 802.15.4 interface [12]. Wireless sensor networks are an emerging technology based on the progress of electrical and mechanical engineering, as well as computer science, in the last decade [13]. Mobile Ad Hoc networks allow autonomy and independence from 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 manage-

ment [14]. An Ad Hoc network is considered a very particular network, since it is a self-organizing network with no pre-deployed infrastructure and no centralized control; instead, nodes carry out basic networking functions such as routing. With this flexibility, Ad Hoc networks have the ability to be formed anywhere and at any time. In addition to traditional uses, such as for military battlefield applications, these networks are being increasingly used in everyday applications, such as in conferences, personal area networking and meetings [15]. Many routing protocols that are compatible with the characteristics of Ad Hoc networks have been proposed. In general, they can be divided into two main categories: topology-based and position-based. Topology-based routing protocols use information about links that exist in the network to perform packet forwarding. In general, topology-based routing protocols are considered not to scale in networks with more than several hundred nodes [16].

In this paper, the WSN is based on Ad Hoc structure, as aforementioned. The data transfer of liquid crystal display (LCD) and motor control are achieved by way of the MCU control methodology.

2 Introduction to WSN

The software development of the WSN is the most important issue. In this paper, the free software called Code::Blocks is used. This includes many application program interfaces (API). The Application Queue API provides a queue-based interface between an application and both the IEEE 802.15.4 stack and the hardware drivers (for the Jennic JN51xx wireless microcontroller):

• The API interacts with the IEEE 802.15.4 stack via the Jennic 802.15.4 Stack API (which sits on top of the 802.15.4 stack).

• The API interacts with the Peripheral Hardware Drivers via the Jennic Integrated Peripherals API (which sits on top of the Peripheral Hardware Drivers). This architecture is shown in Fig. 1. The Application Queue API handles interrupts coming from the MAC sub-layer of the IEEE 802.15.4 stack and from the integrated peripherals of the Jennic JN51xx wireless microcontroller, saving the application from dealing with interrupts directly.

The API implements a queue for each of three types of interrupt:

• Medium Access Control (MAC) Common Part Sub-layer (MCPS) interrupts coming from the stack. This is used for the MAC Data Services.

• MAC sub-Layer Management Entity (MLME) interrupts coming from the stack. This is used for the MAC Management Services.

• Hardware interrupts coming from the hardware drivers.

The prototype for the MCPS and MLME callbacks is a function that takes no parameters and returns void. The prototype for the hardware indications takes two 32-bit values as parameters and returns void. The application polls these queues for entries and then processes the entries [17-20].

A variety of network topologies are possible with IEEE 802.15.4. A network must consist of a minimum of two devices, of which one is the co-ordinator, referred to as the personal area network (PAN) co-ordinator. The possible network topologies are star topology, tree topology and mesh topology. The basic type of network topology is the star topology. A star topology consists of a central PAN co-ordinator surrounded by the other nodes of the network, often referred to as end devices. The tree network topology has an implicit structure based on parent-child relationships. Each node (except the PAN co-ordinator) has a parent. The node (including the PAN co-ordinator) may also (but not necessarily) have one or more children. Each node can communicate only with its parent and its children (if any). Any node which is a parent acts as a local co-ordinator for its children. In the mesh network topology, all devices can be identical (except that one must have the capability to act as the PAN co-ordinator) and are deployed in an ad hoc arrangement (with no particular network structure). Some (if not all) nodes can communicate directly. The nodes may not all be within range of each other, but a message can be passed from one node to another until it reaches its final destination.

A data transfer between network nodes can be unsolicited or the result of a request:

• When transferring data from a co-ordinator to a node, the node may not always be ready to receive data, since it may be in sleep mode for some of the time. In this case, responsibility may be given to the node to request data when it is able to receive. Therefore, the node polls the co-ordinator for data, and the co-ordinator then checks whether data is available and, if so, transmits a data frame. Acknowledgments may also be optionally implemented.

• When transferring data from a node to another node where reception is likely to be guaranteed (for example, from a node to a co-ordinator), it is usual to send a data frame directly (i.e., unsolicited). Again, acknowledgments may be optionally implemented. The data transfer methods are shown in Fig. 2 [17-20].

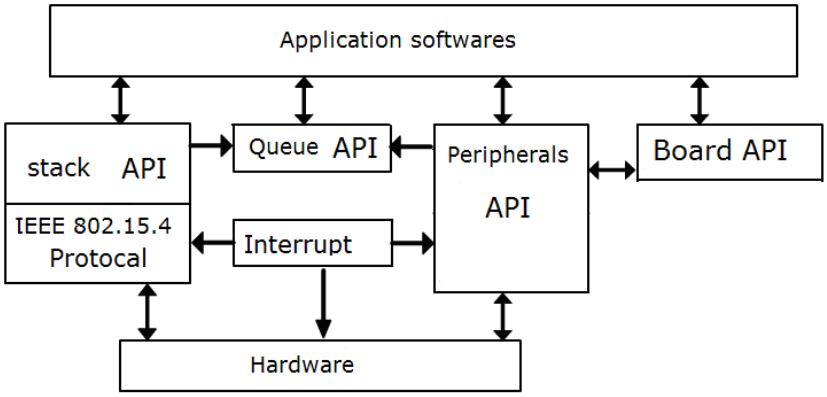


Figure 1
The architecture diagram of Jennic 802.15.4 API

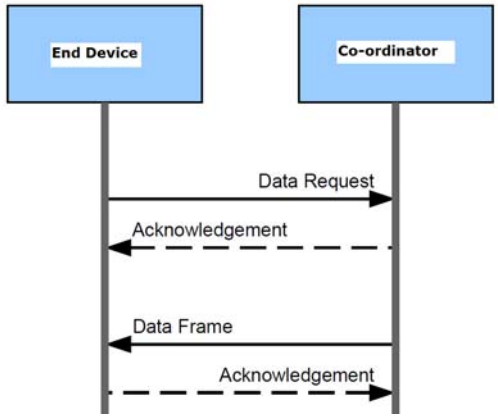


Figure 2
The diagram of data transfer methods of WSN

3 Experiment Result

The program is developed on the free software of Code::Blocks. First, the program for the co-ordinator and then the end device program are developed. Every network must have one and only one PAN co-ordinator, and one of the tasks in setting up a network is to select and initialize this co-ordinator. The network set-up process is shown in Fig. 3. The main co-ordinator and end device programs are shown in Fig. 4. The configure program diagram is shown in Fig. 5. The personal

area network identify (PAN-ID) must be set adequately, such as in line 64 of this program. The development board is produced by Fontal Technology Inc. This is a high power ZigBee Kit (FT-6200). It can provide all the software tools and hardware required to get first-hand experience with wireless sensor networks (WSN). The entry-level kits contain one base development board (BDB) and one sensor development board (SDB). Each board is equipped with a high-power IEEE 802.15.4 RF module based on JN-5121 CPU (produced by Jennic Technology Inc.), which provides a much higher covering range, using a 2.4 GHz RF antenna that has an IPEX connector for easy mechanical design, rather than the normal power RF module. For I/O expansion ports, it has 10 useful pins of GPIO including UART, ADC, DAC and Comparator. The sensor development board features temperature and humidity sensors [12]. The development board is shown in Fig. 6.

For the software, Jennic Technology Inc. provides free Application Programming Interface (API) software for the peripheral devices on the JN5121 and JN513x single-chip IEEE 802.15.4 compliant wireless microcontrollers. This is known as the Integrated Peripherals API. It details the calls that may be made through the API in order to set up, control and respond to events generated by the peripheral blocks, such as UART, GPIO lines and timers, among others. The software invoked by this API is present in the on-chip ROM. This API does not include support for the Zigbee WSN MAC hardware built into the device; this hardware is controlled using the MAC software stack that is built into the on-chip ROM [17-20].

ZigBee can be used with different sensors, such as in home automation, security management, industrial or environmental controls, and personal medical care. The design concept diagram is shown in Fig. 7. Using UART, the data can be presented in the LCD in different sensors. First, the LCD test is implemented as in Fig. 8. Then, as temperature monitoring is the experiment's main purpose, the temperature sensor on the end device will transmit data to the co-ordinator and then also appear in the LCD through UART. The real implementation of temperature monitoring in the laboratory is shown in Fig. 9. This shows a measured temperature of 26 °C. If the temperature is higher than this, the motor should start up to drive a fan to lower the temperature. In this experiment, a light-emitting diode (LED) is used to identify the signal of the starting motor. The WSN's control of temperature monitoring is successfully established and good motor control performance is also demonstrated.

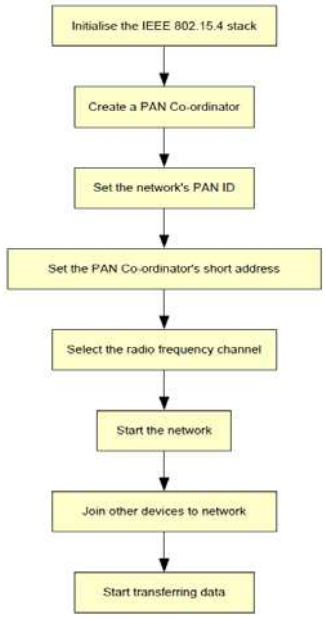


Figure 3
The diagram of network setup process

```
PUBLIC void AppColdStart(void)
{
   vWUART_Init();
   while(1)
   {
      vProcessEventQueues();
      switch (sCoordData.sSystem.eState)
      {
      case E_STATE_INIT:
         sCoordData.sSystem.u8Channel = CHANNEL_MIN;
         sCoordData.sSystem.eState =
E_STATE_START_ENERGY_SCAN;
         break;
      case E_STATE_START_ENERGY_SCAN:
         vStartEnergyScan();
         sCoordData.sSystem.eState =
E_STATE_ENERGY_SCANNING;
         break;
      case E_STATE_ENERGY_SCANNING:
         break;
      case E_STATE_START_COORDINATOR:
         vStartCoordinator();
         sCoordData.sSystem.eState =
E_STATE_RUNNING_UART_APP;
         break;
      case E_STATE_RUNNING_UART_APP:
         break;
      }
   }

}
```

Figure 4 (a)
The main program of co-ordinator

```
PUBLIC void AppColdStart(void)
{
   vWUART_Init();
   vStartActiveScan();
   while(1)
   {
      vProcessEventQueues();
   }
}
```

Figure 4 (b)
The main program of end device

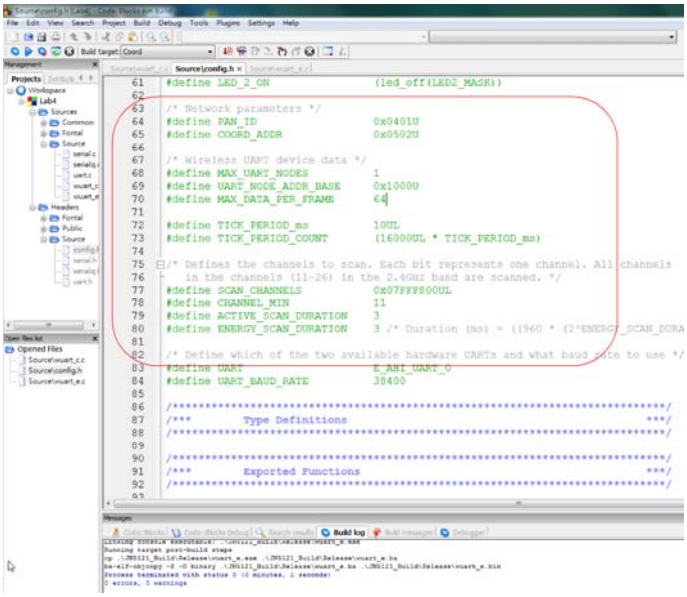


Figure 5
The configure program diagram

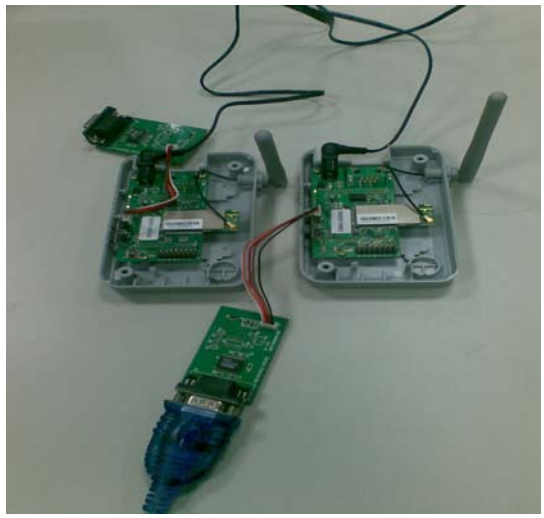


Figure 6
The development boards of WSN

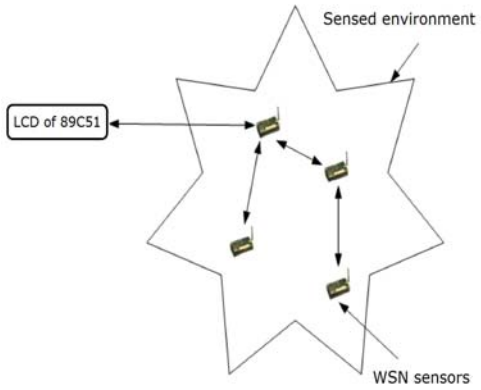


Figure 7
The concept diagram of WSN control

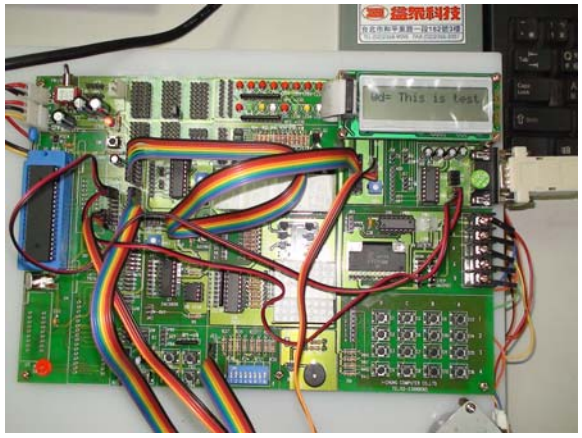


Figure 8
The diagram of LCD test

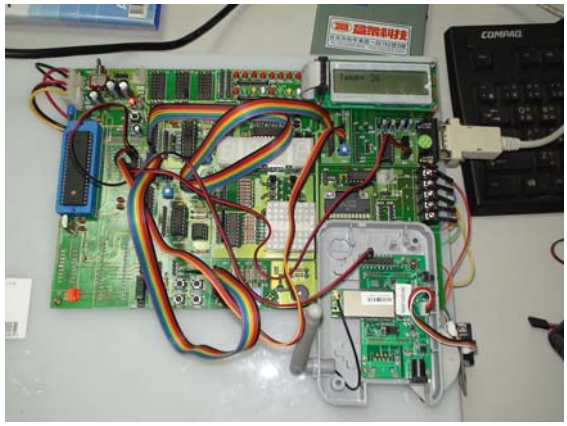


Figure 9
The implementation diagram of WSN for temperature monitor

Conclusions

In this paper, the design method for a temperature monitoring application using a wireless sensor network (WSN) is proposed. This paper has successfully demonstrated the application of the WSN to monitor the indoor temperature. The coordinator and end-device programs are developed using Code::Blocks software. The UART transmission and physical verification applications are also successfully demonstrated to possess good performance in data collection, temperature monitoring, motor control and display.

Acknowledgement

This paper is partially funded by teacher's research project of Taoyuan Innovation Institute of Technology.

References

[1] FT-6200 User Guide, 2012 (http://surewin.com.tw)

[2] Atmel 89C51 Microcontroller with 4Kbytes Flash Datasheet, Atmel Company, 2012 (http://www.datasheetcatalog.com/datasheets_pdf/8/9/C/5/89C51.shtml)

[3] A. Kalra and S. K. Kalra, Architecture and Programming of 8051 Microcontroller, Laxmi Pub. Ltd. 2010

[4] A. Sanz: A Complete Node for Power Line Communications in a Single Chip, International Symposium on Power Line Communications and Its Applications, 2005, pp. 285-289

[5] S. Y. Don: MCS-51 Practices and Designed by C language, I-Gung Publications, Ltd., Taiwan, 2008 (Trad. Chin. ver.)

[6] Simple type-A PGMSA (PGMSx IC WRITER) User Guide, I-Gung Publications, Ltd., Taiwan, 2008. (Trad. Chin. ver.)

[7] Keil u-Vision 2 Development Tools User Guide, ARM Inc., 2012. (http://www.keil.com/support/man_c51.htm)

[8] J. Kim, J. W. Choi and S. Lee: Universal I/O Design for Customizing MCU, Journal of Measurement Science and Instrumentation, Vol. 1, 2010, pp. 121-122

[9] W. Y. Chunga and S. J. Oh: Remote Monitoring System with Wireless Sensors Module for Room Environment, Senor and Actuator B, Vol. 13, 2006, pp. 64-70

[10] N. Golmie and I. Matta: Applications and Services in Wireless Networks, Computer Communications, Vol. 28, 2005, pp. 1603-1064

[11] I. J. Su, C. C. Tsai and W. T. Sung: Area Temperature System Monitoring and Computing Based on Adaptive Fuzzy Logic in Wireless Sensor Networks, Applied Soft Computing, Vol. 12, 2012, pp. 1532-1541

[12] P. Baronti, P. Pillai, V. W. C. Chook, S. Chessa, A. Gotta and Y. F. Hu: Wireless Sensor Networks: A Survey on the State of the Art and the 802.15.4 and Zigbee Standars, Computer Communications, Vol. 30, 2007, pp. 1655-1695

[13] L. Aguilar, G. Licea and J. A. García-Macías: An Experimental Wireless Sensor Network Applied in Engineering Courses, Computer Applications in Engineering Education, Vol. 19, 2011, pp. 777-786

[14] R. Belbachir, Z. M. Mekkakia and A. Kies: Towards a New Approach in Available Bandwidth Measures on Mobile Ad Hoc Networks, Acta Polytechnica Hungarica, Vol. 8, 2011, pp. 133-148

[15] L. K. Qabajeh, M. L. M. Kiah and M. M. Qabajeh: Secure Unicast Position-based Routing Protocols for Ad-Hoc Networks, Acta Polytechnica Hungarica, Vol. 8, 2011, pp. 191-214

[16] M. L. M. Kiah, L. K. Qabajeh and M. M. Qabajeh: Unicast Position-based Routing Protocols for Ad-Hoc Networks, Acta Polytechnica Hungarica, Vol. 7, 2010, pp. 19-46

[17] Jennic Board API Reference Manual (JN-RM-2003), Jennic Inc., 2007

[18] Jennic 802.15.4 Stack API Reference Manual (JN-RM-2002), Jennic Inc., 2007

[19] Jennic Application Queue API Reference Manual (JN-RM-2025), Jennic Inc., 2006

[20] Jennic Integrated Peripherals API Reference Manual (JN-RM-2001), Jennic Inc., 2007

A Stochastic Approach to Fuzzy Control

Károly Nagy*, Szabolcs Divéki*, Péter Odry*, Matija Sokola, Vladimir Vujičić*****

* Subotica Tech - College of Applied Sciences, Marka Oreškovića 16 24000 Subotica, Serbia, e-mail: [nagyk, diveki, odry]@vts.su.ac.rs

** The School of Higher Technical Professional Education in Novi Sad Školska 1, 21000 Novi Sad, Serbia, e-mail: sokola@vtsns.edu.rs

*** Department/Institute for Power, Electronics and Communications Engineering, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, e-mail: vujicicv@uns.ac.rs

Abstract: The paper presents the utilization of low-resolution data for control purposes. The control is based on fuzzy logic, with the deployment of stochastic digital low-resolution time arrays. Every control decision contains a degree of imprecision, being derived from measured low-resolution data. The imprecision is eliminated by stochastic noise superimposed during the data gathering, while the negative effects of noise are suppressed both by the fuzzy nature of the decision-making process and by the energy inertia in the controlled object. The proposed stochastic fuzzy control is extremely fast, robust and so simple that it practically does not need a microprocessor. This approach is validated by a simulation of holding upright an inverse pendulum.

Keywords: fuzzy control; fuzzy inference systems; approximate reasoning; fuzzification; alpha-cuts; stochastic

1 Introduction

Fuzzy logic and fuzzy reasoning have been shown to be a very effective approach in various control applications, especially when the control problem is multi-dimensional; when the plant model is unknown or time-varying; and/or when the feedback measured data are unreliable or unavailable [1]. In many control approaches, the measured feedback is extensively processed in order to eliminate measurement uncertainties and other errors, and such a processed feedback signal is used in the chosen control algorithm. Such processing of high-resolution data either puts further demands on processing capabilities or forces the reduction of the refresh rate of the controller output [2]. Hence, the utilization of accurate high-resolution data may become unsuitable for the control of fast multi-variable processes.

Representing analogue variables by a time-array of 1-bit binary signals has been researched for various purposes. Direct-stream digital, based on sigma-delta modulation, is employed in audio-technology for sound recording, and pulse-width modulation is widely used in power electronics, pulse position modulation, pulse density modulation, etc. All of these methods prove that it is possible to establish a strong correlation between an analogue value and a sufficiently long binary time array. Low-bit digital time arrays have also been researched in metrology and successfully utilized for fast low-resolution measurements [3], [4], [5], [6], [7], [8]. One of the key components is the introduction of stochastic dither, superimposed onto the input signal. Even 2-bit devices, with very coarse instantaneous measurements, will provide extremely accurate results [6]. In similar fashion, dither is utilised in [9] to enhance the quality of feedback signals for the fuzzy logic controller. The drawback is that high precision cannot be achieved by short time arrays of low-resolution data [3], [6], thus using such signals will lead to an imprecision in fast control decisions. However, a new control decision is arriving very soon. Can we use the ideas of making reliable overall systems from unreliable elements as proposed in [10] and utilise them to generate fast control decisions from imprecise low-resolution data, knowing that the control actions will, in time, converge to a stable state?

Over the last several decades, fuzzy logic and fuzzy reasoning have been shown to work effectively with imprecise data ([6], [10]). How to combine the stochastic signal processing with fuzzy control? Papers of Zadeh [11] and Goodman et al. [12], [13], [14] show theoretical possibilities of connecting Boolean algebra, conditional algebra, stochastic concepts and fuzzy logic in complementary ways. For control applications, fuzzy logic operations that include comparison of two or more membership functions are needed. How to incorporate the stochastic dimension that is inherently carried by the low-resolution nature of the feedback signal? One possible method is α-cuts [3], [14], [15], [16], which are used to best represent a certain feature of a set, i.e. to form a relationship between fuzzy sets and crisp sets. The stochastic feature is ensured by employing a randomly varying level of α in every control cycle. In this way it is possible to generate binary time-arrays, which can be compared in order to execute some fuzzy logic operations (for instance min, max operations).

The contribution of this paper is to show that it is possible to realize a simple controller in which probability theory and fuzzy logic complement each other, as theoretically suggested in [11]. In this case, the classical binary logic is utilised in a spirit of fuzzy logic philosophy. The link between these two logics is provided by a novel combination of stochastic principles and α-cuts. The resulting 1-bit time arrays are processed by classical Boolean algebra, necessary for individual control decisions. This approach is suitable for some control applications and offers some groundwork ideas for further development.

The proposed system differs from both classical fuzzy control and various improvements of fuzzy control [2], [9], [17]-[25]. A major aspect of difference is

the utilisation of very raw feedback signals for control. Although many control systems try to find the optimal output control signal in every decision cycle, the proposed system makes individual control decisions in such a way not to worsen the controlled variable. This means that we accept that the controlled object cannot be always brought to the required state within only a few control cycles - we are just driving it in an acceptable direction. Nevertheless, the overall control within a sufficient time interval converges towards an accurate control.

The paper is organized as follows: Section 2 presents the theory of using the α-cut sets approach for control purposes; Section 3 discusses the fuzzy sets and utilization of low-resolution signals for control of the inverse pendulum in the upright position and shows the simulation results.

2 α-cut Set as a Control Element

2.1 Decomposition Principles – the Model for Obtaining a Stochastic Array from a Membership Function

In classical fuzzy control, a membership function is an analogue value between 0 and 1 [1]. In the stochastic approach proposed in this paper, this analogue value is substituted by a time array of 1-bit signals (zeroes and ones) in such a way that the analogue value representing the membership function is the probability of appearance of value 1 in the 1-bit time array.

The model for obtaining such an array can be illustrated by the decomposition principle: An α-cut set is a discrete (crisp) set made up of members whose membership is greater than α [1], [3], [15], [16]:

$$A_\alpha = \{x \mid \mu_A(x) > \alpha\},\ \alpha \in [0,1) \tag{1}$$

Theoretically, the original continuous fuzzy set A can be decomposed into an infinite number of crisp α-cut sets.

The fuzzy set can be represented as a union of discrete sets expressed as:

$$A = \bigcup_\alpha \alpha A_\alpha,\ \alpha \in [0,1] \tag{2}$$

which means that the membership function is calculated using:

$$\mu_A(x) = \sup_{\alpha \in [0,1)} \left[\alpha \wedge \chi_{A_\alpha}(x)\right] \tag{3}$$

where $\chi_{A_\alpha}(x)$ is the characteristic function of the α-cut set A_α. This function represents a crisp discrete set, and hence the value of the membership level is 1 for

all its elements. When, for a chosen value of α, intersection with function $\chi_{A_\alpha}(x)$ is calculated, function αA_α is obtained. When αA_α functions for all α values from 0 to 1 are calculated and subjected to sup (finding the maximum) operation, a union operation is practically performed. In this way the original fuzzy set can be constructed.

In a similar fashion, it is possible to determine the membership function to a fuzzy set for individual elements if there is a finite number N of α-cut sets (N is also the number of samples of the feedback signal), but such that α is a random number of uniform probability distribution in the interval 0 to 1 ([11], [12], [13], [14]).

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\chi_{A_\alpha}(x)_i=\mu_A(x),\qquad \alpha\in[0,1) \tag{4}$$

This means that the characteristic function $\chi_{A_\alpha}(x)$ can be considered as a random variable, which assumes value 1 with probability $\mu_A(x)$, otherwise it assumes value 0. When α is lower than $\mu_A(x)$, then $\chi_{A_\alpha}(x)=1$.

$$\chi_{A_\alpha}(x)=\begin{cases}1 & \alpha<\mu_A(x)\\ 0 & \alpha\geq\mu_A(x)\end{cases} \tag{5}$$

Let us assume that instead of a fuzzy set there is an array of α-cut sets of the fuzzy set, such that α is a random variable with a uniform distribution $\alpha\in[0,1)$. Depending on the value of α, one element can be a member of an α-cut set or it can be outside the α-cut set, thus giving the required time array of zeroes and ones.

Signals obtained by low-bit quantization carry two pieces of information – the accurate value and the random error. The accurate value is extracted as an average of the array. The random error cannot be determined in every element of the array, but the random error for a whole array can be estimated ([5], [6], [7], [8], [26], [27]).

2.2 Comparison of Low-Resolution Signals – Minimum, Maximum, AND, OR Operations

In control systems of "if-then" type, the "if" part of the rule is formed from the membership functions of the input variables. The conditions and rules are formed in a shape of a logic expression, which contains the membership functions of the input variables [1]. If the input variables, at every digital tact cycle, can assume values of 1 and 0 only, then various operations on those variables can be executed within one cycle – extremely fast.

The actual elements of *n* input stochastic signal arrays can be considered as random variables, which assume value 1 with probabilities p_1, p_2, ..., p_n. If those random variables are uncorrelated, classical AND, OR, NAND NOR logic operations can be performed. However, any correlation between variables will change the meaning of logic operations.

It is possible to form logic terms from the elements of stochastic arrays. Knowing the probability of assuming the value of 1 in individual stochastic arrays and whether the variables are correlated or not, it is possible to determine the probability of the logic term output assuming the value of 1. In this way, the output stochastic array of similar properties as the two input arrays is obtained.

If the input arrays are uncorrelated, classical logic operations are valid. However, if every input variable is compared with the same α value of the α-cut set, the resulting operations become minimum, maximum or difference operations.

2.2.1 Uncorrelated Random Variables

Let us consider two input variables and at least two fuzzy sets. The membership function of the first input variable to the first fuzzy set is $\mu_A(x)$, while the membership function of the second input variable to the second fuzzy set is $\mu_B(\omega)$ where $0 \le \mu_A(x), \mu_B(\omega) \le 1$; α_1 and α_2 are independent random variables of uniform probability distribution $0 \le \alpha_1, \alpha_2 \le 1$, utilised for determining the α-cut sets of the corresponding fuzzy sets.

2.3 Logic Operations on Corresponding Elements of the Arrays which Describe Membership Functions $\mu_A(x)$ and $\mu_B(\omega)$

Based on the α-cut set models, an event belongs to the α-cut set if $\alpha_1 < \mu_A(x)$, while if $\alpha_1 \ge \mu_A(x)$ the event doesn't belong. (Ψ_1 is the actual element of the first stochastic array). Event $\Psi_1 = 1$ has the probability $P(\Psi_1 = 1) = \alpha_1$, while the probability of the event $\Psi_1 = 0$ is $P(\Psi_1 = 0) = 1 - \alpha_1$. Similarly, the second array has features $P(\Psi_2 = 1) = \alpha_2$ and $P(\Psi_2 = 0) = 1 - \alpha_2$.

If the input variables are uncorrelated, the probabilities of possible combined events are:

$$P(\Psi_1 = 1, \Psi_2 = 1) = \alpha_1 \cdot \alpha_2 \tag{6}$$

$$P(\Psi_1 = 1, \Psi_2 = 0) = \alpha_1 \cdot (1 - \alpha_2) \tag{7}$$

$$P(\Psi_1 = 0, \Psi_2 = 1) = (1 - \alpha_1) \cdot \alpha_2 \tag{8}$$

$$P(\Psi_1 = 0, \Psi_2 = 0) = (1 - \alpha_1) \cdot (1 - \alpha_2) \tag{9}$$

If the four above expressions are added, their combined probability equals 1, confirming that a complete field of possible events is described.

Equation (6) represents AND operation, while the combination of (7), (8) and (9) represents OR operation.

2.3.1 Minimum and Maximum Operations

Let us consider two membership functions $\mu_A(x)$ and $\mu_B(\omega)$ and assume $\mu_A(x) > \mu_B(\omega)$. If those two signals are compared with the same random number α, the following states can be obtained:

$$\mu_A(x) > \alpha \wedge \mu_B(\omega) > \alpha \tag{10}$$

$$\mu_A(x) > \alpha \wedge \mu_B(\omega) < \alpha \tag{11}$$

$$\mu_A(x) < \alpha \wedge \mu_B(\omega) < \alpha \tag{12}$$

but the event

$$\mu_A(x) < \alpha \wedge \mu_B(\omega) > \alpha \tag{13}$$

is an impossible event.

Performing logic AND operation on such random variables, the lowest value of membership functions can be obtained, since all membership functions can be greater than α if the lowest function is greater than α. The output array is equivalent to the array of lowest membership function,

$$(\mu_B(\omega) > \alpha) \Rightarrow (\mu_A(x) > \alpha) \tag{14}$$

The minimum operation on the two membership functions is performed by AND logic operation:

$$\min(\mu_A(x), \mu_B(\omega)) = \mu_A(x) \wedge \mu_B(\omega) \tag{15}$$

By OR operation the largest of the two compared membership functions are chosen. Consequently, if one the largest membership function is greater than the random number α, then at least one of the compared functions is larger than α.

The maximum operation is performed by the logic OR operation:

$$\max(\mu_A(x), \mu_B(\omega)) = \mu_A(x) \vee \mu_B(\omega) \tag{16}$$

It is also possible to calculate the difference between two membership functions:

$$\mu_A(x) - \mu_B(\omega) = \mu_A(x) \wedge \overline{\mu_B(\omega)} \tag{17}$$

Proof: for

$$\begin{aligned} &\alpha > \mu_A(x) && \Psi_A = 0 \;\; \Psi_B = 0 \\ &\mu_A(x) > \alpha > \mu_B(\omega) && \Psi_A = 1 \;\; \Psi_B = 0 \\ &\alpha < \mu_B(\omega) && \Psi_A = 1 \;\; \Psi_B = 1 \end{aligned} \tag{18}$$

The purpose of the above MIN, MAX and DIF operations is to reduce the amount of calculations on stochastic arrays so that very simple logic circuits can be used instead of a microprocessor. This will enormously increase the speed of sampling and processing, resulting in the possibility that the control output is being updated in every processor cycle.

2.3.2 Choosing the Membership Functions to Fuzzy Sets in the Form of Stochastic Arrays

Let us consider an arbitrary trapezoidal shaped fuzzy set and the random number α (dashed line), Figure 2 [1], [28].

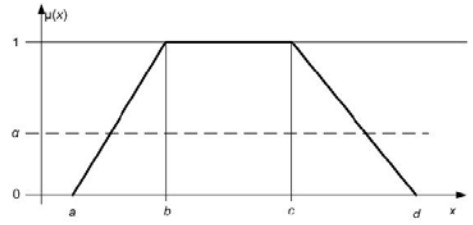


Figure 2
A trapezoidal fuzzy set and the random number α

The input variable *x* belongs to the α-cut set if: $\mu_A(x) > \alpha$. In order to determine the α-cut set interval of the input variable, it is necessary to determine in which part of the membership function the input variable is greater than the random variable α. The intersection points are determined as:

$$\alpha = \frac{x-a}{b-a} \Rightarrow x = a + \alpha(b-a) \tag{19}$$

$$\alpha = \frac{d-x}{d-c} \Rightarrow x = d - \alpha(d-c) \tag{20}$$

Therefore, the condition for belonging to an α-cut set is:

$$\left[a + \alpha(b-a)\right] < x \wedge x < \left[d - \alpha(d-c)\right] \tag{21}$$

This can be rearranged as:

$$a < x - \alpha(b-a) \wedge x + \alpha(d-c) < d \tag{22}$$

i.e.

$$\frac{b+a}{2} < x - \alpha(b-a) + \frac{b-a}{2} \wedge x + \alpha(d-c) - \frac{d-c}{2} < \frac{d+c}{2} \tag{23}$$

As α is a random number between 0 and 1, terms $\alpha(b-a)$ and $\alpha(d-c)$ are random numbers of uniform distribution.

When the sum of the measured variable and a random number is compared with a decision trigger level, this is a process very similar to the stochastic additive A/D conversion against decision levels PO_1 and PO_2 ([5], [6], [7], [8]). If the random dithers $h_1(t)$ and $h_2(t)$ are added to the measured analogue signal *x*, then the digitizing process is defined by:

$$PO_1 < x - h_1(t) \wedge x + h_2(t) < PO_2 \tag{24}$$

$$h_1(t) = \alpha(b-a) - \frac{b-a}{2} \tag{25}$$

$$h_2(t) = \alpha(d-c) - \frac{d-c}{2} \tag{26}$$

As in this case $(b-a) = (d-c)$, it follows that $h_1(t) = h_2(t) = h(t)$.

3 Utilization of Low-Resolution Signals for Control

To illustrate the feasibility of the proposed control system, the classic example of holding the inverse pendulum in the upright position was chosen. Similarly, [2], [9], [17] – [25] use the inverse pendulum in order to validate their proposed fuzzy controllers.

The authors have investigated a practical application of the proposed stochastic fuzzy control system for the very fast and accurate control of arc welding. In such applications, problems with fast-changing plant parameters and the presence of very high levels of noise are pronounced. On the other hand, actuation is very fast: the PWM modulation of welding current is performed by transistors operating in switching mode at frequencies up to 100 kHz. Such requirements can be met by the proposed low-resolution control system. The practical implementation that is under development employs analogue summation of analogue dither and analogue measurement signal, followed by 1-bit digitalization.

3.1 The Control Problem of Inverse Pendulum

The simplest one degree-of-freedom (DOF) inverse pendulum system on a trolley, shown in Figure 3, is considered. The actuator is either a single-level or multiple-level impulse actuator, acting in bidirectional bang-bang mode, which may be described as push/do_nothing/pull (F_+, 0, F_-) type action,. In order to control the pendulum, it is necessary to monitor/measure and control two variables: θ - the angle of the pendulum from the vertical axis and ω - the angular velocity of the pendulum, similarly to [9].

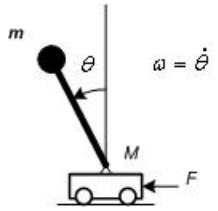


Figure 3
Illustration of the inverse pendulum with one DOF

3.2 Control Utilising One-Level Actuator

The first control system is designed for an actuator which has only one level of possible output in each direction - single level bidirectional bang-bang. The control system operates at a higher frequency, in order to accomplish the control task with a moderate actuator force, even in cases of unfavourable initial conditions or strong disturbances.

3.2.1 Definitions of Fuzzy Sets and Membership Functions

As both measured variables are single-dimensional, the fuzzy sets can be defined as both fuzzy numbers and fuzzy intervals. Three fuzzy sets ("negative", "zero" and "positive") are chosen within the measured angle interval, but with triangular rather than trapezoidal membership functions. Hence values b and c from Figure 2 are identical, b=c. Furthermore, the target value for the control system in this case is the vertical position, i.e. zero, thus $\theta b = \theta c = 0$. In such a case, the membership functions for the pendulum **angle control** are defined as follows:

For negative angle Θ_{NN}, $\mu_{\Theta_{NN}}(\theta) = \begin{cases} 1 & if \quad \theta \le \theta a \\ \dfrac{\theta}{\theta a} & if \quad \theta a < \theta < 0 \\ 0 & if \quad \theta \ge 0 \end{cases}$ (27)

For "zero" angle Θ_{ZZ}, $\mu_{\Theta_{ZZ}}(\theta) = \begin{cases} 0 & if \quad \theta \le \theta a \\ \dfrac{\theta a - \theta}{\theta a} & if \quad \theta a < \theta \le 0 \\ \dfrac{\theta d - \theta}{\theta d} & if \quad 0 < \theta < \theta d \\ 0 & if \quad \theta \ge \theta d \end{cases}$ (28)

For positive angle Θ_{PP}, $\mu_{\Theta_{PP}}(\theta) = \begin{cases} 0 & if \quad \theta \le 0 \\ \dfrac{\theta d - \theta}{\theta d} & if \quad 0 < \theta < \theta d \\ 1 & if \quad \theta \ge \theta d \end{cases}$ (29)

These membership functions are shown in Figure 4:

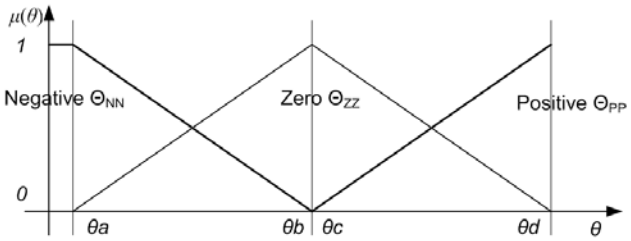


Figure 4
Fuzzy sets of the pendulum angle

The fuzzy sets and the membership functions for the **control of angular velocity** are defined in the same way; only the lower threshold velocity is denoted *ωa* and the upper threshold is denoted *ωd*. With the identical control target of resting (zero velocity) in the upright position, equations for negative, "zero" and positive angular velocity (Ω_{NN}, Ω_{ZZ} and Ω_{PP} respectively), are similar to (27)-(29), and the membership functions are as shown in Figure 5.

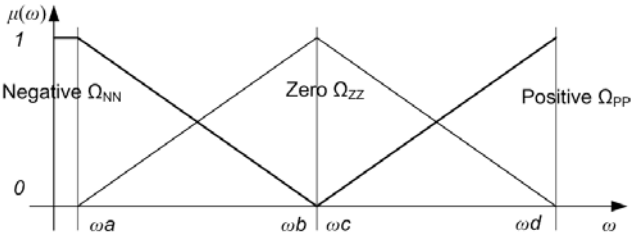


Figure 5
Fuzzy sets of pendulum angular velocity

3.2.2 Fuzzy Rules for One-Level Actuator

The control problem of holding the inverse pendulum upright is defined by rules of fuzzy decision as follows:

$$
\begin{array}{lllll}
\text{R1:} & \text{If} \quad \theta > 0 & \text{and} \quad \omega > 0, & \text{then} & F = F_{+}, \\
\text{R2:} & \text{If} \quad \theta > 0 & \text{and} \quad \omega = 0, & \text{then} & F = F_{+}, \\
\text{R3:} & \text{If} \quad \theta > 0 & \text{and} \quad \omega < 0, & \text{then} & F = 0, \\
\text{R4:} & \text{If} \quad \theta = 0 & \text{and} \quad \omega > 0, & \text{then} & F = F_{+}, \\
\text{R5:} & \text{If} \quad \theta = 0 & \text{and} \quad \omega = 0, & \text{then} & F = 0, \\
\text{R6:} & \text{If} \quad \theta = 0 & \text{and} \quad \omega < 0, & \text{then} & F = F_{-}, \\
\text{R7:} & \text{If} \quad \theta < 0 & \text{and} \quad \omega > 0, & \text{then} & F = 0, \\
\text{R8:} & \text{If} \quad \theta < 0 & \text{and} \quad \omega = 0, & \text{then} & F = F_{-}, \\
\text{R9:} & \text{If} \quad \theta < 0 & \text{and} \quad \omega < 0, & \text{then} & F = F_{-}.
\end{array}
\tag{30}
$$

where: F_{-} is the constant force in negative direction and F_{+} is the constant force in positive direction.

On the basis of the above fuzzy logic rules, the logic circuit with only 12 logic gates, shown in Figure 6, can be constructed.

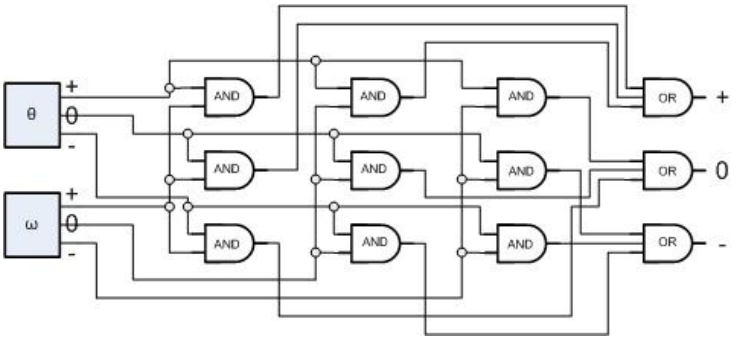
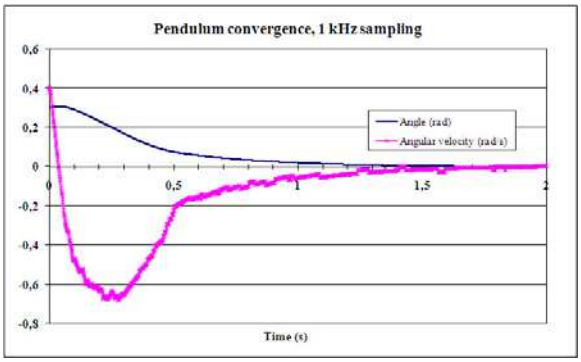


Figure 6
Arithmetic-logic scheme of fuzzy control for a single-level actuator

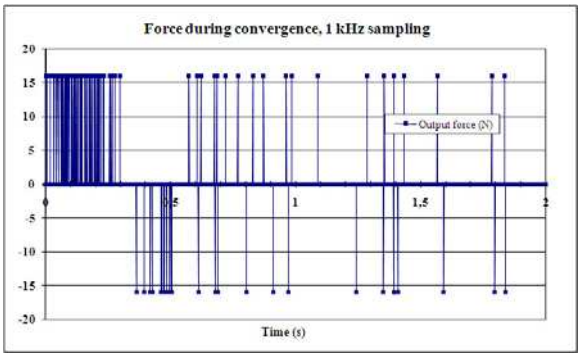
3.2.3 Simulation Results of the Control Utilizing One-Level Actuator

The above fuzzy rules have been faithfully modelled into a simulation of the inverse pendulum system. Although such simple control hardware can be extremely fast, a very moderate frequency of 1 kHz has been chosen for the simulations. The physical parameters are: trolley mass M=1 kg, pendulum mass m=0.1 kg, pendulum height h=1 m, available actuator force F= ±16 N.

The results of the first two seconds of bringing out-of-balance pendulum into a stable upright position are shown in Figure 7. The initial conditions are quite challenging: the pendulum is 0.3 radians out of balance, falling further with 0.4 rad/s angular velocity.



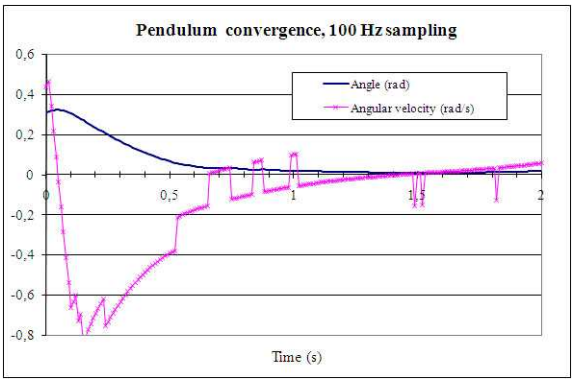
a) angle and angular velocity during the simulation



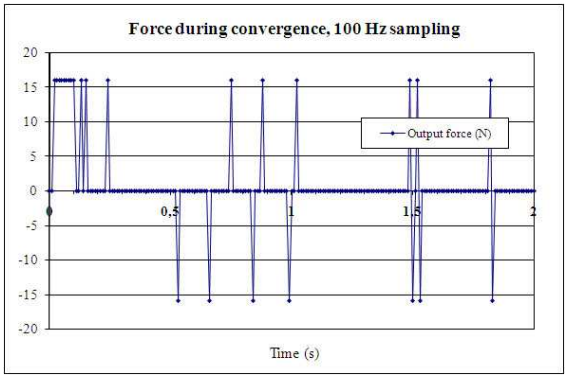
b) b) actuator force during the simulation

Figure 7
Aligning the inverse pendulum into the upright position, sampling 1 kHz, 16N actuator

The simulated values of the pendulum angle and its angular velocity are shown in Figure 7a, and it can be seen that the control system is very effective. It stops the pendulum falling further after less than 0.1 seconds (angular velocity becoming negative), brings it very close to the upright position in less than 1.5 seconds without overshooting, and keeps it stable afterwards. The actuator force, shown in Figure 7b, displays a lot of activity in the first 0.35 s, then moderate activity for the next second, and then is required to act just occasionally afterwards. The stochastic nature of the controller can be observed, at around 1.83 seconds, and although the pendulum is upright and not moving away, there is one positive impulse and then immediately one negative impulse of the actuator. This is a waste of energy: two burst were applied when none was really needed.



a) angle and angular velocity during the simulation



b) actuator force during the simulation

Figure 8
Aligning the inverse pendulum into the upright position, sampling 100 Hz

The required force impulses are further investigated in 10 simulation runs; with identical initial conditions, the stochastic nature of the controller makes every simulation slightly different. Overall, only around 60% of the actuator force output is used for lifting the pendulum from unbalanced to the upright position, while around 40% of the actions is wasted due to the stochastic nature of the controller.

The effects of reducing the sampling frequency were investigated next. Figure 8 shows the position and the angular velocity for the sampling frequency of 100 Hz, with an original actuator force of 16 N. It can be seen that the system is still performing well, converging to a near upright position within less than 2 seconds and holding it upright afterwards. However, higher fluctuations of angular velocity can be observed in a steady state. These fluctuations are more pronounced because the duration of every actuator action is 10 times longer, and the energy

inserted during one control cycle is now 10 times larger. Further reductions in operating frequency further increases velocity fluctuations, thus compromising the accuracy of tracking and eventually would result in an unstable system.

3.2.4 Comparison with High-Resolution Fuzzy Control

A wide comparison of the proposed control approach based on low-resolution (LR) data against the fuzzy control based on high-resolution (HR) data has been performed. Many simulations runs have been performed, at different sampling frequencies and with different actuator forces. Following on from the discussion in the previous paragraph, the expected deviations in velocity response increase with a reduction in sampling frequency; therefore the results at 100 Hz are shown in Figure 9. Responses of the pendulum angle and its angular velocity obtained with the HR control system are depicted in bold lines, while the thin lines are responses of the proposed LR stochastic fuzzy controller, for three randomly chosen simulation runs.

It is interesting to note that simulation runs of the LR control differ from each other, due to the stochastic dither. This confirms our initial idea that individual control decisions do not need to be always the best in every time instant, but the proposed control method will provide the overall convergence of the controlled plant towards the required state.

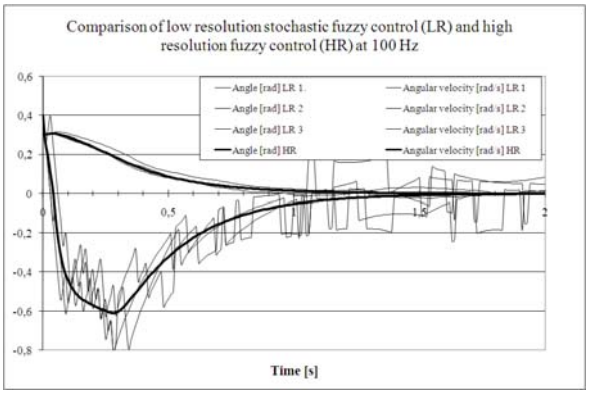


Figure 9
Comparison of high-resolution and low-resolution (3 simulation runs) control, sampling 100 Hz

3.3 Control Utilising a Three-Level Actuator

From the results shown in Section 2.3.2, as well as from many conducted simulations with different sampling frequencies and/or actuator force levels, a collision of three features can be observed:

- a low control sampling rate is beneficial for the reduction of imprecision in individual control decisions, but it increases the fluctuations around the steady-state position,
- a lower actuator force is good for minimising the total force impulses but it limits the maximum system capabilities,
- aggregate force input increases with both too high and too low control sampling rates.

In order to optimise, rather than compromise between sampling frequency, tracking accuracy, actuator available force and energy efficiency, a three-level actuator is implemented. The actuator output force has three digital levels (*low*, *medium, high*), in two directions, so that it can assume seven possible levels.

The control strategy is adapted so that it operates in two modes of control:

1) fine control, when both angular position and angular velocity are within their threshold limits, or
2) forceful control, when at least one of the controlled variables is outside the threshold limits.

The fuzzy rules are adapted so that only the low-level force is applied in the fine control mode, while medium and high force levels can be applied in the forceful control mode.

3.3.1 Definitions of Fuzzy Sets and Membership Functions

The three fuzzy sets for the forceful **angle control** are identical as before, as given by (27)-(29) and shown in Figure 4. When the angle is within the fine regulation thresholds, $\theta \in (\theta af, \theta df)$, then the three fuzzy sets for fine regulation are negative fine angle Θ_N, "zero" fine angle Θ_Z and positive fine angle Θ_P. All six fuzzy membership functions are shown in Figure 10.

Using an equivalent approach, the fuzzy sets for **angular velocity control** are defined for forceful control and for fine control, as shown in Figure 11.

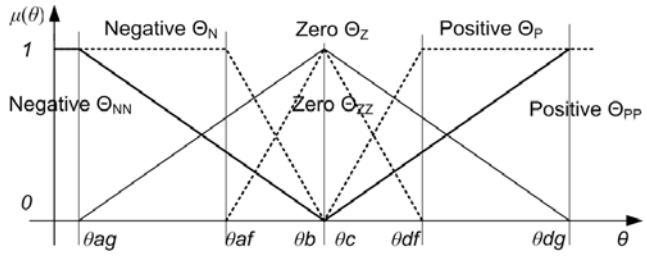


Figure 10
Fuzzy sets of the pendulum angle for forceful/fine control

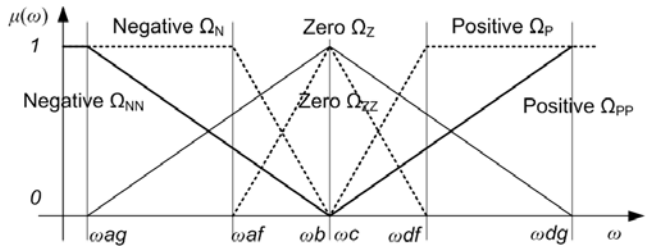


Figure 11
Fuzzy sets of pendulum angular velocity

3.3.2 Fuzzy Rules for a Three-Level Actuator

The rules of fuzzy decision are defined as follows:

In the fine control mode, when $\theta \in (\theta af, \theta df)$ and $\omega \in (\omega af, \omega df)$, fuzzy rules are:

R11: If θ>0 and ω>0, then F= + *low* ,
R12: If θ>0 and ω=0, then F= + *low* ,
R13: If θ>0 and ω<0, then F= 0 ,
R14: If θ=0 and ω>0, then F= + *low* ,
R15: If θ=0 and ω=0, then F= 0 , (31)
R16: If θ=0 and ω<0, then F= - *low* ,
R17: If θ<0 and ω>0, then F= 0 ,
R18: If θ<0 and ω=0, then F= - *low* ,
R19: If θ<0 and ω<0, then F= - *low* .

Otherwise, forceful control is performed, using the following membership functions:

R21: If θ>>0 and ω>>0, then F= + *high* ,
R22: If θ>>0 and ω=0, then F= + *medium* ,
R23: If θ>>0 and ω<<0, then F= 0 ,
R24: If θ=0 and ω>>0, then F=+ *medium* ,
R25: If θ=0 and ω=0, then F=0 , (32)
R26: If θ=0 and ω<<0, then F= - *medium* ,
R27: If θ<<0 and ω>>0, then F= 0 ,
R28: If θ<<0 and ω=0, then F= - *medium* ,
R29: If θ<<0 and ω<<0, then F= - *high* .

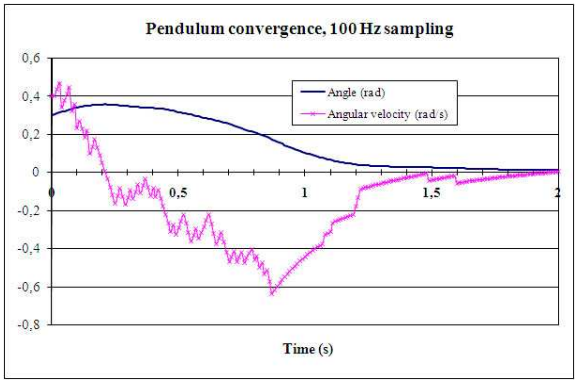
3.3.3 Simulation Results for a Three-Level Actuator

Simulations have been conducted for the data as in section 3.2.3, except the following:

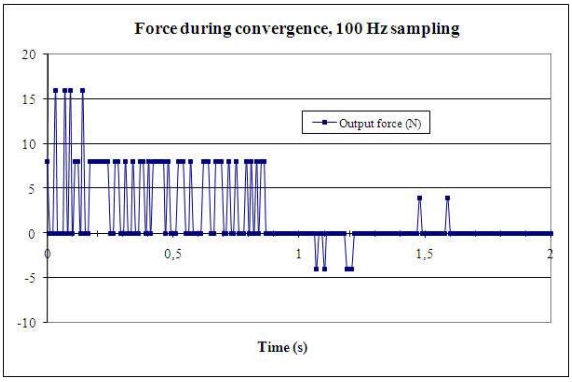
- the three levels of actuator force are chosen as high=16 *N*, medium = 8 *N* and low = 4 *N*,

- the thresholds values for fine control are $\theta af = -\theta df = 0.2\ rad$ and $\omega af = -\omega df = 0.8\ rad/s$
- the sampling frequency is set to 100 *Hz*.

A sample of simulation results is shown in Figure 12: the pendulum angle and angular velocity (Fig. 12a) converge to a steady upright position within 1.5 seconds, with very small tracking errors after that time. The actuator force (Fig. 12b) displays very little activity once the control system enters the fine control mode. The benefit of this is that the overall actuator output (energy requirement) is around 3 times lower than in the case of single-level actuator with 100 *Hz* sampling, Figure 8.



a) angle and angular velocity during the simulation



b) actuator force during the simulation

Figure 12
Aligning the inverse pendulum into the upright position, 100 Hz sampling, 16/8/4 N actuator

Conclusions

The paper shows the possibility to construct a stochastic fuzzy control system that utilizes low-resolution signals. The method for decomposing a fuzzy set into a time array of stochastic α-cut sets, which enables representation of the fuzzy membership function by a stochastic array of zeroes and ones, is practically implemented. This is a way of combining the probability theory with fuzzy logic, in a complementary manner.

For any fuzzy set, α-cut sets can be uncorrelated and then classical logic operations can be applied. Otherwise, it is possible to determine an α-cut set of all fuzzy sets for the same value of α and then utilize MIN and MAX logic operations. In this way, fuzzy rules of if-then type are transformed into logic expressions. As a result, control is reduced to determining the output values of these logic expressions. In MIN and MAX operations, one of the input variables is passed as an output. Hence features of one input are mirrored to the output; in this way the character of the random error in the input variable is unchanged. During the feedback measurement process, i.e. while gathering the input data for the controller, the chosen α value is introduced as a random error. The same error is reflected in every instantaneous control decision, but the error is sufficiently suppressed after an array of control decisions is made. In essence, the proposed control procedure follows a novel philosophy. Within every cycle the procedure is: dithering, coarse digitalization, making rough (inaccurate) control decisions, execution of those control actions. If this cycle is repeated in a very fast manner, the overall control errors will not exist.

The resulting control system is very simple and robust; it doesn't perform complex mathematical operations and can operate without a microprocessor; and it can very quickly make an array of control decisions. The effectiveness of a stochastic fuzzy control approach is validated by simulation.

Acknowledgement

This work was supported by the Serbian Ministry of Education and Science under Grant TR 32019.

References

[1] G. J. Klir, B. Yuan: Fuzzy Sets and Fuzzy Logic Theory and Applications, Prentice Hall, 1995

[2] T. Lazar, P. Pástor: Factors Limiting Controlling of an Inverted Pendulum *Acta Polytechnica Hungarica,* Vol. 8, No. 4, 2011, pp. 23-34

[3] A. Pedrycz, F. Dong, K. Hirota. Finite Cut-based Approximation of Fuzzy Sets and its Evolutionary Optimization. *Fuzzy Sets and Systems*, Vol. 160, 2009, pp. 3550-3564

[4] K. Nagy, M. Takács: Type-2 Fuzzy Sets and SSAD as a Possible Application, *Acta Polytechnica Hungarica,* Vol. 5, No. 1, 2008, pp. 111-120

[5] V. Vujičić: Generalized Low Frequency Stochastic True rms Instrument, *IEEE Trans. Instrum. Meas.*, Vol. 50, 2001, pp. 1089-1092

[6] V. Vujičić, S. Milovančev, M. Pešaljević, D. Pejić, I. Župunski: Low Frequency Stochastic True rms Instrument, *IEEE Trans. Instrum. Meas.*, Vol. 48, 1999, pp. 467-470

[7] V. Vujičić, I. Župunski, Z. Mitrović, M. Sokola. Measurement in a Point Versus Measurement Over an Interval, *Proc. of the IMEKO XIX World Congress*, paper No. 480, 2009

[8] B. Santrač, M. Sokola, Z. Mitrović, I. Župunski, V. Vujičić: A Novel Method for Stochastic Measurement of Harmonics at low Signal-to-Noise Ratio, *IEEE Trans. Instrum. Meas.*, Vol. 58, No. 10, 2009, pp. 3434-3441

[9] J. Mao: Reduction of the Quantization Error in Fuzzy Logic Controllers by Dithering, Master's thesis, Ottawa-Carleton Institute for Electrical and Computer Engineering, University of Ottawa, November 1998

[10] J. von Neumann: Probabilistic Logic and the Synthesis of Reliable Organisms from Unreliable Components, In Automata studies. CE Shannon, Ed. Princeton, NJ, Princeton University Press, 1956

[11] L. A. Zadeh: Discussion: Probability Theory and Fuzzy Logic Are Complementary Rather Than Competitive, *Techmometrics*, Vol. 37, No. 3, 1995

[12] H. T. Nguyen: On Foundations of Fuzzy Theory for Soft Computing *International Journal of Fuzzy Systems*, Vol. 8, No. 1, 2006, pp. 39-45

[13] W. C. Torrez, D. Bamber, I. R. Goodman, H. T. Nguyen: "A New Method for Representing Linguistic Quantifications by Random Sets With Applications to Tracking and Data Fusion" *Proceedings of the Fifth International Conference on Information Fusion (FUSION 2002)*, July 8-11, 2002, Annapolis, MD: Volume 1, pp. 1308-1315

[14] R. Goodman: Random Sets and Fuzzy Sets: A Special Connection, *FUSION '98 International Conference,* pp. 93-100

[15] S. Bodjanova: A Generalized α-cut. *Fuzzy Sets and Systems*, Vol. 126, 2002, pp. 157-176

[16] Van-Nam Huynh, Y. Nakamori, J. Lawry: A Probability-based Approach to Comparison of Fuzzy Numbers and Applications to Target-Oriented Decision Making, *IEEE Trans. on Fuzzy Systems,* Vol. 16, No. 2, 2008, pp. 371-387

[17] Y. Beceriklia, B. K. Celik: Fuzzy Control of Inverted Pendulum and Concept of Stability using Java Application, *Mathematical and Computer Modelling 46*, 2007, pp. 24-37

[18] E. Minnaert, B. Hemmelman, D. Dolan: Inverted Pendulum Design with Hardware Fuzzy Logic Controller, *Systemics, Cybernetics and Informatics,* Vol. 6, No. 3, 2011 pp. 34-39

[19] M. Akole, B. Tyagi: Design of Fuzzy Logic Controller For Nonlinear Model of Inverted Pendulum-Cart System, *XXXII National Systems Conference, NSC 2008,* pp. 750-755

[20] M. I. El-Hawwary, A. L. Elshafei, H. M. Emara, H. A. Abdel Fattah: Adaptive Fuzzy Control of the Inverted Pendulum Problem, *IEEE Trans. on Control Systems Technology*, Vol. 14, No. 6, 2006, pp. 1135-1144

[21] M. Dotoli, B. Maione, D. Naso, B. Turchiano: Fuzzy Sliding Mode Control for Inverted Pendulum Swing-up with Restricted Travel *Fuzzy Systems 2001, the 10th IEEE International Conference*, Vol. 3, pp. 753-756

[22] C. H. Huang, W. J. Wang, C. H. Chiu: Design and Implementation of Fuzzy Control on a Two-Wheel Inverted Pendulum, *IEEE Trans. on Industrial Electronics*, Vol. 58, No. 7, 2011, pp. 2988-3001

[23] C. W. Tao, J. S. Taur, T. W. Hsieh, C. L. Tsai: Design of a Fuzzy Controller with Fuzzy Swing-Up and Parallel Distributed Pole Assignment Schemes for an Inverted Pendulum and Cart System, *IEEE Trans. on Control Systems Technology*, Vol. 16, No. 6, 2008, pp. 1277-1288

[24] J. Yi, N. Yubazaki: Stabilization Fuzzy Control of Inverted Pendulum systems, *Artificial Intelligence in Engineering* 14 (2000) pp. 153-163

[25] N. Muškinja, B. Tovornik: Swinging Up and Stabilization of a Real Inverted Pendulum, *IEEE Trans. on Industrial Electronics*, Vol. 53, No. 2, 2006, pp. 631-639

[26] B. Widrow and I. Kollár: Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control and Communications, Cambridge University Press, 2008

[27] A. Patel, B. Kosko: Noise Benefits in Quantizier-Array Correlation Detection and Watermark Decoding, *IEEE Trans. on Signal Processing,* Vol. 59, No. 2, 2011, pp. 488-505

[28] R. Gayakwad: Optimized Fuzzy Logic for Motion Control, *Acta Polytechnica Hungarica,* Vol. 7, No. 5, 2010, pp. 161-168

The Digital Self-Tuning Control of Step a Down DC-DC Converter

Fatima Tahri, Ali Tahri, Ahmed Allali and Samir Flazi

LGEO Laboratory of Electrical Engineering, Department of Electrotechnics, University of Sciences and Technology of Oran, BP 1505 El Mnaouar (31000 Oran), Algeria
E-mail: Tahri-f@univ-usto.dz, tahri-ali@univ-usto.dz, allali@univ-usto.dz, flazi@univ-usto.dz

Abstract: A digital self-tuning control technique of DC-DC Buck converter is considered and thoroughly analyzed in this paper. The development of the small-signal model of the converter, which is the key of the control design presented in this work, is based on the state-space averaged (SSA) technique. Adaptive control has become a widely-used term in DC-DC conversion in recent years.
A digital self-tuning Dahlin PID and a /Keviczky PID controller based on recursive least-squares estimation are developed and designed to be applied to the voltage mode control (VMC) approach operating in a continuous conduction mode (CCM).
A comparative study of these two digital self-tuning controllers for step change in input voltage magnitude or output load is also carried out.
The simulation results obtained using a Matlab SimPowerSystems toolbox to validate the effectiveness of the proposed strategies are also given and discussed extensively.

Keywords: Continuous Conduction Mode (CCM); Digital Self-tuning Controller; Dahlin PID; Bányász/Keviczky PID; State-Space Averaged (SSA); Voltage Mode Control (VMC)

1 Introduction

Usually power electronic systems consist of one or more power converters which convert one form and/or level of electrical energy into another form or level of electrical energy at the load, thanks to power semiconductor devices controlled by switching action. The advances and availability of modern power semiconductor devices used in power converters have made the switching converter a popular choice in power supplies.

Since the early 1970s, a large number of DC-DC converter circuits have been thoroughly analyzed and designed [1]. Such a converter can increase or decrease the magnitude of the DC voltage and/or invert its polarity. The Buck converter [2],

which uses the switch in series with the supply voltage, is a topology that gives a lower voltage at the load. In contrast, in the topology known as the Boost converter [3], the positions of the switch and inductor are interchanged, which allows this converter to produce an output DC voltage that is greater in magnitude than the input voltage. In the Buck-Boost converter [4], the switch alternately connects the inductor across the power input and output voltages. This converter inverts the polarity of the voltage and can either increase or decrease the voltage magnitude. The Ćuk converter [5] contains inductors in series with the converter input and output ports. The switch network alternately connects a capacitor to the input and output inductors.

The conversion ratio is identical to that of the Buck-Boost converter. Hence, this converter also inverts the voltage polarity, while either increasing or decreasing the voltage magnitude. The single-ended primary inductance converter (SEPIC) can also either increase or decrease the voltage magnitude. However, it does not invert the polarity [6]. These converters are extensively used in electronic equipment such as computer power supplies and battery chargers, and in medical, military and space applications [1].

The DC-DC converter represents different circuit topologies or configurations within each switching cycle. For the continuous conduction mode (CCM), there are two topologies. For the discontinuous conduction mode (DCM) of operation, a third configuration has to be added to yield a total of three topologies. In each configuration, the system can be described by linear state equations. Switching between the different topologies will vary from cycle to cycle depending on the output of the system, and this further complicates the analysis. This converter presents a nonlinear dynamical due to switching power devices and passive components.

The main approach to modeling DC-DC converters is the state-space averaging method [1], [7], [8]. The averaged continuous-time model uses the duty cycle as an input and describes the system's slow dynamics, to avoid difficulties posed by the hybrid nature of the system. This model is nonlinear due to the presence of multiplicative terms involving the state variables and the duty cycle. The averaging procedure hides all information about the fast dynamics of the system and fast instabilities (subharmonic oscillations are not captured). In all switching converters, the output voltage is a function of input line voltage, duty cycle, and the load current as well as the converter circuit element values. The scope is to achieve output voltage regulation with voltage mode control (VMC) approach in the presence of input voltage and output load variations.

Many digital controllers using discrete models of converters have been developed in the literature [9], [10] but it is clear that the Proportional-Integral-Derivative (PID) control has been used successfully for regulating processes in industry for more than 60 years. This is so because the design method can be easily grasped and its implementation is very simple.

Such regulators can usually meet demands, but when the dynamic characteristics of the controlled process vary, the PID regulator's parameters must be readjusted to follow suite. In this case, an adaptive controller should be designed to follow the changes of the operation conditions [11]. The question arises as to how these controllers can be adaptively tuned? In general, self tuners that are capable of automatically adjusting the control loop coefficients are based on recursive least-squares estimation [12], [13].

Many methods and formulae have been developed for tuning the PID controller. Some important examples are the Ziegler and Nichols formula in 1942 [14], the Cohen and Coon formula in 1953 [15], and the Åström and Hägglund in 1985. In 1991 Hang *et al.* introduced the refined Ziegler and Nichols settings [16], the dominant pole design, proposed by Åström and Hägglund, the internal model control (IMC) design method [17], the ITAE integral of the time weighted absolute error [18], and ISE integral of the squared error. There was also the optimal formulae and the gain and phase margin (GPM) design method [19]. The Dahlin controller was proposed by Dahlin [20] and Higham [21], independently. The Dahlin controller is a distinctive algorithm for the control of single input/single output (SISO) plants with dead time. With advances in digital hardware and digital control techniques, it is becoming feasible to implement control schemes, such as self-tuning control for power converters. In recent years there has been increasing interest in the development of efficient control strategies to improve the dynamic behavior of systems by using digital self-tuning controllers. The Bányász/Keviczky digital self-tuning PID controller was proposed by Cs. Bányász and L. Keviczky in 1982. This algorithm is based on the explicit identification of a second order process model with time delay [22], [23].

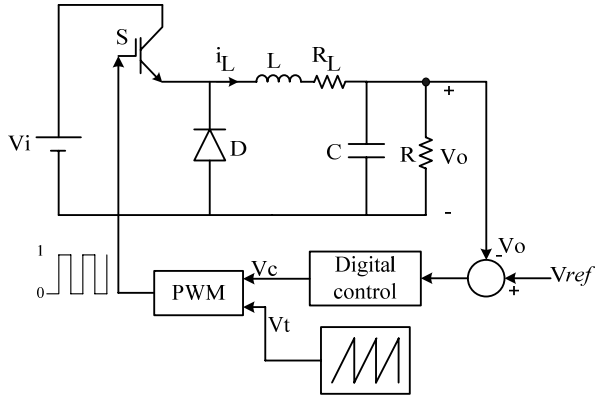


Figure1
Synoptic scheme of the VMC for the Buck converter

The Dahlin digital self-tuning PID is also used in this study. The major advantage of this algorithm is the reduction of the tuning from three to two parameters [24].

Hence, the contribution of this paper is to introduce a mathematical model using the state-space averaged (SSA) technique for the Buck converter in continuous conduction mode (CCM).

A comparative study between two digital self-tuning PID controllers, the Dahlin PID and the Bányász/Keviczky PID, is carried out.

In this paper, VMC approach for Buck converter has been analyzed and developed using these two controllers. The simulation results using a MATLAB SimPowerSytems toolbox to validate the effectiveness of the proposed control strategies are also included.

2 A State-Space Averaged Model of Buck Converter

The basic power circuit of Buck topology is shown in Figure 1. It consists of two semi conductor devices: a controlled power device, such as a power MOSFET or IGBT and an uncontrolled device, such as a power diode and passive elements. They consist basically of an inductor in series with a parallel combination of a capacitor and resistor [25].

The state space model is given by the following equation:

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{1}$$

While taking as state vector $x = \begin{bmatrix} i_L & v_o \end{bmatrix}^T$

In the case where the switch S is closed and the diode *D* is opened, state space model matrices $A_1,\ B_1,\ C_1$ during d of the switching time became:

$$A_1 = \begin{bmatrix} -\frac{R_L}{L} & -\frac{1}{L} \\ \frac{1}{C} & -\frac{1}{RC} \end{bmatrix}, B_1 = \begin{bmatrix} \frac{v_i}{L} \\ 0 \end{bmatrix}, C_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} \tag{2}$$

In the second case where the switch *S* is opened and the diode *D* is closed, state space model matrices $A_2,\ B_2,\ C_2$ exist during $1-d$ of the switching time interval:

$$A_2 = \begin{bmatrix} -\frac{R_L}{L} & -\frac{1}{L} \\ \frac{1}{C} & -\frac{1}{RC} \end{bmatrix},\ B_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ C_2 = \begin{bmatrix} 0 & 1 \end{bmatrix} \tag{3}$$

Combining both set of matrix by using the equation (1) we obtain:

$$\dot{x} = \left[A_1 d + A_2 \left(1-d\right) \right] x\left(t\right) + \left[B_1 d + B_2 \left(1-d\right) \right] u\left(t\right) \tag{4}$$

The nonlinear equation (4) leads to:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -\frac{R_L}{L} & -\frac{1}{L} \\ \frac{1}{C} & -\frac{1}{RC} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \frac{d}{L} \\ 0 \end{bmatrix} v_i \tag{5}$$

The SSA technique introduces a small AC modulating signal represented by "^" in the DC steady-state quantities (which are represented by the upper case letters) [26].

Therefore:

$$d = D + \hat{d} \tag{6}$$

$$x = X + \hat{x} \tag{7}$$

$$v_i = V_i + \hat{v}_i \tag{8}$$

Inserting equations (6) through (8) in equations (4) and recognizing that in steady state, $\dot{X} = 0$. Moreover in the AC equation, terms containing products of $\hat{x}$ and $\hat{d}$ can be neglected (small multiplied by small gives an even smaller result). Based on these facts, the DC and AC equations can be obtained as follows:

DC equation:

$$0 = A_0 X + B_0 V_i \tag{9}$$

AC equation:

$$\dot{\hat{x}} = A_0 \hat{x} + B_0 \hat{v}_i + E\hat{d} \tag{10}$$

Where:

$$A_0 = \left[A_1 D + A_2 \left(1-D\right) \right] \tag{11}$$

$$B_0 = \left[B_1 D + B_2 \left(1-D\right) \right] \tag{12}$$

$$E = \left[\left(A_1 - A_2\right) X + \left(B_1 - B_2\right) V_i \right] \tag{13}$$

Back to the notation of equations (6) through (8), we finally have:

$$\dot{\hat{x}} = A_0 x + B_0 v_i + E\hat{d} \tag{14}$$

Combining the matrix according to equation (14) gives:

$$\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \end{bmatrix} = \begin{bmatrix} -\frac{R_L}{L} & -\frac{1}{L} \\ \frac{1}{C} & -\frac{1}{RC} \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + \begin{bmatrix} \frac{D}{L} \\ 0 \end{bmatrix} \hat{v}_i + \begin{bmatrix} \frac{V_i}{L} \\ 0 \end{bmatrix} \hat{d} \tag{15}$$

From equation (15), we obtain a set of two linearized equations corresponding to the small-signal model:

$$\dot{\hat{x}}_1 = -\frac{R_L}{L}\hat{x}_1 - \frac{1}{L}\hat{x}_2 + \frac{D}{L}\hat{v}_i + \frac{\hat{d}}{L}V_i \tag{16}$$

$$\dot{\hat{x}}_2 = \frac{1}{C}\hat{x}_1 - \frac{1}{RC}\hat{x}_2 \tag{17}$$

D represents the static duty cycle given by $D = V_{out}/V_{in}$

The transfer function relating the output voltage to the duty cycle is given by:

$$\frac{\hat{x}_2(s)}{\hat{d}(s)} = \frac{V_i}{LCs^2 + \left(CR_L + \frac{L}{R}\right)s + 1 + \frac{R_L}{R}} \tag{18}$$

The transfer function of the system with a zero-order hold is given by:

$$\frac{\hat{x}_2(z)}{\hat{d}(z)} = Z\left\{\frac{1-e^{-sT_s}}{s}\frac{\hat{x}_2(s)}{\hat{d}(s)}\right\} = \left(1-z^{-1}\right)Z\left\{\frac{1}{s}\frac{\hat{x}_2(s)}{\hat{d}(s)}\right\} \tag{19}$$

$$\frac{\hat{x}_2(z)}{\hat{d}(z)} = \frac{B\left(z^{-1}\right)}{A\left(z^{-1}\right)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{20}$$

Where:

$$\begin{aligned} b_0 &= \frac{V_i}{LC}\left[A_1 + A_2 + A_3\right] = 0 \\ b_1 &= \frac{V_i}{LC}\left[-\left(A_1 + A_3\right)e^{s_1 T_s} - \left(A_1 + A_2\right)e^{s_2 T_s} - A_2 - A_3\right] \\ b_2 &= \frac{V_i}{LC}\left[A_1 e^{(s_1+s_2)T_s} + A_2 e^{s_2 T_s} + A_3 e^{s_1 T_s}\right] \\ a_1 &= -e^{s_1 T_s} - e^{s_2 T_s} \\ a_2 &= e^{(s_1+s_2)T_s} \end{aligned} \tag{21}$$

And:

$$\begin{cases} A_1 = \dfrac{1}{s_1 s_2} \\ A_2 = \dfrac{1}{s_1 (s_1 - s_2)} \\ A_3 = \dfrac{1}{s_2 (s_2 - s_1)} \end{cases} \tag{22}$$

s_1, s_2 represent the poles of the system.

T_s represents the sampling time.

3 Voltage Mode Control

In the DC-DC converter, the output supply must be adjusted to be constantly equal to a fixed value; considering that the input supply and load can vary, this control is called VMC (Voltage Mode Control). In this way the link between the input and the output of the system can be modified by a controller addition, which is generally calculated according to certain criteria defining the type of desired response for the system in closed loop (stability, response time, …).

Figure 1 shows the synoptic scheme of the VMC for the Buck converter, the control signal v_c fed into a pulse-width modulator (PWM) that compares v_c with the signal v_t (saw wave). The modulator produces a switched voltage waveform that controls the gate of the semiconductor S. The duty cycle d of this waveform is proportional to the control voltage v_c.

4 Digital Self-Tuning Controller

A digital self-tuning controller is a controller that during each sample interval performs three major steps shown in Figure 2.

- Estimates the parameters of the discrete plant model.

- Calculates the controller parameters using the estimated plant model parameters.

- Calculates and implements the new control signal.

In this work, two digital self-tuning controllers based on recursive least-squares estimation are used to improve the robustness of the controlled system.

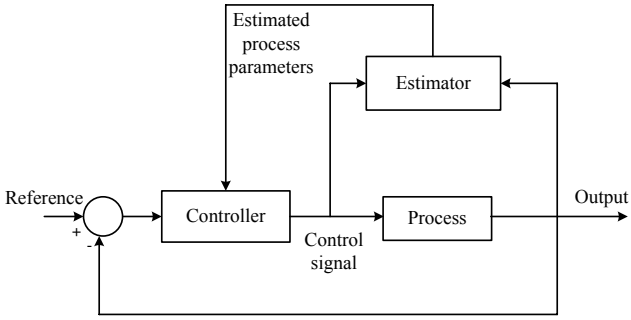


Figure 2
Block diagram of control loop for the Digital Self Tuning system

4.1 Digital Self-Tuning Bányász/Keviczky PID Controller

Discrete PID regulators can be implemented in many different ways. Different structures correspond to different continuous PID regulators. The most common sampled data PID regulator is given by the discrete transfer function:

$$C(z) = \frac{U(z)}{E(z)} = \frac{Q(z^{-1})}{P(z-1)} \tag{23}$$

$$Q(z^{-1}) = q_0 + q_1 z^{-1} + q_2 z^{-2} \tag{24}$$

$$P(z^{-1}) = 1 + p_1 z^{-1} + p_2 z^{-2} \tag{25}$$

Where q_0, q_1, q_2, p_1 and p_2 are the controller parameters.

The controlled process is described by the discrete transfer function:

$$G_p(z) = \frac{B(z^{-1})}{A(z^{-1})} = \frac{b_1(1+\gamma z^{-1})}{1 + a_1 z^{-1} + a_2 z^{-2}} z^{-d} \tag{26}$$

With:

$\gamma = \frac{b_2}{b_1}$, where $b_1 \neq 0$ and $d > 0$ is the discrete time delay of the process.

Since the process is stable and a second order model is used, a very good robust design idea is to choose $Q(z^{-1})$ proportional to the denominator of the estimated process model:

$$Q\left(z^{-1}\right) = q_0\left(1 + a_1 z^{-1} + a_2 z^{-2}\right) = q_0 A\left(z^{-1}\right) \tag{27}$$

Which means that

$$q_1 = q_0 a_1, \quad q_2 = q_0 a_2 \tag{28}$$

Which in turn means all poles are canceled and that it is applicable for stable processes only. In practical tuning, this means that the regulator cancels the two largest time constants in the process dynamics [26], and this idea allows us to simplify the control loop of the system:

$$C\left(z^{-1}\right)G_p\left(z^{-1}\right) = \frac{k_I\left(1+\gamma z^{-1}\right)}{1-z^{-1}} z^{-d} \tag{29}$$

Where the integrator gain is

$$k_I = q_0 b_1 \tag{30}$$

The coefficients of the controller are given by:

$$\begin{aligned} \gamma &= \frac{b_2}{b_1} \\ q_0 &= \frac{k_I}{b_1} \\ q_1 &= q_0 a_1 \\ q_2 &= q_0 a_2 \\ p_1 &= -1 + \gamma \\ p_2 &= -\gamma \end{aligned} \tag{31}$$

$$\text{For } \gamma = 0 \rightarrow \; k_I = \frac{1}{2d-1} \tag{32}$$

$$\text{For } \gamma > 0 \rightarrow \; k_I = \frac{1}{2d(1+\gamma)(1-\gamma)} \tag{33}$$

if $\gamma < 1$, equation (32) is used together with a serially connected digital filter, which is easy to be realized in digital control systems [27]:

$$G_F\left(z\right) = \frac{1}{1+\gamma z^{-1}} \tag{34}$$

From Equation (23) the control law can be written as follows:

$$U\left(z\right) = \frac{Q\left(z^{-1}\right)}{P\left(z^{-1}\right)} E\left(z\right) \tag{35}$$

By inserting polynomials (24) and (25) into equation (35), the relations that calculate the controller output become:

$$u(k) = q_0 e(k) + q_1 e(k-1) + q_2 e(k-2) - p_1 u(k-1) - p_2 u(k-2) \tag{36}$$

4.2 Digital Self-Tuning Dahlin PID Controller

The velocity PID control algorithm is represented by the equation:

$$u(k) = u(k-1) + K_p \left[e(k) - e(k-1) + \frac{T_s}{T_i} e(k) + \frac{T_d}{T_s} \left(e(k) - 2e(k-1) + e(k-2) \right) \right] \tag{37}$$

In the adaptive control scheme, the controller tuning parameters K_p, T_i, and T_d are computed directly from the parameters of the second-order model of the process by the controller synthesis method proposed by Dahlin [20], [11]. The adapter formulas developed by this method for a process modeled by equation (19) require that this controller uses parameters estimation vector in the form:

$$\Theta^T(k) = [a_1, a_2, b_1] \tag{38}$$

And the parameter b_2 must be forced to zero by the estimator. Assuming this to be the case, the following relationships have been developed for the tuning parameters of the PID algorithm:

$$\begin{aligned} K_p &= -\frac{(a_1 + 2a_2)Q}{b_1} \\ T_i &= -\frac{T_s}{\frac{1}{(a_1 + 2a_2)} + 1 + \frac{T_d}{T_s}} \\ T_d &= \frac{T_s a_2 Q}{K_p b_1} \end{aligned} \tag{39}$$

For a PI algorithm, a_2 must also be forced to zero so that $T_d = 0$.

The variable Q is defined by:

$$Q = 1 - e^{-T_s/B} \tag{40}$$

Where B is the tuning factor that represents the time constant of the desired closed-loop response. The smaller value of B leads to a faster response of the closed control loop.

The relation that calculated the controller output is given by equation (36).

To identify the unknown coefficients of the controller, equation (36) is equal to equation (37) as follows:

$$\begin{aligned} q_0 &= K_p\left(1+\frac{T_s}{T_i}+\frac{T_d}{T_s}\right) \\ q_1 &= -K_p\left(1+2\frac{T_d}{T_s}\right) \\ q_2 &= K_p\frac{T_d}{T_s} \\ p_1 &= -1,\ p2 = 0 \end{aligned} \tag{41}$$

5 Simulation Results

In order to compare the performance of these two digital PID self-tuning regulators using a VMC control methodology of the Buck converter, some simulation results are given with the system parameters:

$V_i = 15volts$, $L = 3.716mH$, $R_L = 0516\Omega$, $C = 100\mu F$, $R = 7.5\Omega$. Switching frequency $f = 20KHz$.

The parameters of the digital self-tuning Bányász/Keviczky PID Controller are:

Dead time $d = 5$, $T_s = 10^{-5}$.

The parameters of the design of the digital self-tuning Dahlin PID are:

Adjustment factor $B = 9.10^{-4}$, $T_s = 10^{-6}$.

Figure 3 shows the output voltage response obtained by the digital self-tuning controllers Bányász/Keviczky PID and Dahlin PID. It is clear that the output voltage response obtained via the Bányász/Keviczky PID is faster than that obtained by the Dahlin PID. It is found that spikes occur in the output voltage controlled by the Bányász/Keviczky PID more than in that obtained with the Dahlin PID.

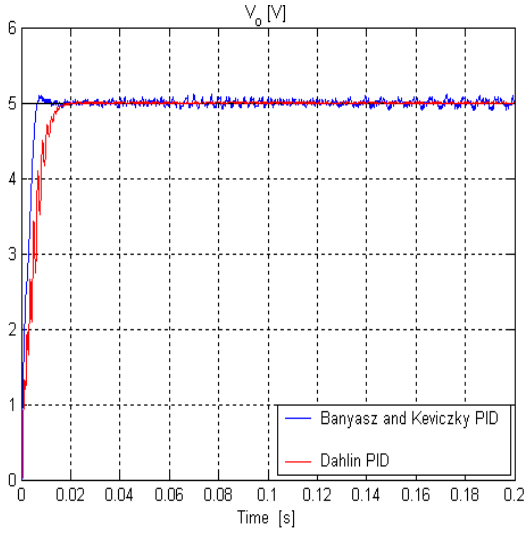


Figure 3
Output voltage responses using the two digital self-tuning controllers

To verify the effectiveness of the proposed controllers, first the system responses are driven initially with $15volts$ as the input voltage when a step change in the input voltage from $15volts$ to $23volts$ is applied at $t = 0.1s$ as can be seen from Fig. 4.

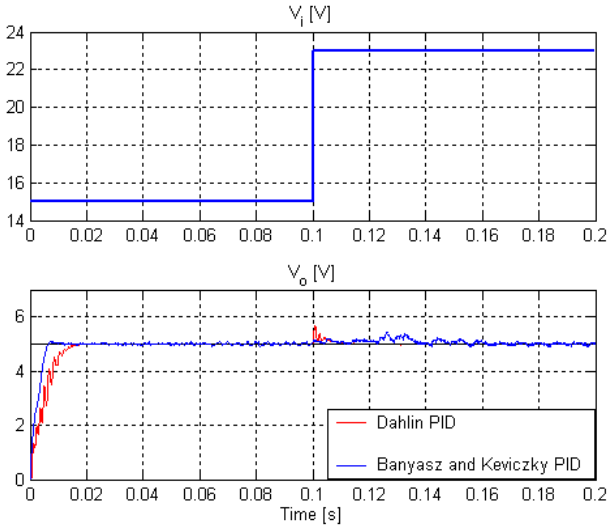


Figure 4
Output voltage responses under step input voltage change from $15volts$ to $23volts$ at $t = 0.1s$

Figure 5 shows the zoom of the output voltage response of Figure 4 obtained by the digital self-tuning controllers Bányász/Keviczky PID and Dahlin PID. It is clear that the voltage increase is recovered more quickly with the self-tuning digital Dahlin PID controller for this sudden change in the input voltage than with the Bányász/Keviczky PID.

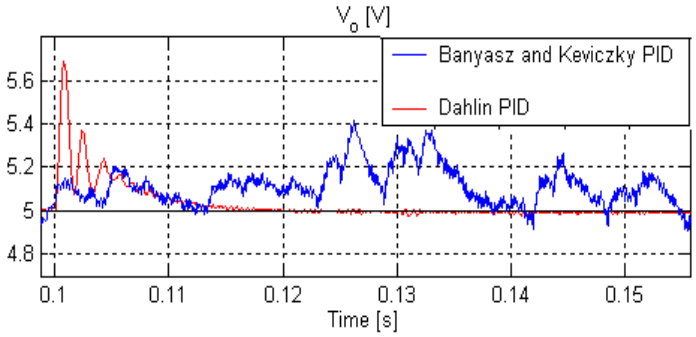


Figure 5
Zoom of the output voltage response of the Figure 4

To check the hardiness of the digital self-tuning Dahlin PID controller, a very important change in the input voltage from *15 volts* to *30 volts* is applied at $t = 0.1s$ as can be seen from Figure 6. It is clear that the output voltage response obtained by the Dahlin PID is faster than that obtained by the Bányász/Keviczky PID.

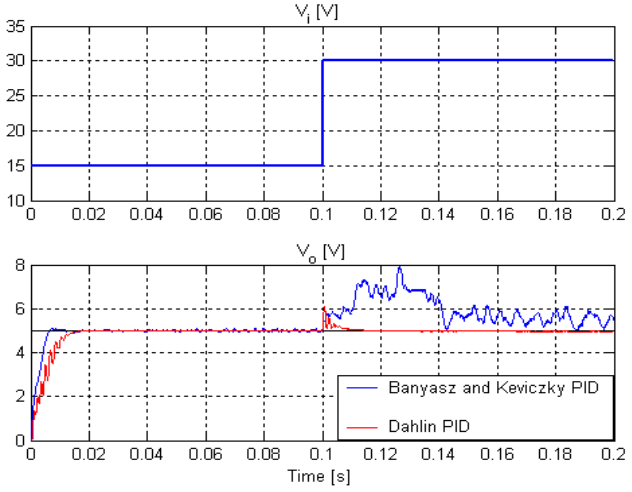


Figure 6
Output voltage responses under step input voltage change from $15volts$ to $30volts$ at $t = 0.1s$

Finally, we examine a crucial aspect of the controllers' operation, namely the system's behaviour against low load current. The load resistor is increased from its nominal value 7.5Ω to 10Ω at $t = 0.1s$. The simulation results in Figure 7 show that the voltage drop is recovered more quickly with the self-tuning digital Dahlin PID controller for this sudden change in the load than with the Bányász/Keviczky PID.

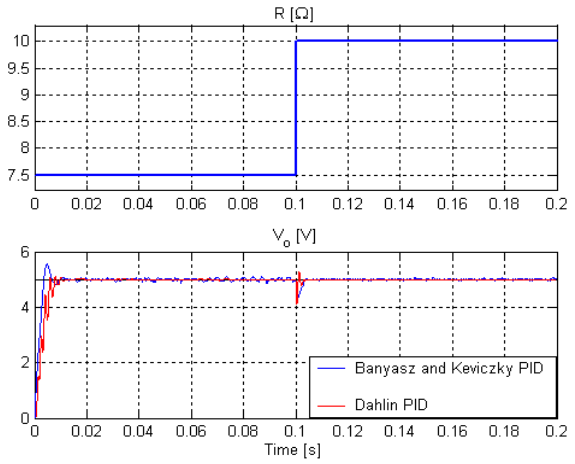


Figure 7

Output voltage response under step load change from 7.5Ω to 10Ω at $t = 0.05s$

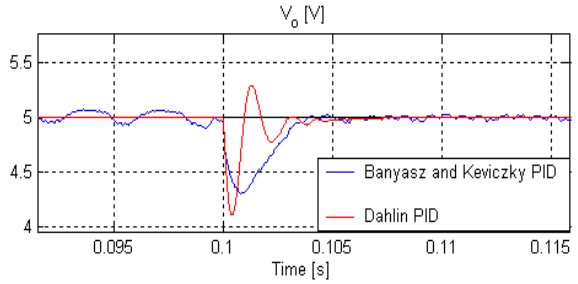


Figure 8

Zoom of the output voltage response of Figure 7

Conclusion

In this paper the state-space averaged technique to derive the small-signal model of the DC-DC Buck converter is made.

Tow digital self-tuning Bányász/Keviczky PID and Dahlin PID controllers based on recursive least-squares estimation are designed to regulate the output voltage using a VMC strategy.

Comparative studies were made with these two digital self-tuning PID controllers for a sudden change in input voltage magnitude and/or load change. The digital self-tuning Dahlin PID controller gives the better performance and is more robust for model inaccuracies and disturbances in comparison with the Bányász/Keviczky PID controller.

Simulated results obtained with a MATLAB SimPowerSytems toolbox validate the effectiveness of the proposed control strategy.

References

[1] S. Ang A. Oliva. Power-Switching Converters. Taylor and Francis, Group, 2005

[2] M. H. Rachid. Power Electronics, Circuits, Devices, and Applications. Pearson Education, Singapore, 2004

[3] R. W. Erickson. Fundamentals of Power Electronics. New York, Chapman and Hall, 1997

[4] N. Mohan, T. Undeland and W. Robbins. Power Electronics: Converters, Applications, and Design. New York, John Wiley and Sons, 1995

[5] Robert. W. Erickson. Dc-dc power converters. Technical report, Wiley Encyclopedia of Electrical and Electronics Engineering Department of Electrical and Computer Engineering University of Colorado Boulder, CO 80309-0425

[6] R. D. Middlebrook. Power Electronics: Topologies, Modeling, and Measurement. Proc. IEEE Int. Symp. Circuits Syst., April 1981

[7] A. Ž. Rakić T. B. Petrović. Linear Robust Approach to dc/dc Converter Modelling1: Deterministic Switching, Electrical Engineering. IEEE Transactions on Industry Applications, Springer-Verlag, DOI 10.1007/s00202-003-0210-6, 2003

[8] G. Svensson U. Svanberg. State Space Controlled Buck Converter. Master's thesis, Chalmers Tekniska Högskola Institutionen för Elteknik, Göteborg, Sweden, 2003

[9] I. Dogan. Microcontroller Based Applied Digital control. John Wiley and sons, Ltd, West Sussex, England, 2006

[10] R. S. Burns. Advanced Control Engineering. Integra Software Services Pvt. Ltd., Pondicherry, India, 2001

[11] V. Bobál, J. Böhm, J. Fessl and J. Macháček. Digital Self-tuning Controllers. Springer-Verlag London Limited, 2005

[12] M. Milanovic, M. Truntic, P. Slibar and D. Dolinar. Reconfigurable Digital Controller for a Buck Converter Based on FPGA. Science Direct, Microelectronics Reliability, page 150–154, November 2006

[13] M. Zhang D. M Gorinevsky and G. A. Dumont. Tuning Feedback Controller of Paper Machine for Optimal Process Disturbance Rejection. In Control Systems'98, Porvoo, Finland, September 1998

[14] J. G. Ziegler N. B. Nichols. Optimum Settings for Automatic Controllers. Trans. ASME, Vol. 64, pp. 759-768, 1942

[15] G. H. Cohen G. A. Coon. Theoretical Considerations of Retarded Control. Transactions of the ASME, pp. 827-834, 1953

[16] C. C. Hang, K. J. Åström and W. K. Ho. Refinement of the Ziegler Nichols Tuning Formula. IEE Proceedings - D, Vol. 138, No. 2. pp. 11 1–1 18, 1991

[17] I. L. Chien and P. S. Fruehauf. Consider Imc Tuning to Improve Controller Performance. Chemical Enginnering Progress, Vol. 86, No. 10, pp. 33-41, 1990

[18] O'Dwyer. Pi and Pid Controller Tuning Rules for Time Delay Processes: a Summary. In Irish Signals and System Cofiference, NUI Gaiway. 1999

[19] C. C. Hang, W. K. Ho and L. S. Cao. A Comparison of Two Design Methods for Pid Controllers. ISA Transactions, Elsevier, 33, pp. 147-151, 1994

[20] E. B. Dahlin. Designing and Tuning Digital Controllers. Instrum. Control Systems, Vol. 41, pp. 77-83, 1968

[21] J. D. Higham. Single Term Control of First and Second Order Processes with Time Delay. Control, pp. 136-140, Feb. 1968

[22] Cs. Bányász L. Keviczky. Direct Methods for Self-Tuning Pid Regulators. 6th IFAC Symp. on Ident. and Syst. Par. Est., Washington D.C. (USA), pp. 1249-1254, 1982

[23] Cs. Bányász J. Hetthéssy and L. Keviczky. An Adaptive Pid Regulator Dedicated for Microprocessor-based Compact Controllers. 7th IFAC Symp. on Ident. and Syst. Par. Est., York (UK), pp. 1299-1304, 1985

[24] A. B. Corripio, P. M. Tompkins. Industrial Application of a Self-Tuning Feedback Control Algorithm. ISA Transactions, Vol. 20, No. 2, pp. 3-10, 1981

[25] F. Tahri, A. Tahri and S Flazi. Digital Self-tuning Control of DC-DC Buck Converter. ICEL'09 Quatrième Conférence Internationale sur l'Electrotechnique, Oran-Algerie, Nov. 10-11, 2009

[26] C. P. Basso. Switch-Mode Power Supplies Spice Simulations and Practical Designs. McGraw-Hill, 2008

[27] Cs. Bányász L. Keviczky. Pid Regulator Tuning for Factorable Nonlinear Plants. 10th Mediterranean Conference on Control and Automation - MED2002, Lisbon, Portugal, July 9-12, 2002

Biogas and Energy Production by Utilization of Different Agricultural Wastes

Attila Meggyes

Budapest University of Technology and Economics
Department of Energy Engineering
Műegyetem rkp. 3, H-1111 Budapest, Hungary
e-mail: meggyes@energia.bme.hu

Valéria Nagy

University of Szeged, Faculty of Engineering
Moszkvai krt. 9, H-6725 Szeged, Hungary
e-mail: valinagy78@mk.u-szeged.hu

Abstract: Sustainable agricultural development and increasing the rate of renewable energy sources have become an economic issue after Hungary joined the EU. Under the present economic conditions the private sector cannot solve in its complexity the problems of environment protection and energy from its own sources. The paper introduces biogas production and utilization methods that are suitable for providing continuous operation of existing biogas plants and also for determining the parameters of establishing biogas plants. Experimental variants (mixtures of liquid pig manure and plant additives) were developed to produce biogas and intensify biogas yield, and then gas engine tests were done for the energy utilization. The eco-energy system can be built up by taking into consideration the specific local conditions. It does not require any change or transformation of agricultural structure. The system can be expanded by the utilization of other organic materials, so it supports efficient operation. Furthermore, it can be the pillar of energy independence of rural life, because during the establishing of the eco-energy system, ecological aspects are taken into consideration, which makes this system sustainable. Waste disposal requirements can be integrated, too. We created a complex biogas production and utilization system by developing variants, so that both the energy and the waste disposal goals can be achieved together. This system was presented as an alternative agricultural system for an animal farm.

Keywords: biogas production; biogas utilization; energy; complex system; environment; waste disposal

1 Introduction

Producing and utilizing renewable energy – both in a global and a national context – is necessitated by the synergistic effect of climate change and the long term, continuous price rise of fossil fuels. The main reasons for the spread of renewable energy sources are to increase the security of the energy supply or, in optimal case, to realize total energy independence. Our paper deals with production and utilization of biogas for energy. The importance of biogas is, in addition to energy aspects, justified environmentally by EU requirements and by economic considerations, because conservation of the state of our environment and efficient, economically satisfying energy demands can be solved by the harmonized application of traditional and renewable energy sources. It is necessary to create a complex biogas production and utilization system for energy while we focus on the environment and waste disposal. The centre of the system must be integrated waste-management and environmental energy utilization. The subject of this research is: how to meet collectively the energy and waste disposal requirements without transformation of the system.

2 Research Task, Object

This paper presents the complexity of the production and utilization of biogas. The research work was done at the Budapest University of Technology and Economics, Department of Energy Engineering, and at Szolnok University College, Technical and Machinery Department. In the course of our research work, we supposed that there is an energy-producing and energy-utilizing technology which can be suitable for the circumstances and initial conditions. The objective of the research task was to support the approach that it is necessary to examine the biogas producing and utilizing technological processes together as a complex system. It is needed to analyze these, considering that the principle of the complex optimization focuses just on the environment and waste disposal. The tasks performed during the research were:

- Proving with experiments the yield-increasing and quality improving effects of different kinds of plant additives added to pig manure in the fermenters (mesophylic bioreactors).
- Testing the utilization of biogases from different kinds of liquid pig manure and additives for energy gaining in gas engines, with particular regard to the emission.

Environment-friendly utilization of organic pollution materials and energy production can be realized together by biogas production and utilization. We showed through a specific example how to develop a pig farm to a steady energy supply and waste disposal unit by treatment of liquid pig manure in a pig farm.

3 Scientific Background

3.1 Biomass-based Energy Production – Biogas

Energy is a complex system; hence, energy-production and energy-conversion require systemic thinking, for which firstly a change of aspect is necessary. The primary view-point is to satisfy the energy demands with the lowest possible stress on the environment. Furthermore, ecological thinking should prevail increasingly during the planning and operating of the different kinds of technical equipment and facilities [1]. It can be determined that the biomass-based energy system can mean the necessary transformation of the energy structure. Researchers of several domestic and foreign universities and research institutes have specialized in the feasibility of biogas production from different kinds of biomass. Braun [2] examined the types and degradation features of basic material, while Llabrés and Müller et al. and Borbély [3, 4, 22] examined the anaerobe degradation of different kinds of substrates based on liquid pig manure and examined the output-increasing effect of pre-treatment processes and the kinetics of cellulose. Gunaseelan, Lehtomäki et al, Mata et al and Panichnumsin et al [5-8] studied the fermentation of manures and different kinds of plant additives. In their experiments, the positive synergistic effects created the possibility of higher methane productivity. They established that the top of sugar-beet doping causes higher hydrogen-sulphide content in biogas and that increasing the ratio of additives causes a decrease methane productivity. Houdková et al built a laboratory fermentation unit for the experimental determination of biogas production [9]. Kalyuzhnyi et al [10] studied the integrated mechanical, biological and physical-chemical treatment of liquid pig manure. He created a mathematical model of anaerobe decomposition and revealed and described the principal controller factors with numerical experiments. Meanwhile Misra [23] also emphasizes the elaboration of a practical model for empirical validation.

3.2 Utilization of Biogas in Gas Engines

Biogas is a gaseous matter similar to natural gas and can be utilised in many ways. Biogas has different combustion and compositional characteristics compared to natural gas, so it needs a different system of preconditions compared to the combustion of natural gas [11]. One way to use biogas is its utilization in internal combustion engines. The basic national research institutes for combustion engineering examined the feasibility of the combustion of low heat value gases with inert content – among them biogases, the technical and economic effect of their application, and the combustion properties of biogas [12, 13]. In the 1960s, Sándor [14] undertook experiments to verify the necessity of gas-engines. Neyeloff-Gunkel [15] specialized in the modelling and simulation of the

combustion of biogas. Porpatham [16] examined how bio-fuel from biomass can be applied as a fuel in internal combustion engines. They tested the operation and the emission of a one-cylinder, four-stroke engine fuelled by natural gas, biogas and their mixtures.

Huang-Crookes [17] did experiments on a one-cylinder, four-stroke, spark ignited engine at constant speed, using increasing compression-ratios, given CO_2 content (37.5%) and air access ratio (0.97). The increasing compression-ratio caused intensively increasing NO_x and HC emission. Crookes [18] made further tests with changing CO_2 content. He determined that increasing the CO_2 content results in a decrease in NO_x emission, which can be due to decreasing combustion speed and combustion peak-temperature.

4 Methods and Results of the Research Work

Through the simultaneous presentation of the research into biogas production and utilization, we wished to present the connection between agriculture and energy; we looked at the waste-problem of pig-farms with a few hundred pigs and the possibility of the treatment of liquid pig manure and, further, to the feasibility of building a stable energy-production unit.

The results of our research work, undertaken at the Mezőtúr Campus of Szolnok College, demonstrate the quantity and composition of the generated biogas via fermentation of liquid pig-manure and different kinds of additives. The results reveal further how the fermenters can be operated to produce the proper quantity of biogas with a composition that complies with added utilization potentialities in plants with similar technology.

In the course of biogas utilization in spark-ignited internal combustion engines, information can be collected as to the effects of biogases from different kinds of basic materials and additives on the operation of gas-engines, considering particularly the emission.

4.1 Methods and Results Concerning the Biogas Production Researches

At the Mezőtúr Campus of Szolnok College, we undertook biogas producing experiments based on liquid pig-manure to develop variants for intensifying biogas yield. The task performed during the research were:

- Planning of the biogas producing experiments, and the preparation of an applied fermentation technology and unit.
- Constructing and continuously developing the instruments for the experiments.

- Undertaking biogas producing experiments for the energy utilization of biogas.
- Evaluating the results (quantity, composition, energy content, etc.) of the biogas producing experiments.

We created 30 experimental variants using different kinds of plant additives to produce biogas based on liquid pig manure. The dry matter content of the liquid pig manure was 4%, and our experiments lasted 43-50 days. The increase did not influence quality (methane content), but methane stability depended on the additives. *Table 1* shows the variants with which biogas can potentially be made in the proper quantity and quality for energy production.

Table 1
Chosen variants

signs of variants	***pig manure***	***bacteria***	***sweet sorghum press residue***	***fruit marc*** (different ratio)	***maize marc***	***average biogas production*** *[dm^3/kg organic dry matter]*	***methane content*** *[%]*
Variant 1	+	–	+	–	–	417	52.0-59.0
Variant 2	+	+	–	–	+	589	52.2-59.5
Variant 3	+	–	–	+ (50%)	–	512	62.5-74.9
Variant 4	+	–	–	+ (25%)	–	453	66.8-77.1

We have developed variants to produce biogases beneficial not only from the point of view of energy but also from that of waste disposal as well. The concrete conclusions of our biogas production experiments can be summarized as below:

- Some additives (fruit marc) have acidic pH, so they can be applied only under certain conditions and in limited quantities in biogas systems.
- The additives which contain volatile organic acids can be yield-increasing, because of the biomass containing 50% (Variants 3) and 25% (Variants 4) alcohol waste (fruit marc), which means a potential biogas producing method which takes into consideration both the quantity of the produced biogas and the methane content of biogas.
- The applied variants can provide favourable conditions for producing biogas, and, simultaneously, waste disposal can also be realized via biogas-production.
- Maize-marc and fruit-marc as additives have the effect of an increased methane production, coupled with a stable gas composition.
- The methane production of the given variants satisfies the conditions of utilization, and so the heat engines can operate properly.

4.2 Methods and Results of Biogas Utilization Researches

We undertook research work in the György Jendrassik Heat Engineering Laboratory at the Budapest University of Technology and Economics, Department of Energy Engineering. The objective of the gas engine tests was to get to know how biogases – as the different kind of experimental variants – influence the operation of gas engines. The experimental gas engine is not a special biogas engine but rather a conventional natural gas engine.

Figure 1 shows the schematic diagram of the engine test set-up. The main parts of the test engine are:

- 24.6 kW, 4 cylinder Wisconsin Motors Continental TM27 type gas engine
- 26.4 kW, 4 pole Marelli CX IM B3 180M type asynchronous dynamometer
- controller box (starter button, mode switch, locking switch etc.)
- indication system
- emission analyzer
- data collecting system

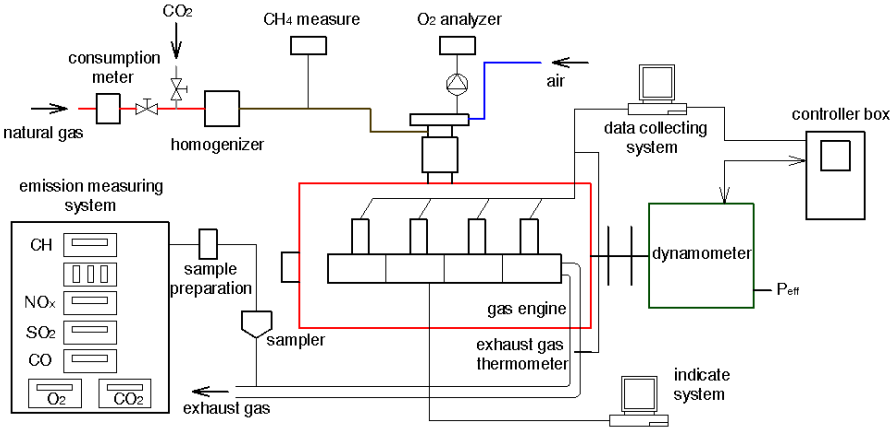


Figure 1
The engine test set-up

The tasks performed during the research were:

- Planning of biogas utilization tests and choosing the applied technology.
- Undertaking biogas utilization tests with the experimental gas engine set-up at disposal.
- Processing the results of the biogas utilization tests.

In the course of choosing the applied fuel to operate heat engines we must take into consideration, in every case, that environmental science, which has an interdisciplinary character, also involves certain areas of the energy. One way to use biogas is in internal combustion engines. The objective of our research work is to examine what effects biogases have on the operation of gas engines and to choose biogases via which both the operation of the gas engine is suitable and the utilization and waste disposal can be realized together.

Out of the biogases produced for the experiments, we chose – considering the yield and the methane content – some optimal appearing biogases (Variants 1, 2, 3, 4). We represented the composition of the biogases with substitute gases (a variable mixture of natural gas and carbon-dioxide). Natural gas contains ~96% methane [21], but the composition of biogases is methane and carbon-dioxide; in addition, they have a changing composition. During our engine laboratory tests, we produced a gas mixture with constant composition almost similar to a biogas variety and we undertook tests with them because variable composition complicates the tests and makes generally valid conclusions uncertain. Ignition time and ignition advance were not changed. During our experiments, we assured that the engine was able to perform with different ranges of biogas composition. Thus, it can designate all the possible ranges, but in the case of a variable composition of biogas, it is more difficult to perform the tests.

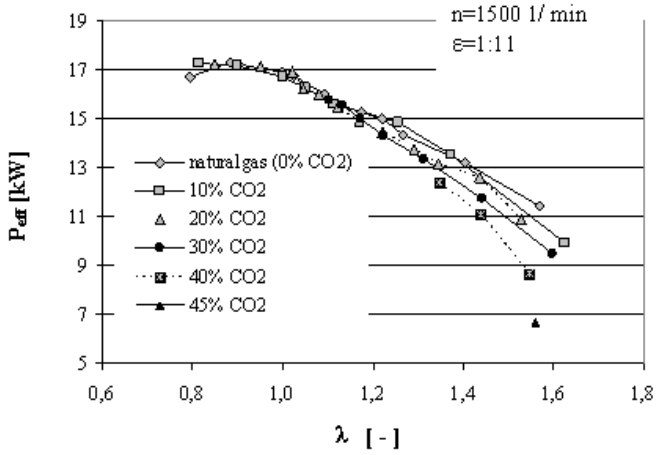


Figure 2
Effective performance

Figure 2 shows the effective performance diagrams. If the engine is fuelled with gas mixtures (with 10-20% carbon-dioxide content) in the range of λ=0.8-1.1 air access ratios, it is able to transmit almost the same values of the effective performance as in the case of operation with natural gas. If the engine is fuelled with gas mixtures with 30% carbon-dioxide content in the range of λ=1.1-1.2 air

access ratios (the gaseous consumption increased though), it is able to produce values of the effective performance similar to natural gas operation. In the range of λ>1.2 air access, the values of the effective performance coefficient fall below the values of natural gas due to the influence of the increase in carbon-dioxide. As the methane content of the biogas decreases, the effective performance values decrease (by 10-15%). We dealt with the operation of the engine in our previous paper in details [19].

Fig. 3 – fig. 6 show the partial results of measurement. Our results were analyzed for energy generation and waste disposal aspects.

In the figure below (*Fig. 3*) it can be seen that in the case of λ>1.1 air access ratios, the cooling effect of the surplus air results in lower NO_x emission. The engine operation with increasing carbon-dioxide content (and thus decreasing methane content) of the gas mixture – on account of delaying of the combustion and the cooling effect of carbon-dioxide – results in a further decrease.

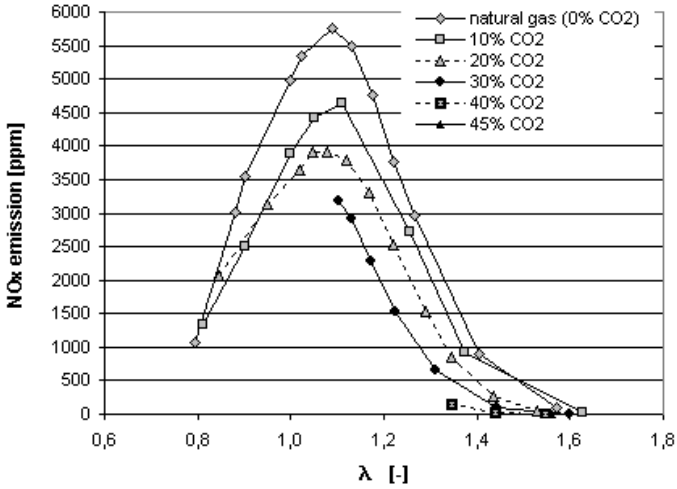


Figure 3
NO_x emission

However, while the methane content of the biogas decreases, the carbon-dioxide content increases at the same time. This means that excess feeding from the biogas with lower methane content is needed for just the same quantity of methane. The quantity of carbon-dioxide fed into the engine with the fuel increases, which appears also in the exhaust gas, producing significantly increased CO_2 (*Fig. 4*).

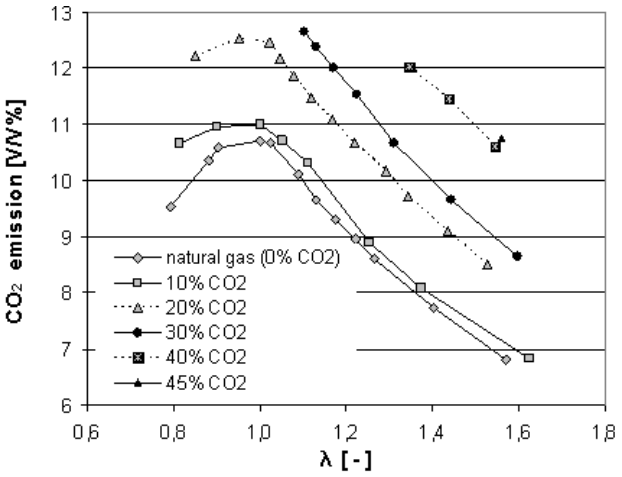


Figure 4
CO_2 emission

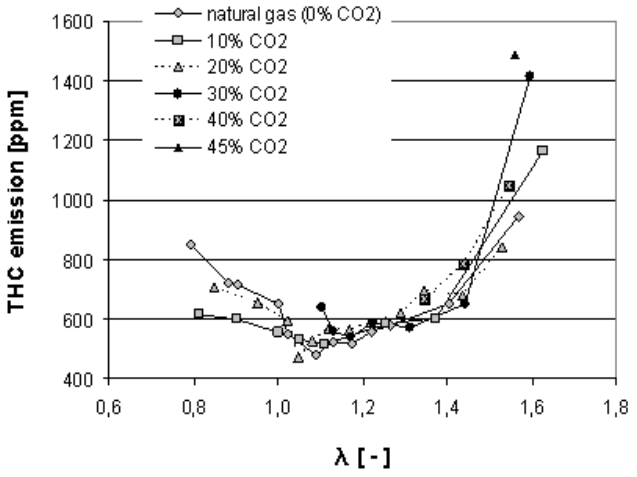


Figure 5
THC emission

With an increase in the carbon-dioxide content of the applied energy carrier, the combustion conditions worsen, which causes a higher unburned hydrocarbons content in the exhausted gas. There is no significant difference between the operation of gas engines fed with natural gas and with gas mixtures with a higher carbon-monoxide content (with lower methane content) in the range of λ=1.2-1.5 air access ratio (*Fig. 5*).

It is noticeable in the following figure (*Fig. 6*), that in case of $\lambda<1.0$ air access ratios, the CO emission increases by leaps and bounds, which can be explained by the production of a richer mixture. However, in range of $\lambda=1.1\text{-}1.4(1.5)$ air access ratios, CO emissions – independently of the carbon-dioxide content of the gas mixture – are stabilized at much lower values. In the case of $\lambda>1.4(1.5)$ air access factors, the drag on the combustion results in increased CO emission. Thus, concerning CO emission, it can be unambiguously determined that for a traditional gas engine operated with a gas mixture with low methane content, there is no effect on CO emission if the gas engine is operated permanently within the range of $\lambda=1.1\text{-}1.4$ air access factors.

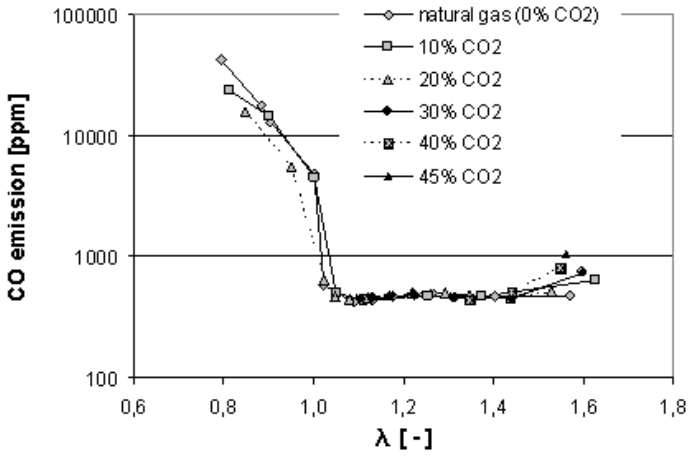


Figure 6
CO emission

The results of tests done with biogases with different methane contents show that biogas with 60-72% methane content – as the result of developed variants – can be combusted in a traditional gas engine. When the methane content of the biogas decreases, the operating range of the gas engines narrows and will shift towards the higher air access ratio (1.2-1.6), the effective performance values decrease (by 10-15%), and the efficiency slightly (by 2-4%) worsens.

We determined by emission tests that the utilization of biogases generated on the basis of our variants results in lesser emission in the operation range of $\lambda=1.2\text{-}1.6$ of the gas engines. The NO_x emission decreases as much as (20-50%), while the CO and THC emissions practically do not change. The increased CO_2 emission can be explained by the CO_2 content of the applied biogas, but the CO_2 content of the biogas is of biological origin.

The concrete conclusions of our biogas utilization tests can be summarized as below:

- The engine, operated with biogas with 70% methane content within the range of λ=1.1-1.2 air access ratio (though with increased gas consumption) is able to produce values of effective performance similar to natural gas operation.
- At higher air access ratios, the effective output and the efficiency would decrease significantly, due to the influence of increasing carbon-dioxide.
- In the course of the operation of the gas engines, in the case of lean burn, the higher CO_2 content of different biogases positively influences the NO_x emission.
- The increasing CO_2 emission is due to the CO_2 content of the biogas. On the other hand, the CO and THC emissions rise suddenly at insufficient air access or at large air-excess.
- A continuous operation in the range of λ=1.2-1.6 air access ratios results in lower emission in total.

5 Complex Energy Production and Waste Disposal

The production and utilization of renewable energy sources are justified not only by energy, political, environmental and competitive aspects, but by rural development aspects as well. Therefore, in our paper we present research results which promise to help in these areas and to provide for the smooth and effective operation of already existing and operating energy producing establishments, as well as waste disposal biogas facilities. The merit of the research work is that we have verified the necessity of the complex ecological aspects via complex biogas production experiments and via biogas utilization tests in gas engines. Based on the results of our research, it can be stated, that:

- Through anaerobe fermentation of biomass, produced from basic material and plant additives, a suitable quantity and quality of biogas can be generated for utilization in gas engines.
- When applying the above demonstrated variants, the favourable conditions for producing a potential renewable energy source – biogas – at workshop level can be created, and waste disposal can be realized simultaneously.

5.1 Complex Biogas Utilization System

We created a complex biogas production and utilization system by developing variants so that both the energy and the environmental goals can be achieved. For the sake of an optimal solution, it is necessary to analyze the two objective functions together, considering that the principle of the complex optimization focuses just on the environmentally-friendly energy utilization. If the quantity and/or quality of the input material necessary for developing variants cannot be provided, the energy output can decrease and waste disposal can be overshadowed too. In *Fig. 7* the sketch model of an energy cycle adapted to the biogas producing and utilizing system can be seen, and we present this system as alternative agricultural system of an animal farm.

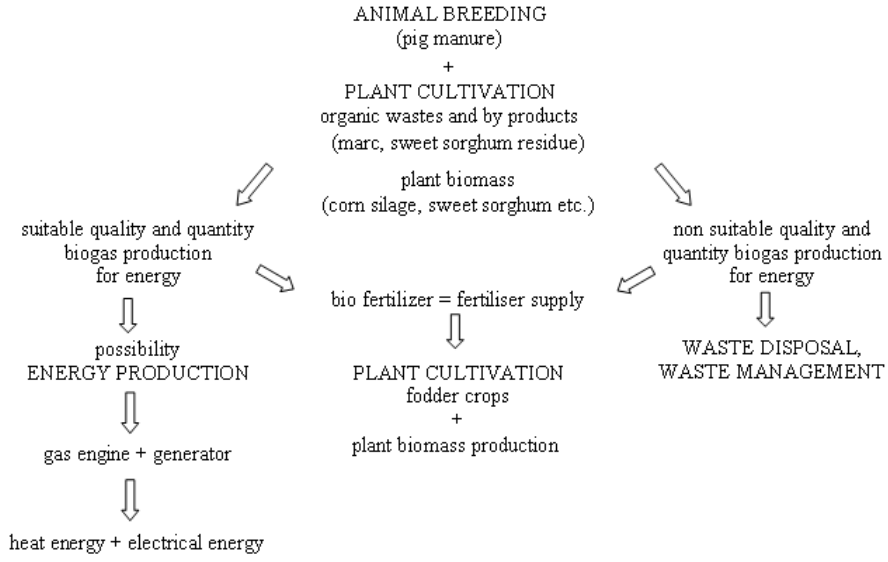


Figure 7
Diagram of biogas producing and utilizing farm system

The centre of the complex system is the integrated waste-management and environmental energy-utilization. Its advantages are local and global environmental results, energy production independent of external influences, and a better "population retention capacity" for rural areas, as well as a near optimum solution both ecologically and economically. It also realizes the achievement of being a closed circuit system.

5.2 Practical Application of the Results

We supposed that there are 500-800 pigs on a pig farm, and therefore the quantity of liquid pig manure is 1277.5 ton/year. The pig (liquid) manure is not enough alone to produce the necessary biogas quantity and fuel quality for the energy supply needs of the farm, so it is necessary to use different kinds of agricultural by-products and wastes, and it is necessary to add a biomass plant. Since biogas plants are in continuous operation, it is necessary to provide a yield enhancing organic matter in the annual production cycle. We applied agricultural by-products and wastes as organic additives during the development of the experimental variants. On the farm, the disposal options for organic wastes played an important role in selection alongside enhanced biogas yield (methane yield). *Table 2* contains the annual quantity of pig liquid manure and organic additives for developing the experimental variants.

Table 2
"Energy" from organic wastes

organic wastes for energy production	***quantity of organic wastes*** *[ton/year]*	***average quantity of produced methane*** *[m^3/day]*
pig (liquid) manure	1277.5	Σ 130.7*
fruit marc	219	
maize marc	43.8	
sweet sorghum residue	18.25	
corn silage	25.55	

** 170-250 m^3 biogas*

Most of the organic wastes in the table can be produced on the farm and are available outside the harvest period (August-November) as well, and so the operation of system can be provided for by agricultural wastes which are stored. But other organic matter (organic waste) is available through agriculture or can be obtained from the alcohol industry.

Based on the above, it can be stated that all organic matter can be biodegraded in the biogas plant, but that due to aspects of biogas production we should consider for energy production only organic matter which biodegrade rapidly and is available in sufficient quantity in the farm.

The available capacity on the farm is: 2 fermenters 115 m^3 volume/fermenter (diameter 7 m, height 3 m) and utilization of approx. 1600 ton/year multi-component biomass, which is proper ~2.22 kg organic dry matter content daily per fermenter m^3. The low heating value of the producing biogas can be determined from methane yield based on dry matter content, so the biogas production and its methane content: energy value also increases if the yield and methane ratio increase. Based on the above, ~130.7 m^3 methane is produced per day. 1 m^3

normal state methane has ~34.014 MJ/Nm^3 low heating value. In the case of Micro F22 AP type gas engine, 18 hours operation time, ~86% engine loading, ~130 m^3 methane/day (170-250 m^3 biogas per day) is needed. It can produce ~27.3 kW heat power and ~14.5 kW electric power considering that the efficiency of the machine units is ~90%. Some electric and heat energy can be used for individual needs (e.g. heating, tempering of the fermenters, etc.) of the farm and the fermenting system. The heat energy consumption is 30%, while the electric energy consumption is 5% from the produced energy. The surplus electricity produced can be used in the electric network while the use of surplus heat is a continuous problem and can be a practical and obstructing factor, but it can be used for heating in winter or for drying, e.g. alfalfa and/or grain in summer time. Based on the above, it can be determined that the farm – in an optimal case – can be energy independent.

The nutrient demand of sweet sorghum, corn and other fodder can be provided by fermentation residue (bio fertilizer). As the application of fermentation residue can help to take the natural cycle, on the one hand it has valuable nutrients which improve the soil structure, and on the other hand we do not produce waste.

The experimental variants that produce biogas with smaller methane content serve the principles of sustainable development via aspects of organic waste management and disposal. We all know about the environmental impacts that threaten our environment, and we also know that all responsible persons must do something to prevent this damage and harm [20].

Conclusions

A complex biogas production and utilization system was created by developing experimental biogas variants, and in such a way both the energy and the environmental goals can be achieved together, as the applied variants can provide favourable conditions for the production and the utilization of biogas. The methane content of biogas satisfies the conditions of utilization so that the heat engines can operate properly. Simultaneously, waste disposal can also be realized. In the interest of a near optimal solution, it is necessary to analyze the production and utilization functions together, considering that the principle of the complex optimization focuses just on the environmental-friendly energy utilization. Thus, if the quantity and/or quality of the input material necessary for developing variants cannot be provided, the energy output can decrease and waste disposal can be overshadowed too.

References

[1] Nemcsics, Á.: Technical Ecology (A műszaki ökológia), *Természetbúvár*, Hungary, Vol. 1 (2003) p. 37

[2] Braun, R: Biogas-Methangärung organischer Abfallstoffe, Springer Wien (1982)

[3] Llabrés-Luengo, P., Mata-Alvarez, J.: Influence of Temperature, Buffer, Composition and Straw Particle Length on the Anaerobic Digestion of Wheat Straw-Pig Manure Mixtures, *Resources, Conservation and Recycling*, Volume 1, Issue 1 (1988) pp. 27-37

[4] Müller, J. et al.: Thermische, chemische und biochemische Desintegrationsverfahren, *Korresp Abwasser*, 50 (2003) pp. 796-804

[5] Gunaseelan, V. N.: Anaerobic Digestion of Biomass for Methane Production: A Review, *Biomass & Bioenergy*, 13 (1997) 1-2, pp. 83-114

[6] Lehtomäki, A., Huttunen, S., Rintala, J. A.: Laboratory Investigations on Co-Digestion of Energy Crops and Crop Residues with Cow Manure for Methane Production, *Resources, Conservation and Recycling* (2006) Nov, pp. 1-19

[7] Mata-Alvarez, J., Mace, S., Llabres, P.: Anaerobic Digestion of Organic Solid Wastes, *Biores Technol*, 74 (2000), pp. 3-16

[8] Panichnumsin, P., Nopharatana, A., Ahring, B., Chaiprasert, P.: Production of Methane by Co-Digestion of Cassava Pulp with Various Concentration of Pig manure, *Biomass and Bioenergy*, Vol. 34 (2010) Issue 8, pp. 1117-1124

[9] Houdková, L., Borán, J., Pěček, J., Šumpela, P.: Biogas – A Renewable Source of Energy, *Thermal Science*, Vol. 12 (2008) No. 4, pp. 27-33

[10] Kalyuzhnyi, S. et al.: Integrated Mechanical, Biological and Physico-Chemical Treatment of Liquid Manure, *Water Science and Technology*, 41 (2000) 12, pp. 175-182

[11] Kapros T.: Biogas Combustion in Industrial Equipment (Biogáztüzelés az ipari berendezésekben), *Biogáz-előállítás és -felhasználás*, Hungary, Vol. 1 (2009) 1, pp. 38-41

[12] Kerek I., Riba D.: Development of Biogas Combustion Equipment (Biogáz tüzelőberendezések fejlesztése), XXXV. Ipari Szeminárium, Miskolc, Hungary, 1999

[13] Selmeci J.: Experiments on Inert Gas Combustion (Inert tartalmú gázok eltüzelésével kapcsolatos kísérleti tevékenység), XXXIV. Ipari Szeminárium, Miskolc, Hungary, 1998

[14] Sándor I.: The Agriculture as an Energy Source for Engine Fuel (A mezőgazdaság, mint motorikus gázenergiaforrás), *Járművek, Mezőgazdasági Gépek*, Hungary, Vol. 12, No. 3 (1965) pp. 107-109

[15] Neyeloff, S., Gunkel, W.: Performance of a CFR Engine Burning Simulated Anaerobic Digester's Gas, ASAE Publication (1981) 2, pp. 324-329

[16] Porpatham, E., Ramesh, A., Nagalingam, B.: Investigations on the Use of Biogas and LPG in a Spark Ignition Engine, PRITHVI International conference on environment friendly transportation, Trivandrum, India, 24-25 February, 2005

[17] Huang, J., Crookes, R. J.: Assessment of Simulated Biogas as a Fuel for Spark Ignition Engine, *Fuel*, Volume 77 (1998) 15, pp. 1793-1801

[18] Crookes, R. J.: Comparative Bio-Fuel Performance in Internal Combustion Engines, *Biomass and Bioenergy*, Volume 30 (2006) 5, pp. 461-468

[19] Meggyes A., Nagy V.: Effect of the Biogases Produced by Different Kinds of Recipes on the Operation of Gas Engines, *Proceedings*, 9th International Conference on Heat Engines and Environmental Protection, Balatonfüred, Hungary, 2009, pp. 71-76

[20] Nagy, V.: Effect of the Biogas Producing Methods on the Operation of the Gas Engine Considering Emission (A biogáz előállítási eljárások hatása a gázmotorok üzemére, különös tekintettel a károsanyag kibocsátásra), Ph.D thesis, Budapest University of Technology and Economics, Budapest, Hungary, 2010

[21] Kovács V. B., Meggyes A.: Energetic Utilization of Pyrolysis Gases in IC Engine, *Acta Polytechnica Hungarica,* Vol. 6, No. 4 (2009) pp. 157-172

[22] Borbély É.: The Kinetics of Cellulose Grafting with Vinyl Acetate Monomer, *Acta Polytechnica Hungarica,* Vol. 2, No. 2 (2005) pp. 67-76

[23] Misra S.: An Approach for the Empirical Validation of Software Complexity Measures, *Acta Polytechnica Hungarica,* Vol. 8, No. 2 (2011) pp. 141-160

Durability of Cellulose and Synthetic Papers Exposed to Various Methods of Accelerated Ageing

Mirica Karlovits[1], Diana Gregor-Svetec[2]

[1] University of Ljubljana, PostDoct researcher, Slovenia, mirica.dk@gmail.com

[2] University of Ljubljana, Faculty of Natural Sciences and Engineering, Snežniška 5, 1000 Ljubljana, Slovenia, diana.gregor@ntf.uni-lj.si

Abstract: The paper presents a study of the stability of cellulose and synthetic papers exposed to various methods of accelerated ageing. Particular consideration was given to the optical and mechanical stability of six paper samples (one film synthetic paper, two fibre synthetic papers, one lignin-free paper of higher quality and two security cellulose papers), which have undergone changes during accelerated ageing. The papers were artificially aged using standard techniques of accelerated ageing, e.g. moist-heat (80 °C and 65% RH), dry-heat (105 °C) and treatment with a xenon arc lamp (35 °C CT, 50 °C BST, 35% RH). The ageing was performed for the periods of 1, 2, 3, 6 and 12 days. The changes in the optical (ISO brightness, Yellowness Index), surface (roughness, paper topography) and mechanical stability (zero-span tensile strength, elongation at break, folding endurance) of papers were measured during the periods of accelerated ageing. The results show that the differences between synthetic and cellulose papers exist. On average, the dry-heat ageing had the highest influence on ISO brightness and Yellowness Index on synthetic papers, while the treatment with a xenon lamp had the strongest influence on cellulose papers. A comparison of mechanical properties showed that synthetic papers are more durable than cellulose paper; they had higher zero-span tensile strength and folding endurance, and showed substantially better ageing resistance to dry-heat and moist-heat accelerated ageing than cellulose papers. It was also noticed that the surface roughness increased after all three accelerated ageing processes.

Keywords: synthetic paper; cellulose paper; accelerated ageing; optical properties; mechanical properties; paper topography

1 Introduction

As with all other organic materials, paper is subjected to a number of fundamental deterioration processes. Under normal conditions of storage, these processes are very slow; however, they eventually and inevitably still lead to the well-known

ageing effects, e.g. yellowing and loss of strength [1]. Heat and moisture are two of the most important environmental influences on the stability of papers [2]. At artificial or accelerated ageing methods, a material is exposed in a climate-chamber to extreme conditions in the terms of temperature and humidity for a certain period of time, during which the changes in the material are measured. Artificial ageing tests are often used to determine the permanence of paper, i.e. its rate of degradation, as well as to predict the long-term effect of a conservation treatment [1]. Moreover, exposure to light can cause fading and can shorten the use-life of paper. The degree of fading varies with the type of illumination and is greater with higher intensity light [2]. During accelerated ageing, the measured variables can include exposure time, exposure to UV irradiation over specific wavelength range and exposure to moisture as a number of cycles or time [3].

J. Malešič et al. [4] studied the photo-induced degradation of cellulose. The research demonstrated that extensive oxidative degradation of cellulose, accompanied by the formation of hydroxyl radicals, occurs during the exposure to light with $\lambda > 340$ nm. The studies showed that the rate of degradation, carbonyl formation and brightness decrease may be reduced with the addition of magnesium carbonate. Dr. S. Kaufmann and A. Bossmann [5] studied the light resistance of synthetic fibres under extreme exposure conditions. M. S. Islam et al. [3] studied the influence of accelerated ageing on the physico-mechanical properties of alkali-treated industrial hemp fibre reinforced poly(lactic acid) (PLA) composites. After the accelerated ageing, tensile strength, flexural strength, Young's modulus, flexural modulus and fracture toughness were found to decrease, whereas impact strength increased for aligned untreated long hemp fibre/PLA (AUL) and aligned alkali treated long hemp fibre/PLA (ALL) composites. B. Havlíonova et al. [6] studied the mechanical properties of papers (one alkaline and three different acidic samples) exposed to various methods of accelerated ageing. The increased temperature and relative humidity caused a significant loss of folding endurance, especially for acidic papers. S. Soares et al. [7] studied the degradation of cellulose in the form of transformer insulating paper and Whatman filter paper in the air at the temperatures from 200 °C to 550 °C with and without the addition of 0.01 wt.% NaCl, $ZnCl_2$ and $CuCl_2$, using the solid-state NMR and FTIR spectroscopy. Major changes occurred at the temperatures higher than 250 °C, resulting in the loss of protons and the development of new saturated and unsaturated structures.

Our research study focused on the investigations of the stability of papers made from synthetic and/or cellulose fibres. The surface, optical and mechanical properties of three different synthetic papers and three different cellulose papers exposed to various methods of accelerated ageing were determined.

2 Experimental

2.1 Materials

In the present study, commercially available papers were used, three types of synthetic papers and three types of cellulose papers.

Paper 1: Yupo (manufacturer: Yupo Corporation, Japan) is a biaxially-oriented film synthetic paper. It consists of three extruded polypropylene (PP) layers with inorganic filler ($CaCO_3$) and does not contain wood pulp or other bio materials.

Paper 2: Pretex (manufacturer: Neenah-Lahnstein Company, Germany) is a double-side coated fibre synthetic paper made from a mixture of selected pulp and synthetic fibres (polyamide – PA and polyester – PES) in a combination with a special binder system.

Paper 3: Neobond (manufacturer: Neenah-Lahnstein Company, Germany) is a double-side coated fibre synthetic paper made from a mixture of selected pulp and synthetic fibres (polyamide – PA, polyester – PES and viscose – CV), reinforced with a special impregnation.

Paper 4: G-print (manufacturer: Stora Enso, Finland) is a coated lignin-free paper made from 100% virgin cellulose fibres.

Paper 5: Catanelle (manufacturer: Fabriano, Italy) is an uncoated security paper made from 100% E.C.F. chemical bleached pulp. The paper has multitonal watermark and contains fluorescent security fibres.

Paper 6: Small Money (manufacturer: Gmund, Germany) is a lignin-free security paper made from a mixture of old german marks, waste paper and cellulose fibres.

2.2 Methods

The experiments, which were focused on the surface, optical and mechanical paper properties, were performed in compliance with ISO standards. The paper samples were aged using standard techniques for accelerated ageing: moist-heat based on the standard SIST ISO 5630-3 (80 °C and 65% relative humidity), dry-heat based on the standard SIST ISO 5630-1 (105 °C) and ageing with a xenon lamp in a Xenotest® Alpha apparatus based on the standard ISO 12040 (35 °C Chamber Temperature, 50 °C Black Standard Temperature, 35% relative humidity). For all three types of accelerated ageing, the exposure time was 1, 2, 3, 6 and 12 days [8-10].

The optical properties of papers were evaluated based on ISO Brightness and Yellowness Index YI E313 [13-14]. The measurements were performed in

accordance with the standard ISO 2470 (ISO Brightness, R_{457}) with an X-Rite spectrophotometer at D65/10° and in accordance with the ASTM Method 313 (YI E313) with a spectrophotometer Spectroflash 600 – Datacolor International at D65/10°.

The influence of dry-heat and moist-heat ageing on the mechanical properties was determined using an Instron 5567 tensile testing machine. A paper strip with 15 mm in width was clamped with the span length as close to zero as possible. The zero-span tensile strength was determined according to ISO 15361. The tensile properties were measured in the machine (MD) and cross directions (CD) of papers. The folding resistance was measured on the MIT folding endurance tester in the machine and cross direction of papers (load: 2 kg, except for the aged Paper 6 in MD: 0.5 kg).

The chemical composition of paper surface before and after 12 days of accelerated ageing was determined with the ATR-FTIR technique on a FTIR spectrometer PerkinElmer SpectrumGX. The standard FTIR spectrometer settings were as follows: range 4000-500 cm^{-1}, 64 scans, resolution 4.00 cm^{-1}.

The unaged and aged papers were imaged using a Scanning Electron Microscope JOEL, JSM 6060LV at 100× magnification and 10 kV voltage. The captured JPEG images were re-saved into 8 bit images and analysed using ImageJ software (Mean Grey Values, Standard Deviation and Median Value). The topography of samples was evaluated at the area of 1120 × 750 pixels with Surface Plot diagrams. The average surface roughness (R_a) of unaged and aged papers was measured with a Surface Roughness Tester TR200.

3 Results

3.1 Influence of Accelerated Ageing on Optical Properties of Papers

The deterioration of paper upon ageing is initiated with an irreversible change of their mechanical, chemical and optical properties [11]. In the first part of the investigation, the influences of various accelerated ageing techniques on the optical properties (ISO Brightness, Yellowness Index) of papers were investigated. Figures 1-3 summarise the influence of three accelerated ageing techniques (dry-heat treatment, moist-heat treatment and treatment with a xenon lamp) on the ISO Brightness of the papers.

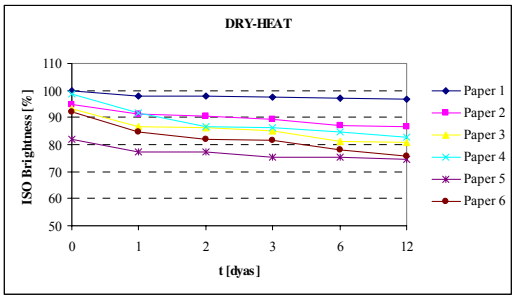


Figure 1
Influence of dry-heat accelerated ageing on ISO Brightness of papers

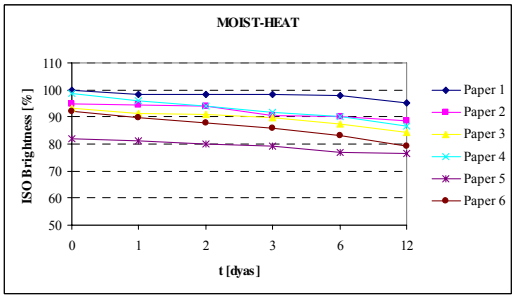


Figure 2
Influence of moist-heat accelerated ageing on ISO Brightness of papers

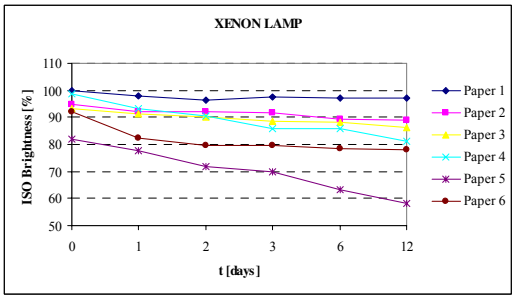


Figure 3
Influence of accelerated ageing with xenon lamp on ISO Brightness of papers

An important feature of cellulose fibres in the paper is their degradation due to ageing [8]. One of the major sources of decay of materials made from natural fibrous materials is the effect of light [4]. Paper made from cellulose fibres has the tendency to undergo yellowing (brightness reversion) upon exposure to the sunlight. There is a general agreement that the coloration occurs due to the remaining lignin constituents in the pulp, although neither the precise nature of the

chromophores responsible for this nor the exact mechanism for their formation has been conclusively established [12]. The ISO brightness values for papers before the ageing were as follows: Paper 1 = 99.9%, Paper 2 = 94.9%, Paper 3 = 93.1%, Paper 4 = 98.6%, Paper 5 =82.0% and Paper 6 = 92.2%. Paper with the highest ISO brightness, i.e. the film synthetic paper (Paper 1), retained the highest brightness at all three accelerated aging techniques after each day of aging. For most papers, the dry-heat treatment (cf. Figure 1) caused a higher loss in ISO brightness than the moist-heat treatment (cf. Figure 2) and the treatment with a xenon lamp (cf. Figure 3). With the dry-heat treatment (105 °C) and at the moist-heat treatment (80 °C, 65% RH), the most obvious decrease in ISO brightness was established for the cellulose paper containing old value paper and waste paper (Paper 6). Within 12 days, brightness dropped by 16.3 units for the dry-heat treatment and by 13.1 units for the moist-heat treatment, which corresponds to a perceptible change. The treatment with a xenon lamp (35 °C CT, 50 °C BST, 35% RH) influenced most substantially the ISO brightness of both cellulose papers (Paper 4, Paper 5). The highest loss in ISO brightness was obtained for the security paper (Paper 5), where the value dropped by 23.9 units. It was established that all three synthetic papers are more durable than the cellulose papers of higher quality. According to the producer, both fibre synthetic papers contain pigment colors, no optical brighteners and still have good light fastness after many years. Normally, temperatures up to 100 °C do not influence the paper properties, and a short-term increase to 200°C only leads to paper surface discoloration [13]. The papers produced from cellulosic fibres with various additives are determined by the extent of oxidative and hydrolytic reactions taking place upon ageing [6]. Two general mechanisms are involved in the degradation of cellulosic materials by light in the visible and ultraviolet ranges. In the short-wave UV region, the breakdown is believed to occur due to the photolysis of cellulosic chains, leading to the cleavage of carbon-to-carbon or carbon-to-oxygen linkages, without any particular evidence that the reaction with oxygen is vital to the process. The other type of reaction is thought to involve the presence of a substance which can act as a photosensitiser and which, in the presence of oxygen and moisture, leads to the production of hydrogen peroxide, which in turn degrades the cellulose by oxidation [14].

In terms of visual appearance, absorption in the blue part of the light spectrum causes yellowness. Visually, yellowness is associated with scorching, soiling and general product degradation by light, chemical exposure and processing [15]. The Yellowness Index is a number calculated from spectrophotometric data that describes the change in color of a test sample from clear or white toward yellow. This test is most commonly used to evaluate the color changes in a material caused by a real or simulated outdoor exposure. Lightfastness normally decreases with increasing atmospheric humidity, the extent of the effect depending on the dye-substrate system, which is very pronounced for cellulosic fibres [16].

The Yellowness Index according to the ASTM Method E313 is calculated as follows:

$$YIE313 = \frac{100\left(C_X X - C_Z Z\right)}{Y} \tag{1}$$

where X, Y, Z are the CIE tristimulus values, C_X and C_Z are coefficients (D65/10°: $C_x = 1.3013$, $C_Z = 1.1498$) [17].

Figures 4-6 summarise the influence of three accelerated ageing techniques (dry-heat treatment, moist-heat treatment and treatment with a xenon lamp) on the Yellowness Index of the papers.

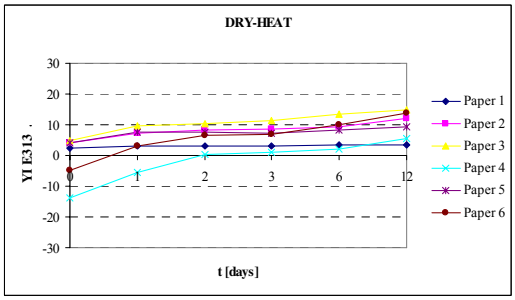


Figure 4
Influence of dry-heat accelerated ageing on Yellowness Index of papers

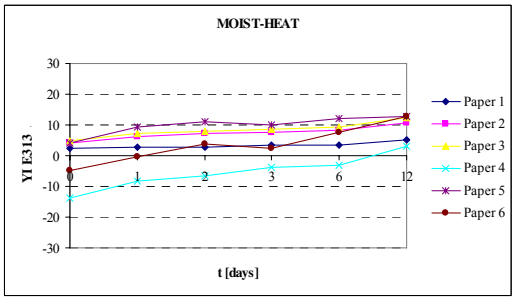


Figure 5
Influence of moist-heat accelerated ageing on Yellowness Index of papers

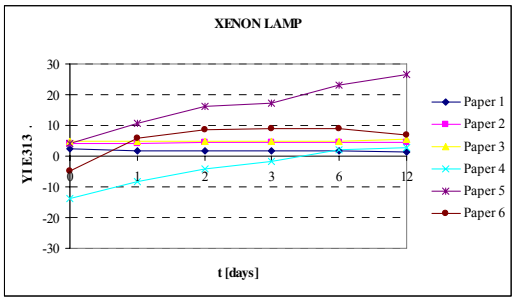


Figure 6
Influence of accelerated ageing with xenon lamp on Yellowness Index of papers

The tested papers slightly differ in yellowness; Paper 1: YI E313 = 2.4, Paper 2: YI E313 = 4.3, Paper 3: YI E313 = 4.9, Paper 4: YI E313 = –13.8, Paper 5: YI E313 = 4.1 and Paper 6: YI E313 = –4.7, the values leading to the conclusion that Papers 1-3 and 5 are more yellowish, while Papers 4 and 6 are more bluish. Paper yellowing is a natural process of paper ageing, which is caused by the sunlight, moisture and air. On average, the best stability among synthetic papers was obtained by the film synthetic paper (Paper 1) made from PP fibres, while among the cellulosic papers, by Paper 5. The results obtained for Paper 5 showed that Paper 5 yellowed more slowly under the dry-heat treatment and the most under treatment with the xenon lamp. After 12 days of treatment with the xenon lamp, the value for Paper 5 was YI E 313 = 26.51. The reason is in the paper structure; Paper 5 is not coated, which influenced the results. The Yellowness Index for Paper 6 for all three ageing methods increased polynomially. Among all the tested papers, only Paper 6 contained recycled fibres. Papers 2 and 3 contained polyamide (PA) and polyester (PES) fibres apart from cellulosic fibres. Polyester (PES) fibres have greater light resistance than polyamide (PA) and cellulosic fibres, and photooxidation takes place at higher temperatures. It was noticed that the yellowing during the dry-heat and moist-heat treatment was the most progressive for Paper 4 and Paper 6, and the treatment with the xenon lamp for Paper 4 and Paper 5. Under all treatments, Paper 4 and Paper 6 turned from bluish to yellowish after accelerated ageing. The loss of brightness and paper yellowing during the ageing procedures is attributed to the presence of chromophores formed by the degradation of paper components (cellulose, hemicellulose, lignin) [11]. The degradation starts under the presence of light or no light and in the presence of increased temperature and humidity [18]. Pure cellulose absorbs visible light only to a small extent, while the absorption in the near UV spectral region is more pronounced. The absorption in the blue spectra causes the yellowing [4]. Lightfastness normally decreases with increasing atmospheric humidity, the extent of the effect depending on the dye-substrate system, which is very pronounced with cellulosic fibres [16]. Photodegradation is influenced by the surface-area-to-volume ratio to such a great extent, since the mechanism of degradation is largely

photo-oxidative, and the surface area of the fibre in contact with air would hence be a very important factor in the chemical reactivity and its inherent kinetics. The greater the surface area facing a light source, the greater the actinic energy absorbed and the more vigorous the photodegradation reaction [14].

3.2 Influence of Accelerated Ageing on Mechanical Properties of Papers

The mechanical properties of paper can vary significantly by selecting different types of fibres and fibre preparation. Tensile strength is indicative of fibre strength, fibre bonding and fibre length. Fibre length and coarseness also influence the tensile strength of paper [11]. The tensile strength of paper is the maximum force per unit width that a paper strip can resist before breaking when applying the load in the direction parallel to the length of a strip. A special case of tensile strength is the zero-span tensile strength, which is a good indicator of the average strength of individual fibres. Table 1 shows the influence of dry-heat accelerated ageing, while Table 2 shows the influence of moist-heat accelerated ageing on the zero-span tensile strength, elongation at break and stress at break of papers.

Table 1
Influence of dry-heat ageing on zero-span tensile strength, elongation at break and stress at break of papers

Samples	Fibre orientation	Days of ageing	Zero-span tensile strength [N]	Elongation at break [mm]	Stress at break [N/mm^2]
Paper 1	*MC*	0	250.3	6.7	131.4
	CD		87.4	15.8	45.9
	MC	*6*	297.7	8.1	156.8
	CD		102.9	20.9	54.2
	MC	*12*	294.6	8.0	155.1
	CD		104.2	19.6	54.9
Paper 2	*MC*	*0*	116.3	1.0	91.2
	CD		78.8	2.0	61.8
	MC	*6*	196.0	1.3	100.5
	CD		140.6	1.7	72.1
	MC	*12*	186.7	1.2	95.8
	CD		126.1	1.5	64.6
Paper 3	*MC*	*0*	65.1	1.7	40.4
	CD		45.3	2.0	28.1
	MC	*6*	79.4	1.6	59.2
	CD		50.3	1.8	31.2
	MC	*12*	79.5	1.6	49.4
	CD		51.2	2.0	31.8

	MC	*0*	95.7	0.7	70.9
	CD		59.1	1.0	43.7
Paper 4	*MC*	*6*	120.1	0.7	89.4
	CD		73.4	0.9	54.7
	MC	*12*	113.8	0.7	84.8
	CD		69.1	0.8	51.5
	MC	*0*	118.0	0.8	73.5
	CD		88.3	1.6	55.0
Paper 5	*MC*	*6*	157.0	0.9	98.5
	CD		109.7	1.2	68.8
	MC	*12*	148.6	0.9	93.2
	CD		102.9	1.1	64.6
	MC	*0*	101.0	0.7	52.6
	CD		71.3	1.1	37.2
Paper 6	*MC*	*6*	118.6	0.7	61.5
	CD		87.4	0.9	45.3
	MC	*12*	118.8	0.7	58.0
	CD		79.2	0.8	41.1

Table 2

Influence of moist-heat ageing on zero-span tensile strength, elongation at break and stress at break of papers

Samples	**Fibre orientation**	**Days of ageing**	**Zero-span tensile strength [N]**	**Elongation at break [mm]**	**Stress at break [N/mm^2]**
	MC	0	250.3	6.7	131.4
	CD		87.4	15.8	45.9
Paper 1	*MC*	*6*	296.5	8.0	156.1
	CD		101.2	19.4	53.3
	MC	*12*	292.2	7.7	153.9
	CD		104.1	20.7	54.8
	MC	*0*	116.3	1.0	91.2
	CD		78.8	2.0	61.8
Paper 2	*MC*	*6*	203.9	1.3	104.6
	CD		142.5	1.7	73.1
	MC	*12*	186.2	1.3	95.5
	CD		133.7	1.7	68.6
	MC	*0*	65.1	1.7	40.4
	CD		45.3	2.0	28.1
Paper 3	*MC*	*6*	79.8	1.6	49.6
	CD		50.5	1.9	31.4
	MC	*12*	80.4	1,6	49.9
	CD		50.4	1.9	31.3

	MC	*0*	95.7	0.7	70.9
	CD		59.1	1.0	43.7
Paper 4	*MC*	*6*	118.9	0.7	88.6
	CD		72.5	1.0	54.0
	MC	*12*	117.5	0.7	87.5
	CD		69.3	0.8	50.1
	MC	*0*	118.0	0.8	73.5
	CD		88.3	1.6	55.0
Paper 5	*MC*	*6*	159.9	0.9	100.3
	CD		114.0	1.1	71.5
	MC	*12*	153.6	0.8	96.3
	CD		109.5	1.1	68.7
	MC	*0*	101.0	0.7	52.6
	CD		71.3	1.1	37.2
Paper 6	*MC*	*6*	123.2	0.7	63.8
	CD		86.8	0.9	45.0
	MC	*12*	118.9	0.7	61.6
	CD		82.3	0.9	42.7

The mechanical properties of investigated papers are substantially influenced by individual characteristics of cellulose and synthetic fibres, concentration and chemical properties of fillers and additives, as well as by the paper network structure. The differences in tensile properties between the papers are high, especially in the machine direction (MD) of papers. The paper properties, such as tensile strength, vary significantly between the machine and cross directions of papers. This is attributed to fibre orientation; however, the fact that in the machine direction paper is dried under much higher resistance than in cross direction is often of even greater importance [19]. The highest zero-span tensile strength and stress at break belonged to the film synthetic paper made from the PP fibres in the form of extruded layers. The film synthetic paper was also much more extensible than other investigated papers. The lowest zero-span tensile strength and stress at break belonged to the film synthetic paper (Paper 3), whereas all other papers had similar tensile strength. The differences in elongation at break are a consequence of paper composition; Paper 2 and Paper 3 contained cellulose and synthetic fibres, whereas Papers 4-6 contained only cellulose fibres, which are less flexible and less extensible. Polyester, polyamide and polypropylene fibres have high strength and excellent strength retention properties, and are mostly used as non-absorbable suture [20]. It is well known that the effect of exposure to moisture and heat or a combination of these parameters may damage paper stiffness and strength [3]. During ageing, the loss of paper strength is a consequence of the degradation processes of its main structural component, the fibre [21]. From Tables 1 and 2, it is evident that the dry-heat and moist-heat ageing influenced the tensile strength and extension of papers. Synthetic papers are more thermally stable and more durable to light and moisture, and they have in normal use higher

dimensional stability than cellulose papers. They also have superior resistance against tear and damage [22]. Moisture generally reduces the tensile strength of the hydrophilic fibre. The exception is the most natural cellulose fibre, where in the wet the tensile strength increases due to the layer structure of the secondary cell wall. Humidity does not affect the tensile strength of hydrophobic fibres (polyester), since in the wet the tensile strength does not change [14]. The values of the zero-span tensile strength reflect the detailed structure of a paper and mainly the properties of its individual fibres, i.e. dimension and strength of fibres, their arrangements and interfibre bonding. The effects of accelerated ageing processes on paper are interpreted in terms of bond scission between fibres, chain scission producing weaker fibres and degradation of smaller molecules. At first, only the degradation in amorphous regions takes place, leaving more ordered structure, resulting in crosslinking by additional bonds leading to increased strength and brittleness.

Table 3 shows the loss in the folding endurance of papers after 6 and 12 days of accelerated ageing.

Table 3
Number of double folds of papers after moist-heat and dry-heat ageing

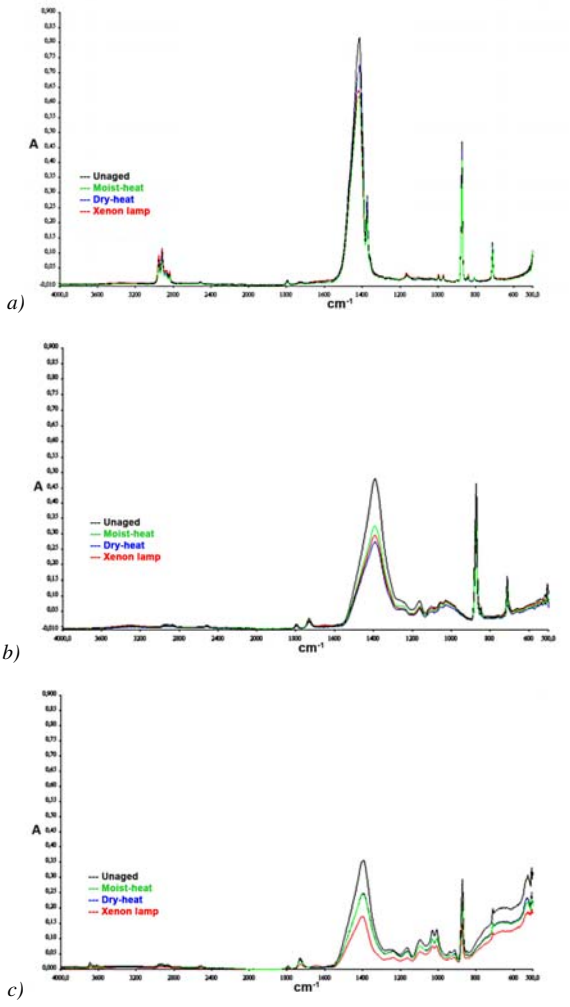
Numbers of double folds after breaking (load: 2 kg)						
Samples	**Fibre orientation**	**Unaged papers**	**Moist-heat ageing**		**Dry-heat ageing**	
			6 days	**12 days**	**6 days**	**12 days**
Paper 1	MC	Did not break	Did not break	Did not break	Did not break	Did not break
	CD					
Paper 2	MC	3543	3479	3139	2335	1807
	CD	743	196	166	119	99
Paper 3	MC	Did not break	Did not break	Did not break	Did not break	Did not break
	CD					
Paper 4	MD	641	643	550	295	246
	CD	280	218	188	193	135
Paper 5	MD	2140	939	755	701	614
	CD	625	75	52	53	22
Paper 6	MD	781	354	217	120	33
	CD	114	69	60	36	14

Folding endurance represents the most sensitive indicator of paper breakage upon ageing. The effects of accelerated ageing processes on paper are interpreted in terms of fibre chain scission producing weaker fibres and covalent crosslinking by additional bonds leading to increased brittlenes [6]. The results presented in Table 3 show the highest folding resistance to ageing for film synthetic paper (Paper 1) and fibre synthetic paper (Paper 3). Neither paper broke under the load of 2 kg at 5,000 double bonds, meaning that these two papers had higher strength, were more flexible and more bonded. It was found that dry-heat ageing processes had a

higher impact on the folding endurance of papers than moist-heat ageing. The most obvious decrease in the number of double folds was obtained for security cellulose papers (Paper 6), as it contained recycled fibres. For all papers, the load was 2 kg, except for Paper 6; it was in MD direction 0.5 kg.

3.3 FTIR Spectroscopy

Infrared (IR) spectroscopy is useful in the elucidation and identification of the molecular structure and in the applications of quantitative analyses. The FTIR spectra of unaged papers, as well as papers after 12 days of various accelerated ageing, procedures are illustrated in Figure 7.



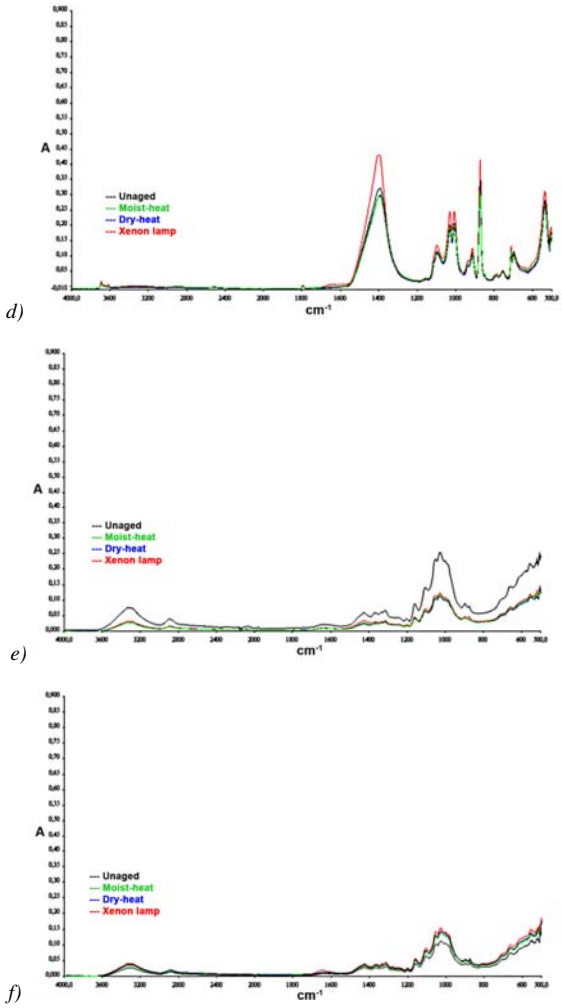


Figure 7

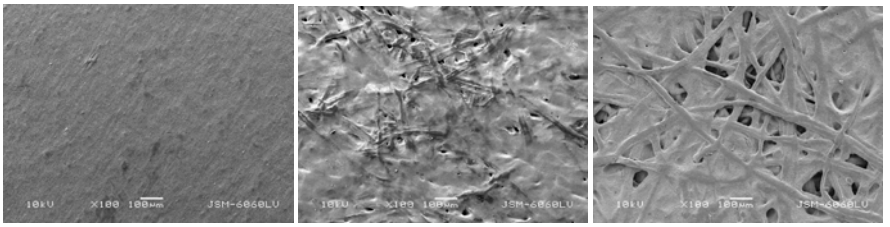
FTIR spectra of unaged and for 12 days acceleratedly aged papers: *a)* Paper 1, *b)* Paper 2, *c)* Paper 3, *d)* Paper 4, *e)* Paper 5 and *f)* Paper 6

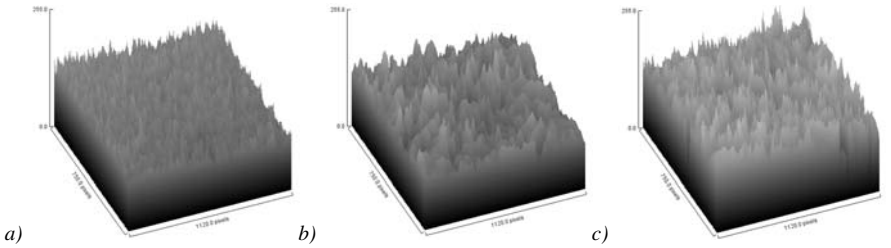
The applications of all accelerated ageing procedures led to the most obvious decrease in absorption for Paper 1 (cf. Figure 7/a) in the regions 1500-1400 cm^{-1}, especially under the treatment with the xenon lamp. The band at the 1400 cm^{-1} for unaged Paper 1 obtained 0.500 of absorption, while for the treatment with the xenon lamp obtained 0.385 of absorption. Only for Paper 1, the absorption peak was noticed at 2922 cm^{-1}, which represents the CH_2-gruoup vibration in the main PP polymer chain [23]. In comparison with all other papers, Paper 1 obtained the highest intensity of absorption. For Paper 2 (cf. Figure 7/b), the most obvious

decrease in absorption after accelerated ageing procedures was observed in the region 1550-1000 cm^{-1}. For example, at the band at 1400 cm^{-1}, the following values were obtained: A = 0.472 (unaged), A = 0.320 (moist-heat), A = 0.269 (dry-heat) and A = 0.290 (xenon lamp). For Paper 3 (cf. Figure 7/c), a decrease in the absorption in the region 1500-500 cm^{-1} was observed. The curves for moist-heat and dry-heat treated papers behaved similarly, while the curve for the xenon lamp treated papers decreased the most. From Figure 7/d, it can be seen that for Paper 4, the absorption under treatment with the xenon lamp increased in the region 1600-500 cm^{-1}, while the curves for the moist-heat and dry-heat treated papers decreased. The intensity of bonds in the cellulose finger print is in the region 1500-800 cm^{-1}.While the hydrolytic degradation results in the breaking of (1→4) β-glycosidic bonds and the occurrence of the formation of aldehyde groups, the oxidative degradation of cellulose results in the opening of the β-D-glucopyranose ring, causing the formation of carboxylic and aldehyde groups. These bands are located in the region 1750-1617 cm^{-1} [24]. For Paper 5 (cf. Figure 7/e), all three types of accelerated ageing behaved similarly. In all regions, the absorption was lower than for unaged paper. In contrast to Paper 5, the unaged Paper 6 obtained lower absorption than the aged samples. From all FTIR spectra, it can be seen that for Papers 1-3 and 5, the absorption of aged samples decreased compared to the unaged sample. The treatment with the xenon lamp reflected the growing absorption for Paper 4 and all three ageing procedures for Paper 6.

3.3 Paper Surface Topography

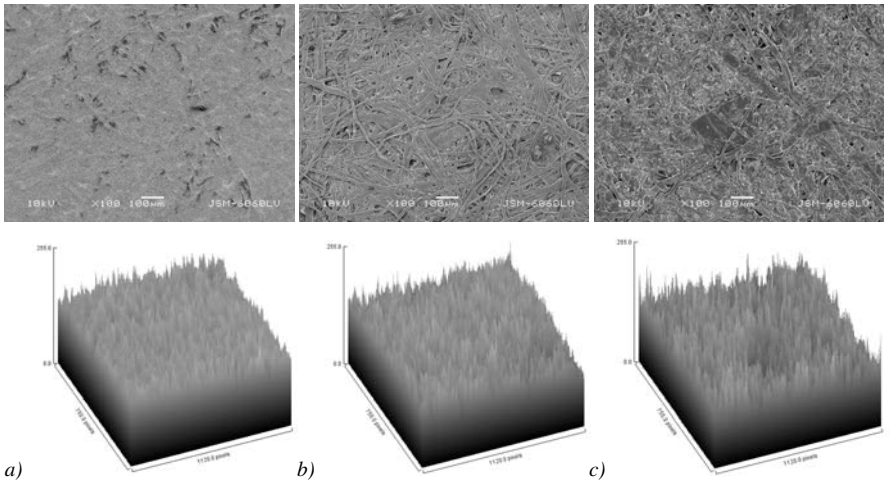
The mechanical properties of different paper samples are substantially influenced by the characteristics of individual fibres, nature, concentration and chemical properties of fillers and additives, and by the paper network structure [6]. Figures 8 and 9 present the SEM images of the unaged papers and their surface plot diagrams. Surface plot displays a three-dimensional graph of intensities of pixels in a greyscale or pseudo color image.





a) *b)* *c)*

Figure 8
SEM images of unaged synthetic papers and their surface plot diagrams: a) Paper 1, b) Paper 2, c) Paper 3



a) *b)* *c)*

Figure 9
SEM images of unaged synthetic papers and their surface plot diagrams: a) Paper 4, b) Paper 5, c) Paper 6

Table 4
Results of analysed SEM images of papers and average surface roughness

Samples	Ageing	Mean	StdDev	Median	Ra [µm]
Paper 1	*Unaged*	118.42	22.95	117	0.44
	Moist-heat	129.29	17.60	128	0.75
	Dry-heat	111.17	22.90	109	0.70
	Xenon lamp	118.49	10.20	117	0.43
Paper 2	*Unaged*	143.13	37.20	142	1.84
	Moist-heat	167.11	31.20	171	1.93
	Dry-heat	140.47	40.50	143	1.81
	Xenon lamp	142.96	27.10	143	2.35

Paper 3	*Unaged*	136.52	35.84	137	4.77
	Moist-heat	150.61	32.50	149	5.64
	Dry-heat	133.27	44.60	133	6.05
	Xenon lamp	117.42	26.40	118	4.95
Paper 4	*Unaged*	140.80	35.55	140	0.85
	Moist-heat	123.62	9.80	125	1.17
	Dry-heat	130.22	35.70	128	0.91
	Xenon lamp	138.63	15.40	138	1.64
Paper 5	*Unaged*	130.43	42.67	129	2.94
	Moist-heat	105.70	30.30	101	2.75
	Dry-heat	109.08	29.40	106	2.90
	Xenon lamp	97.96	22.90	97	4.12
Paper 6	*Unaged*	115.64	46.40	108	3.07
	Moist-heat	147.30	37.60	141	3.14
	Dry-heat	131.12	42.90	124	2.97
	Xenon lamp	139.67	22.30	138	3.29

Legend:
Mean – Average grey value within selection. This is the sum of grey values of all pixels in the selection divided by the number of pixels.
StdDev – Standard deviation of grey values used to generate the mean grey value.
Median – Median value of pixels in the image or selection.
Ra – Average surface roughness.

In the scanning electron images of papers (cf. Figure 12), some differences can also be seen. Plot diagrams support the visual evaluation of roughness on the SEM images. 3D surface plot diagrams show the highest level of uniformity for Paper 1.

From Table 4, it is seen that all three synthetic papers (Papers 1-3) and one cellulose paper (Paper 6) obtained the highest mean grey values and also the median value under the moist-heat treatment. With ageing, the uniformity of surface topography improves, except for the dry-heat treated synthetic papers. For all papers, the highest uniformity was obtained after the treatment with the xenon lamp. It was noticed that the average surface roughness of synthetic and cellulose papers increased after all three accelerated ageing processes. All three cellulose papers (Papers 4-6) and one fibre synthetic paper (Paper 2) became rougher after the treatment with the xenon lamp, while the film synthetic paper (Paper 1) and fibre synthetic paper (Paper 3) under the dry-heat ageing. Accelerated ageing had the highest influence on the fibre synthetic paper (Paper 3), which had among all papers the highest roughness. The environmental action causes oxidation of the coating surface layers and in this way deteriorates the cohesion between the filler particles and polymer matrix, which results in increased roughness of the polymer coating surface [25].

Conclusions

In the present work, a comparison of the surface, optical and mechanical properties of synthetic and cellulose papers exposed to various accelerated ageing (dry-heat, moist-heat and xenon lamp) was studied. For most of the papers, the dry-heat treatment caused a higher loss in ISO brightness than the moist-heat treatment and the treatment with the xenon lamp. The yellowing of the paper and the brightness decrease upon ageing were more pronounced for cellulose papers than for synthetic papers. The film synthetic paper with the highest ISO brightness kept high brightness over 95% and showed no yellowing even after the ageing. The differences between papers were noted also for the mechanical properties of the papers. Ageing had little influence on the tensile strength and elongation at break; zero-span tensile strength increased, whereas a dramatic decrease in folding endurance was observed. Dry-heat ageing had a higher impact on the folding endurance of papers than moist-heat ageing. Synthetic papers had higher mechanical resistance to ageing than cellulose papers. It was also noticed that surface roughness increased after the ageing, especially at the synthetic paper with higher roughness.

The best durability in the dry-heat and moist-heat environment was obtained by the film synthetic paper. Both fibre synthetic papers also showed higher optical and mechanical resistance than cellulose papers.

References

[1] Porck HJ. *Rate of Paper Degradation*. European Commission on Preservation and Access, Amsterdam, pp. 11-23, 2003

[2] A Consumer Guide to Traditional and Digital Print Stability [online] Available at: http://www.imagepermanenceinstitute.org/shtml_sub/consumerguide.pdf; [accessed November 2011]

[3] Islam MS, Pickering KL, Foreman NJ. Influence of Accelerated Ageing on the Physico-Mechanical Properties of Alkali-treated Industrial Hemps Fibre Reinforced Poly(lactic acid) (PLA) Composites. *Polymer Degradation and Stability,* Vol. 95, pp. 59-65, 2010

[4] Malešič J, Kolar J, Strlič M, Kočar D, Fromageot D, Lemaire J, Haillant O. Photo-induced Degradation of Cellulose. Polymer *Degradation and Stability,* Vol. 89, pp. 64-69, 2005

[5] Kaufmann S, Bossmann A. Light Resistance of Synthetic Fibers under Extreme Exposure Conditions. *Chemical Fibers International*, Vol. 45, pp. 188-190, 1995

[6] Havlínová B, Svetozár K, Pertovičová M, Maková A, Brezová V. A Study of Mechanical Properties of Papers Exposed to Various Methods of Accelerated Ageing. Part I. The Effect of Heat and Humidity on Original Wood-Pulp Papers. *Journal of Cultural Heritage*, Vol. 10, pp. 222-231, 2009

[7] Soares S, Ricardo MPSN, Jones S, Heatley F. High Temperature Thermal Degradation of Cellulose in Air Studied Using FTIR and ^{1}H and ^{13}C Solid-State NMR. *Europen Polymer Journal*, Vol. 37, pp. 737-745, 2001

[8] SIST ISO 5630-3:1997 Paper and Board – Accelerated Ageing – Part 3: Moist Heat Treatment at 80° C and 65% Relative Humidity

[9] SIST ISO 5630-1:1997 Paper and Board – Accelerated Ageing – Part 1: Dry Heat Treatment at 105° C

[10] ISO 12040 Graphic Technology: Prints and Printing Inks – Assessment of Light Fastness using Filtered Xenon Arc Light

[11] Havlínová B, Babiaková D, Brezová V, Ďurovič M, Novotná M, Belányi F. The Stability of Offset Inks on Paper upon Ageing. *Dyes and Pigments*, Vol, 54, pp. 173-188, 2002

[12] Durbeej B, Eriksson LA. Photodegradation of Substituted Stilbene Compounds: What Colors Aging Paper Yellow? *The journal of physical chemistry*, Vol. 109, pp. 5677-5682, 2005

[13] Neenah Lahnstein [online]. Available at: http://www.neenah-lahnstein.de; [accessed November 2011]

[14] Rijavec T. Delovanje sončne svetlobe na vlakna. *Tekstilec*, Vol. 43, Nu. 3-4, pp. 86-102, 2000

[15] Hunter RS, Harold RW. *The measurement of Appearance*. 2nd ed. John Wiley&Sons : New York, pp. 205-208, 1994

[16] Feller RL. *Accelerated Aging: Photochemical and Thermal Aspects*. Library of Congress Cataloging-in-Publication Data, Michigan, pp. 183-184, 1996

[17] Yellowness Index per ASTM Method 313

[18] Černič-Letnar M, Scheicher L. Trajnost in obstojnost papirja potiskanega v digitalni tehniki tiska. *Papir*, Nu. 1-2, pp. 14-23, 2000

[19] Nazhad MM, Hariss EJ, Dodson CTJ, Kerekes RJ. The Influence of Formation on Tensile Strenght of Paper Made from Mechanical Pulp. *TAPPI Journal*, Dec, pp. 1-9, 2010

[20] Bierman CJ. *Handbook of Pulping and Papermaking*. 2nd ed., San Diego : Academic Press, pp. 174-180, 1996

[21] Abdessalem SB, Jedda H, Skhiri S, Dahmen J, Boughamoura H. Improvement of Mechanical Performances of Braided Polyester Sutures. *AUTEX Research Journal*, Vol. 6, Nu. 3, pp. 169-174, 2006

[22] Lichtblau D, Strlič M, Trafela T, Kolar J, Anders M. Determination of Mechanical Properties of Historical Paper Based on NIR Spectroscopy and

Chemometrics – a New Instrument [online]. Available at: http://eprints.ucl.ac.uk/8000/1/8000.pdf; [accessed November 2011]

[23] Morent R, De Geyter N, Leys C, Gengembre L, Payen E. Comparison between XPS- and FTIR-Analysis of Plasma-treated Polypropylene Film Surfaces. *Surface Interface Analysis*, Vol. 40, pp. 597-600, 2007

[24] Tomšič B, Simončič B, Vince J, Orel B, Vilčnik A, Fir M, Šurca Vuk A, Jovanovski V. Uporaba ATR IR spektroskopije pri preučevanju strukturnih sprememb celuloznih vlaken. *Tekstilec*, Vol. 50, No. 1-3, pp. 3-15, 2007

[25] Kotnarowska D, Wojtyniak M. Influence of Aging on Mechanical Properties of Epoxy Coatings. *Solid State Phenomena*, Vol. 147-149, pp. 825-830, 2009

Experimental Verification of Cusp Heights when 3D Milling Rounded Surfaces

Balázs Mikó[1], Jozef Beňo[2], Ildikó Maňková[2]

[1]Óbuda University Budapest, Donát Bánki Faculty of Mechanical and Safety Engineering, Népszínház u. 8, 1084 Budapest, Hungary
miko.balazs@bgk.uni-obuda.hu

[2]Technical University Košice, Faculty of Mechanical Engineering, Mäsiarska 74, 040 01 Košice, Slovakia, jozef.beno@tuke.sk, ildiko.mankova@tuke.sk

Abstract: The paper deals with the experimental verification of cusp height when finishing an elementary surface by end ball milling. The relationship expressing the effect of the direction of milling cutter motion on cusp heights has been derived from the geometrical interpretation of inclined elementary surface. Experimental verification has been carried out as 3D milling of surface with definite roundness while surface roughness was measured considering normal vectors of tool-workpiece contact. An approach based on process window has been used to evaluate the measured data.

Keywords: ball end milling; cusp height; roughness data; process window

1 Introduction

Tool making and mould production technology became a key aspect in product innovation because of wide scale application of design features based on formed and free-form surfaces. Software development, CAD and CAM products, programming and control of machine tools as well as advanced metal cutting tools represent main driving forces associated with mould making in various industrial sectors. The production of both formed and free-form surfaces by milling has replaced traditional technologies such as electro-discharge machining and electro-chemical machining, the latter being of high power consumption rate. This fact is a very influential factor and hence contributes to the reduction of product life cycles and an increasing innovation rate in machinery products [1].

Formed and free-form surfaces result from the required shape of the final engineering component in the design of tooling and molding, whereas the form of any surface is usually generated by means of CAD, and are subsequently transferred into the control unit of any CNC milling machine. According to the current state of the art, the scientific sources characterizing the field of formed surfaces may be classified into five main fields as follows:

1 Parametric and geometrical definition of any formed and free-form surface; formal expression, surface slicing, surface segmentation, etc.

2 Decomposition of any surface for the control unit of machine tool motion, and that is subject of programming of tool paths, which include the choice of metal removal strategy, the visual representation of metal removal by tool edge, etc.

3 Graphical and analytical expression of machined surface texture generated by the tool, machined surface inclination in terms of the normal vectors, surface generation due to tool motion, etc.

4 Modeling of material removal as well as establishing of various dependencies based either on one quantity or more; while cutting force components and tool wear represent the data widely discussed in scientific sources.

5 Time of machining of any formed surface, cutting conditions, the influence of cutting conditions on a machined surface, the resultant surface quality and its representation.

On the other hand, the tool edge of metal cutting tools (end milling cutter, end ball milling cutter for formed surface semi-finishing) are considered to be the factors of technology of removal, see in Fig. 1, while influences of tool edge shape/geometry on resultant surface quality of produced surfaces are discussed, e.g. in [2].

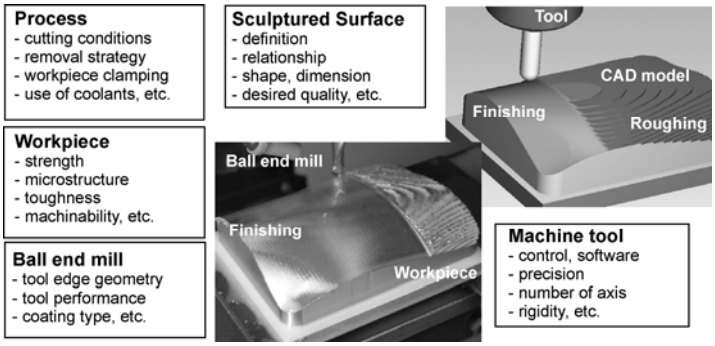


Figure 1

Factors affecting the quality of the product when machining with ball end milling

Based on the factors shown in Fig 1, the following notes are added to the machining of any free-form surface. First, commercial CAD/CAM systems provide various capabilities for surface design and process planning. Second, an efficient process plan requires the establishment of proper operation sequences, including removal strategies, that is, tooling and tool performance. In other words, no firm rules are available to enter any removal strategy for different manufacturability of designed free-form surface. Finally, yet more importantly, the economics of free-form machining imply as well the consideration of the tool performance, surface quality, and production costs.

2 Review of some Recent Results

The machined surface belongs to the most discussed subjects regarding product quality when milling by ball end mill cutters. Due to variability in technology such as cutting conditions, tool material/tool shape, milling strategy, etc., the resultant surface roughness varies widely by being affected by the type of produced surface and workpiece as well; see sources given in Tab. 1. Becze [4] showed that the roughness range varies due to the design of the die cavity, which differentiates it from straight surface, wall, and corner. As realized by [5], three common values of roughness exist which distinguishes the level of finishing, and the best of them, $Ra \approx 0.5$ μm refers to equivalent of manual polishing. According to [6], there is no uniformity of roughness range because of the coating applied to end mill cutter. In general, the best surface quality ($Ra = 0.5 \div 0.8$ μm) results from milling with CBN tools. As found by [7], the resultant roughness depends on tool deflection. Because of very specific appearance of machined surface, the so called topomorphy was introduced by [8]. However, the outer topomorphy by [8] differs greatly between up milling and down milling. Roughness data by [8] from Table 1 characterizes the former and the latter.

Table 1
Some roughness data resulting from ball end milling

Source	Workpiece	Type of surface	Roughness [μm]
Gilles et al. [3]	C 35E steel	planar	Ra =0.90 – 1.26
Becze et al. [4]	Steel 62 HRC	die cavity	Ra = 0.22 – 1.22
Baptista, Simoes [5]	Aluminum	convex, concave	Ra = 0.29 – 3.10
Fallböhmer et al. [6]	P20 steel	planar	Ra = 0.80 – 5.00
de Lacalle et al. [7]	Steel 62 HRC	complex form	Ra = 1.47 – 2.79
Antoniadis et al. [8]	C 60 steel	planar, oblique	Rz = 1.00 – 4.00

The next feature of surface quality during 3D milling is tolerance, often called the allowed errors of both dimension and shape. These tolerances determine the conditions when generating tool path, especially the tolerance for path interpolation and the tolerance of cusp height [9]. The generation of scallop height has been studied in [10] and its dimension varies within app 13.5 ÷ 2000 μm in 3D milling. Jerard et al. [11] pointed out that cutting simulation error is caused by deviation between the actual surface and the polyhedral approximation and protrusion of the tool between the surface points. When the polyhedral approximation lies inside the surface, the simulation error is the sum of the former and the latter. Finally, the resultant tolerance follows from ball end mill diameter and protrusion of the tool between the surface points. Narita et al. [12] showed that machining errors result from the deformation of the used milling tool. Similarly, Kim et al. [13] developed a method of surface error calculation based on beam deflection. Form errors from app zero to app 0.2 mm depend mainly on the inclination angle of machined surface, as the former (error app zero) refers to a

plain surface and the latter refers to inclination angle of 60 deg. Fallböhmer et al. [6] introduced average dimensional errors within 0.02 ÷ 0.06 mm for injection moulds and stamping dies, respectively. Nevertheless, an improved surface finish can be achieved either through an increased number of finishing paths or with a larger diameter cutter. According to [6], cusp height (Ch) is identical with theoretical surface roughness R_{th}. De Lacalle et al. [7] introduced a new methodology for the selection of the tool paths on complex surfaces that minimize dimensional errors due to tool defection in three axes milling. The methodology enabled dimensional errors to fall from 30 mm to below 4 mm in 3D milling. Dimensional errors, however, depend on the deflection force which is defined to be perpendicular to the tool axis and projected on the plane defined by the tool axis and the normal vector to the machined surface.

In addition to the factors associated with surface quality that have been discussed above (roughness, cusp/scallop height, tolerances and errors due to machining), there is another factor in the research of 3D milling of both formed and free-form surfaces, i.e. the design of the tested workpiece. There is no uniqueness in designing a tested workpiece as it varies in quite different shapes. A tested workpiece as a planar object gives results as tool axis orientation, transversal cutting force, tool vibrations, and surface roughness [3]. Planar surface of tested object enables us to use specific techniques such as an inclined slot cutting test [13]. The combination of two such surfaces as convex and/or concave denotes another possibility to combine milling of part of any free form surface, for instance [10], or machined surface enables us to apply so called wavelet-based multi-resolution representation of a series of intermediate shape models according to [14]. The inclined planar surface gained wide applicability in 3D milling; however, the inclination angle of a planar surface is very manageable for copying a very small portion of free form surface. The inclination angle of a planar surface can be equal to 10 ÷ 25 deg [15], though greater inclination up to 75 deg does appear, too. If a machined surface is defined as a portion of a cylinder of axis y and radius R, such setting enables us to combine planar/cylindrical surfaces, as in e.g. [16], while ramping and contouring strategies are applied in preference. However, the testing part may consist of quite different surfaces, such as e.g. surfaces which are smooth and dragged, and hemispheric ones [7].

This contribution deals with the investigation of surface roughness based on the calculation of true cusp height Ch while tested part consists of cylindrical surface with defined rounding. It is commonly known, e.g. from [6], that the dimension of the ball end milling cutter is often limited by the part geometry, and the theoretical surface roughness can only be minimized by decreasing the step-over distance a_e [mm]. On the other hand, the tool path needs not be a straight line, as for instance in "ramping" or in "contouring" strategies. If any element is taken off from the free form surface, it consists of an inclined part, the former allows us to model free form surface formation as end milling cutter motion with different directions A in finishing of the free form surface.

3 Geometrical Interpretation of Cusp Height

Let us consider a planar surface from Fig. 2a with inclination angle α_1 to its base. If the ball milling cutter moves along an inclined surface of a workpiece, the effective diameter of the tool leaves tool edge impressions on the newly machined surface. The tool edge impression consists of two elements: one is the cusp bottom line left by the axis of tool rotation and the other is the true cusp height (Ch) due to radial step-over a_e [mm], a factor resulting from machine tool programming. From the point of view of machining, an end ball milling cutter of effective diameter D_e removes axial depth of cut a_p [mm]

If α_1 equals to zero, the plain surface resulting from removal of a_p achieves theoretical cusp height Ch as:

$$Ch = \frac{D_e}{2} - \sqrt{\left(\frac{D_e}{2}\right)^2 - \left(\frac{a_e}{2}\right)^2} \tag{1}$$

If angle $\alpha_1 \neq 0$, and obviously $\alpha_1 > 0$, the theoretical cusp height depends on α_1 as:

$$Ch = \frac{D_e}{2} - \sqrt{\left(\frac{D_e}{2}\right)^2 - \left(\frac{a_e}{2.\cos\alpha_1}\right)^2} \tag{2}$$

Let us consider the cusp bottom line shown in Fig. 2b, an inclined elementary surface. Any point at cusp bottom line from Fig. 2b includes the normal unit vector **N** expressing the relationship of the cusp bottom line to the tool axis. The inclination angle α_1 has respect to triangle ECD whereas ED is another cusp bottom line. The motion of the tool axis along line DE refers to removal by "ramping", and that means the direction of milling A = 90 deg. The triangle BDE includes line BD, which has respect to removal by "contouring" and that assumes step-over a_e being perpendicular to BD, i.e., A = 0 deg. Variable direction of motion for ball end milling cutter can be found within A = 0 ÷ 90 deg, and therefore, the true cusp height must differ from those of Equations (1) and (2) because of vector **N**. Thus, the actual inclination angle results from angle A as a simple ratio:

$$\operatorname{tg}\alpha_3 = \overline{CE} / \overline{CF} \tag{3}$$

or

$$\alpha_3 = \operatorname{arctg}\,[\sin A.\ \operatorname{tg}\alpha_1] \tag{4}$$

while

$$\overline{CD} = \overline{CE}\ \operatorname{tg}\alpha_1 \tag{5}$$

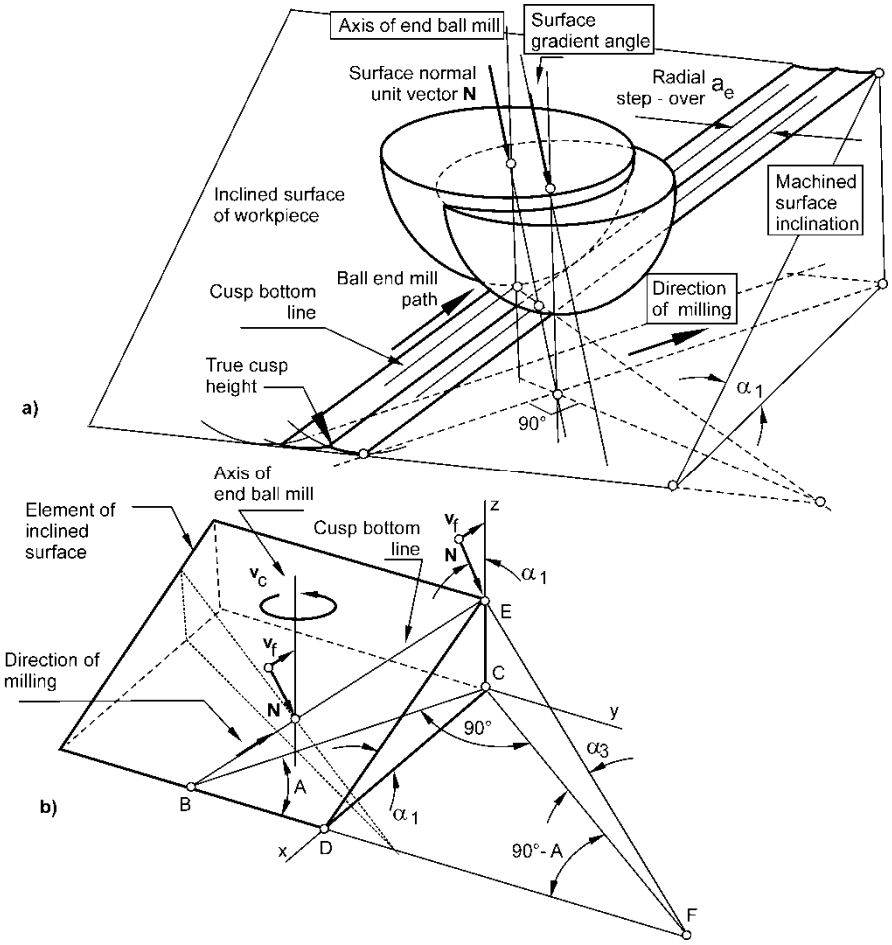


Figure 2

Geometrical interpretation of cusp formation: a) inclined surface with common direction of milling motion; b) element of inclined surface with cusp bottom line as the way of expressing true cusp height

and

$$\overline{CF} = \frac{\overline{CE}}{\sin A.\mathrm{tg}\,\alpha_1} \tag{6}$$

Thus, combining Equations (2) with those of (3) ÷ (6), the true cusp height depending on angle A will be as follows:

$$Ch = \frac{D_e}{2} - \sqrt{\left(\frac{D_e}{2}\right)^2 - \left(\frac{a_e}{2.\cos\alpha_3}\right)^2} \tag{7}$$

Let us consider the boundary cases of using Equation (7). If A = 90°, or milling of elementary surface in "ramping" strategy of milling, the cusp height does not depend on the vector **N**. If A = 0°, the angle α_3 depends on the roundness of the machined surface, i.e. cusp height Ch depends on the actual angle α_3 derived from α_1. If A = 45°, the true angle $\alpha_3 = \alpha_1$; the latter, however, is not identical with that of from as shown in Fig. 2a. In this case, the true cusp height is not constant.

4 Experimental Study and Process Window Approach

The scope of our experimental research was to determine the effect of the variables in 3D milling on surface roughness data as well as to compare the results with the calculation of true cusp heights according to Equation (7). The tested parts of dimensions 175 x 165 x 45 mm have been designed in such a way that enabled pre−machining of cylindrical semi surfaces with 100 mm surface rounding. It is shown in Fig. 3.

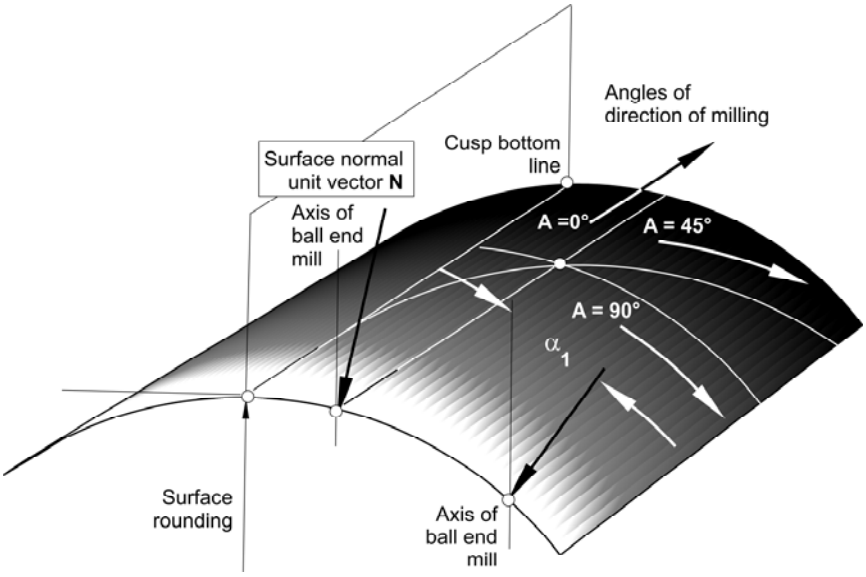


Figure 3

Designed semi-cylindrical surface for experimental research of surface roughness and true cusp height

The semi-cylindrical surface was made from 42CrMo4 steel (1.7225) by rough milling, leaving an allowance of a_p = 0.2 mm for surface finishing. Ball end milling cutters of diameter D = 12 mm with number of edges z = 2 (Fraisa U5286.501) were chosen for the surface finishing. The needed CNC programming was done by Pro/Engineer WF4 CAM and experimental work were conducted by

Mazak Nexus 410A−II machining centre. The surface roughness data were measured with a Mitutoyo Surftest SJ−301 machine.

The first step of exploring the true cusp height was to calculate Ch by the normal vector angle α_1 and milling direction A. Three steps-over, a_e = 0.20, 0.50, and 0.80 mm, as well as three angles, A = 0°, 45° and 90°, in Fig. 3 have been chosen at the outset for cusp height's not depending on cutting conditions. Of cutting conditions, the feed rate v_f [mm/min] is the most influencing quantity for involving feed per tooth f_z [mm] and spindle revolutions n [1/min], and that are factors being set by the machine tool. Thus, three feed rates (v_f = 630; 950 and 1265 [mm/min]) have been used. Thus, the applicable space of exploring quantities is shown in Fig. 4, assuming angles α_1 and α_3 are constant.

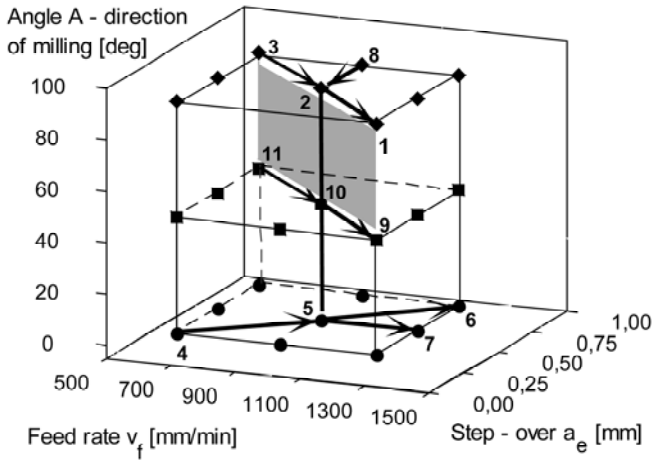


Figure 4
Applicable space of exploring quantities applied to experimental research

Roundness of machined semi-cylindrical surface by Fig. 3 have been prepared in such a way as to enable variation of angle α_1 within $0 \div 45°$. Five values of angle α_1 from Fig. 5 have been used in experimental works for covering completely roundness of machined of only half of semi−cylindrical surface.

Fig. 4 implies that the space of three variables $v_f - a_e - A$ gives $N = 3^3$ measuring points, which correspond to the only value of angle α_1. If all angles α_1 are assumed to be included into the aim of work, the whole number of measurement would have been equal to N = 185, providing that replication of measuring equals to number one. Therefore, an approach based on process windows has been used to test true cusp height, which combines part of applicable space from Fig. 4 with proper arrangement of measuring shown in Fig. 5.

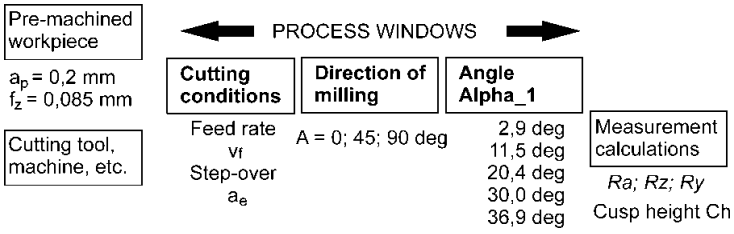


Figure 5

Basic arrangement of influencing quantities producing process windows

Applicable space of exploring quantities in Fig. 4 may be divided into three levels, while A = 45 deg produces a boundary between two strategies from Fig. 3, i.e. "ramping" strategy wherein A = 90 deg ($\alpha_3 = 0$) and "contouring" strategy with A = 0 deg ($\alpha_3 = 2.9$, 11.5, 20.4, 30, 36.9 deg). It must be noted that A = 45 deg gives $\alpha_3 = 2.05$, 8.19, 14.73, 22.21 and 27.96 deg. In order to reduce the great number of measurements, three windows express the relationship between variables and measured (or calculated) data, while the numbers in Fig. 4 label the measuring points:

− axis of space in Fig. 4, which is created by line 5 − 10 − 2

− panel of space in Fig. 4, which is created by line 1 − 3 − 9 − 11

− diagonals at the bottom/top rectangles complemented by single checking point.

The purpose of windows taken off from space of exploring is to explore the relationships as well as proper comparisons. The former and the latter are very suitable tools for choosing a proper machining strategy for expressing circumstances where undesirable cusp heights appear and thus contribute to the reduction of additional finishing operations when 3D milling.

5 Evaluation Results

Overall results include measured data (roughness average *Ra,* average maximum height of profile *Rz,* maximum height of profile *Ry*) and calculated values of Ch. The former can be reviewed either as qualitative understanding or as any form of dependence. Quantitative understanding means two ways of data expressing. The first is the combination of angles A, α_1 and α_3 and determines the surface without needing additional finishing. Three points in Fig. 4 (i.e., 1, 4 and 10), all of them performing metal removal with a_e = 0.2 mm, produce roughness *Ra* no greater than 1.50 μm, a condition where there is no need to apply additional surface finishing, see in [5]. However, such a condition must be accompanied with further data as shown in Tab. 2. It is obvious that there is no uniformity in ranging of the further roughness data and calculated cusp Ch does not fit with measured *Ry* either. All cases in Tab. 2 return Ch within a range 0.83 ÷ 1.30 μm, i.e. *Ry* >> Ch.

On the other hand, the unacceptable cases as two points 6 and 8 have been found out in Fig. 4, where measured $Ry > 20$ μm, and Ch = 13.35 ÷ 20 μm. Finally, point 11 in Fig. 4 is the only case wherein measured *Ry* (13.32 ÷ 19.35 μm) approaches in very way to the calculated Ch (13.37 ÷ 17.12 μm).

Table 2

The ranges of the best roughness data depending on milling direction

A = 90° v_f = 1265 m/min	A = 0°; v_f = 630 m/min	A = 45°;v_f = 1265 m/min
Ra = 0.53 ÷ 0.93 μm	*Ra* = 0.62 ÷ 1.22 μm	*Ra* = 0.67 ÷ 1.25 μm
Rz = 2.48 ÷ 3.45 μm	*Rz* =4.03 ÷ 6.22 μm	*Rz* = 3.42 ÷ 6.33 μm
Ry = 2.77 ÷4.05 μm	*Ry* =6.52 ÷ 11.05 μm	*Ry* = 3.96 ÷ 8.25 μm

Axis of the explored space, line 5 – 10 – 2 gives a way of expressing the relationship between the surface data, and hence can be considered as the one of various process windows. The measurement of machined surface leads to the well-known *Ra* vs. *Rz* relationship which characterizes either machining process (e.g. turning, etc.), or applied tool material [17]. If the position of the edge of a rounded ball end mill depends on normal angles α_1 and α_3, angle of milling direction A has two definite consequences see in Fig. 6.

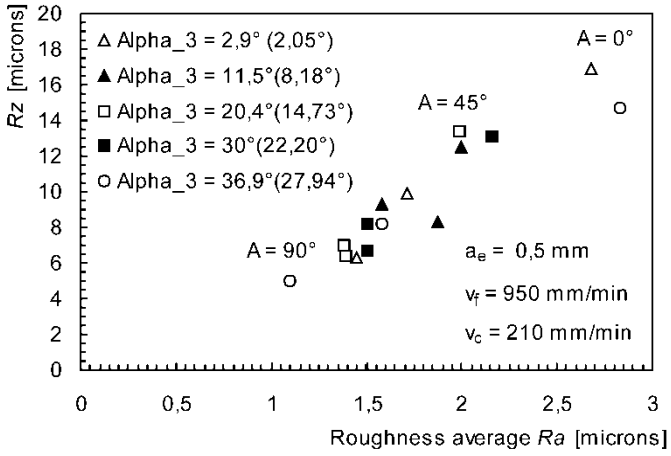


Figure 6

Relationship between *Ra* and *Rz* when milling surface with definite roundness (angle α_3 in parentheses are valid for A = 45 deg only)

The first consequence of the above is that the relationships in Fig. 6 are expressible in such a way as those in common theory of machining, though the dividing line is given by angle A. Considering angle α_3, another consequence is that the greater α_3 is the better the final roughness occurs without appearing tool edge marks. Concerning *Ra*, data resulting from A = 90° draw on results from [3] and [4] but data *Rz* are comparable with those from [8].

Panel 1 − 3 − 9 − 11 in Fig. 4 displays another process window which is treated particularly as a comparison. Let us follow the line 1 − 9 expressing the influence of direction of milling on cusp height A on both *Ry* and Ch. While calculation of Ch gives results of order 10^{-1}, the measured roughness – irrespective of *Rz* or *Ry* – appears in or of 10^{0} as shown in Fig. 7.

Another fact is that the removal by "ramping" strategy of milling (A = 90 deg) is not affected by angle α_3, i.e. the maximum height of the profile holds roughly constant. If the direction of milling is A = 45°, the resultant maximum height of profile *Ry* is strongly reduced by the angle $\alpha_3 = \alpha_1$. Nevertheless, it can be seen that there is a very different variation span of roughness in terms of values *Ry*. The greater angle α_3, as a rule, reduces *Ry* due to true contact between rounded tool edges of ball end mill cutter. However, the distortion of resultant roughness results from marks of the cutter left on the machined surface [18].

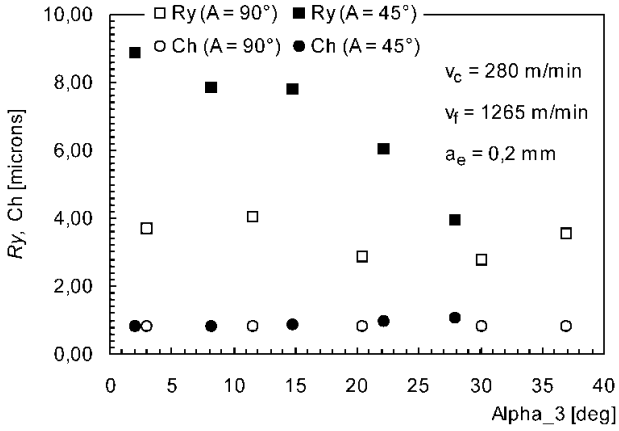


Figure 7

Process window as effect of angle α_3 on maximum height of the profile *Ry* and its comparison with calculated cusp height Ch

Fig. 8 shows the overall panel 1 − 3 − 9 – 11, and that is the process window expressing spans of roughness *R;* the latter depends on both angles α_3 and A. There are two points: 3 (thick line) and 11 (bold line) in Fig. 8 wherein measured *Ry* values roughly approach the calculated cusp height Ch. However, the wide span of measured *Ry* from point 3 results rather from a small cutting speed. It is obvious that the increase of feed rate (and cutting speed) brings about a reduction of span of measured *Ry* data if A = 90 deg, whereas calculated Ch is far from the measured *Ry* in such case. Nevertheless, the variation of angle α_3 along the surface with definite roundness shows quite different effects when the direction of tool motion is equal to A = 45°. As indicated by bold lines in Fig. 8, there is perhaps a reversal progress of spans in the measurement of *Ry*: the greater the feed rate, the greater are the spans of the measured roughness values *Ry*.

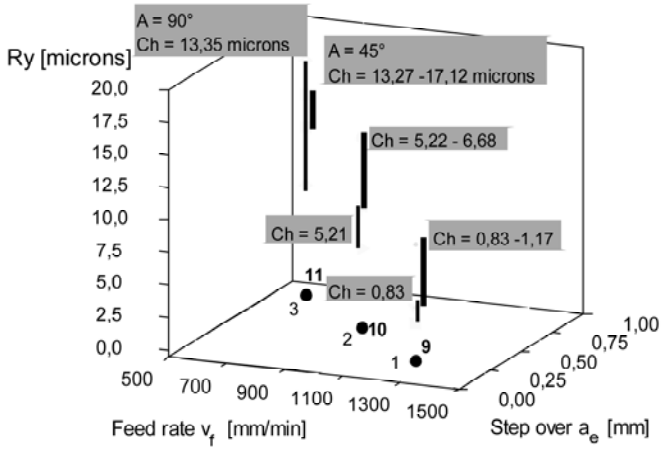


Figure 8

Spans of measured *Ry* and their comparison with calculated cusp height Ch

Figs. 9a and 9b introduce a way of reducing cusp height due to angle α_3, the variable appearing in the calculation of Ch. In the case of A = 90 deg from Fig. 9a, the chosen angles α_3 of overall process window 1 − 3 − 9 − 11 give the sequence of how tool edge marks merge into resultant tool edge footprint producing desired roughness *Ra*. The results from A = 45° shown in Fig. 9a give a similar sequence, yet yield better data of *Ra* without evidence of tool edge marks for point No. 9 of process window 1 − 3 − 9 − 11. On the other hand, the disappearance of tool edge marks due to angle α_3 means no direct improvement in the quality of the machined surface as shown in Fig. 9b. The resultant roughness still depends on step–over a_e, and in such case the calculated Ch fits well to measured roughness values *Ry*.

6 Discussion and Outlook

The bottom panel of exploring space, or points 3 – 4 – 5 – 6 – 7 with A = 0° give very poor correlation with the calculated value of Ch with measured values of *Ry*. Nevertheless, the smallest feed rate gives the best results of *Ra* shown in Tab. 2. There is always a possibility to reduce resultant *Ra* by increasing the feed rate up to 1265 m/min; however, step-over must be held constant in such a case. Once the proper agreement of Ch with measured *Ry* has been discovered, when A equals to 45 deg, a point 11 of exploring space. Such a result, however, can be expected when the smallest step-over a_e is used. The angle A = 90 gives an approach Ch to measured *Ry* twice for Ch not being affected by angle α_3 in such case. There is only the highest feed rate which avoids appearance of undesired cusp height Ch, see e.g. data from Tab. 2.

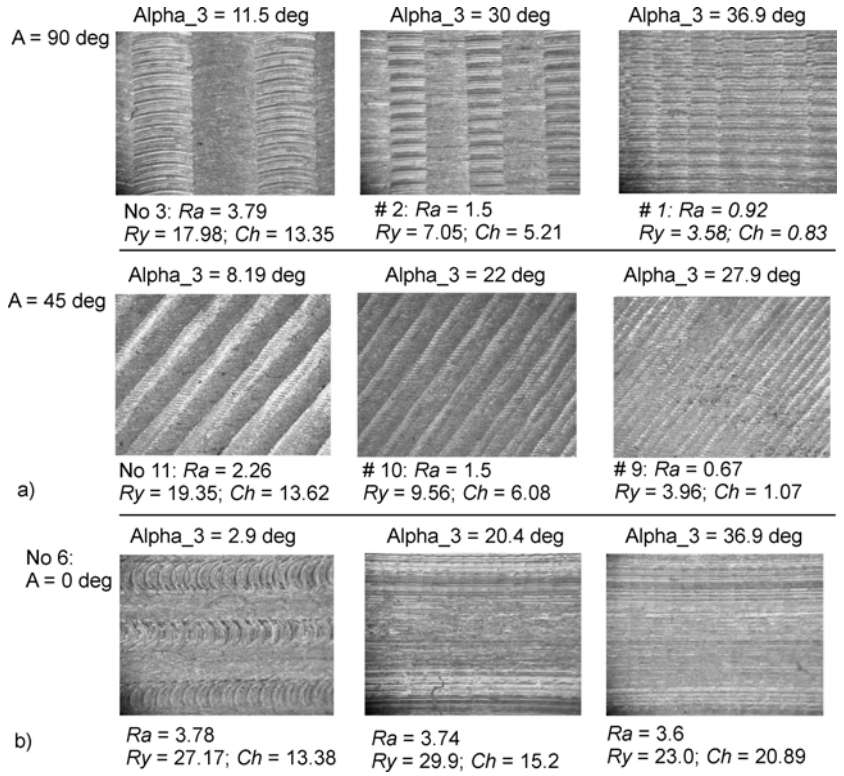


Figure 9

Effect of cutting conditions (feed rate v_f and step – over a_e) on resultant surface texture being affected by angle α_3: a) process window from panel 1 − 3 − 9 − 11 from exploring space; b) surface texture in point 6 of exploring space as a fragment of bottom process window (all roughness data in microns)

Broadly speaking, an arbitrary strategy of removal by Fig. 3 enables us to obtain a desired roughness that avoids further finishing operations; however, there are roughness variations for angle α_3 changed progressively. Thus, not only does the measured roughness value characterize the machined surface of definite rounding. An option is to indicate true results *Ra, Rz, Ry*, etc., for instance, *Ra* = 0.61/30°, *Ry*/25°, and so on. Another option is to display the results as spans in terms of Fig. 8. However, the surface normal vectors must be accompanied with measured roughness values as shown in Fig. 6.

An examination of about sixty microscopic photo shots of surface texture, as well as the examination of measured data, showed that the achieved results conform well to data presented in Tab. 1. For instance, the measured results $Rz = 2.50 \div 3.22$ from upper panel of exploring space (point No. 1 in Fig. 4) match well to results from [8], where obtained data are based on the milling of an oblique surface. As far as roughness average in molding and tool making, the best *Ra* data obtained never exceeded those from [4].

Because of the wide flexibility of designed free form surface, measured roughness represents only one criterion of product quality. Thus, further research will aim to examine mechanisms that accompany the formation of free form surface due to the variability of cutting forces. Surface formation and related forces shall contribute to explain formation of precision of any free form surfaces. The initial steps were carried out and presented in [18].

Conclusions

Strategies chosen when machining free-form surfaces affect the resulting quality of the final product as well as cost per piece. Suitable combinations of strategies reduce total machining time, and do influence the surface quality of the free form surface. Thus, based on achieved experimental results, the following conclusions can be drawn.

- Considerable effects of surface normal vector on resultant roughness produced by ball end milling tool have been found. The greater the leading angle of surface normal, the better the surface finish with definite roundness.
- The desired surface quality results from merging tool edge marks into continuous surface based on the tool footprint because of suppressing cusp height. Such effect is accomplished by combinations between cutting conditions (a_e, v_f) and vector N.
- Evaluation by the Process Window Approach showed that there are diverse effects of angle A, direction of milling surface with defined rounding. Reversal progress of roughness of *Ry* spans was proved when applying the angles A = 45 and A = 90 deg.
- Calculated cusp height Ch approaches exceptionally to the measured roughness *Ry* because of the surface formation's mechanisms in front of the tool edges. An improvement in surface finish appears due to merging of tool edge marks.

Acknowledgement

Research works have been supported by Agency of Research and Development APPV, under contracts No DO7RP−0014−09 and contracts of Bilateral cooperation Slovakia – Hungary SK−HU 0015−08 IMPRICAM. The Slovak authors express their thanks for projects VEGA 1/0500/12 “Quality improvement when milling form surfaces by advanced milling tools” and VEGA No. 1/0279/11 "Integration of trials numerical simulation and neural network to predict cutting tool performance", supported by Scientific Grant Agency of the Ministry of Education, Science and Research of Slovakia.

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

References

[1] Byrne G. et al. Advancing Cutting Technology, Annals of the CIRP, 52 (2003) 2, 483-507

[2] Taylan A et al.: Manufacturing of Dies and Molds. CIRP Annals − Manufacturing Technology Volume 50, 2 (2001) pp. 404-422

[3] Gilles P. et al: Dynamic Behaviour Improvement for a Torus Milling Cutter using Balance of the Transversal Cutting Force. Int J Adv Manuf Technol (2009) 40:669-675

[4] Becze, C. E. et al.: High-Speed Five-Axis Milling of Hardened Tool Steel International Journal of Machine Tools & Manufacture 40 (2000) 869-885

[5] Baptista, R, Antune Simoes, J. F.: Three and Five Axes Milling of Sculptured Surfaces. Journal of Materials Processing Technology 103 (2000) 398-403

[6] Fallböhmer, P. et al.: High-Speed Machining of Cast Iron and Alloy Steels for Die and Mold Manufacturing. Journal of Materials Processing Technology 98 (2000) 104-115

[7] Lopez de Lacalle L. N. et al.: Toolpath Selection Based on the Minimum Deflection Cutting Forces in the Programming of Complex surfaces Milling. International Journal of Machine Tools & Manufacture 47 (2007) 388-400

[8] Antoniadis A. et al.: Prediction of Surface Topomorphy and Roughness in Ball-End Milling. Int J Adv Manuf Technol (2003) 21:965-971

[9] Lartigue, C. et al.: CNC Tool Path in Terms of B-Spline Curves. Computer-aided Design 33 (2001) 307-319

[10] Warkentin, A. et al.: Comparison between Multi-Point and Other 5-Axis Tool Positioning Strategies. International Journal of Machine Tools & Manufacture 40 (2000) 185-208

[11] Jerard, R. et al.: Methods for Detecting Errors in Numerically Controlled Machining of Sculptured Surfaces. IEEE Computer Graphics & Applications, 10 (1989) 1, 26-39

[12] Narita H. et al.: Trial-Less Using Virtual Machining Simulator for Ball End Milling Operations. JSME International Journal Series C, 49 (2006) 1, 50-55

[13] Ozturk B. et al.: Machining of Free-Form Surfaces. Part II: Calibration and Forces. International Journal of Machine Tools & Manufacture 46 (2006) 736-746

[14] Date H. et al.: Wavelet-based Multiresolution Representation of a Geometric Model for Free-Form Surface Machining. Proc. of the 2000

Japan−USA Flexible Automation Conf., July 23-26, 2000, Ann Arbor, Michigan, 2000JUSFA−13035, pp 1-8

[15] Imani, B. M. et al.: An Improved Process Simulation System for Ball-End Milling of Sculptured Surfaces. International Journal of Machine Tools & Manufacture 38 (1998) 1089-1107

[16] Kim G. M. et al.: Cutting Force Prediction of Sculptured Surface Ball-End Milling Using Z-Map. International Journal of Machine Tools & Manufacture 40 (2000) 277-291

[17] Beňo, J.: Theory of Innovative Technology. Vienala Kosice (2010) 175, pp. (in Slovak language)

[18] Ižol P., Beňo J., Mikó B.: Precision and Surface Roughness when Free Form Milling. Manufacturing Engineering, 10 (2011) 1, pp. 70-73

Elastic Analysis of Heterogeneous Thick Cylinders Subjected to Internal or External Pressure Using Shear Deformation Theory

Mehdi Ghannad

Mechanical Engineering Faculty, Shahrood University of Technology, University Boulevard, Haft Tir Square, Shahrood, Iran
email: mghannadk@shahroodut.ac.ir

Mohammad Zamani Nejad

Mechanical Engineering Department, Yasouj University, Daneshjo Street, P. O. Box: 75914-353, Yasouj, Iran
email: m_zamani@yu.ac.ir

Abstract: An analytical formulation based on the first-order shear deformation theory (FSDT) is presented for axisymmetric thick-walled heterogeneous cylinders under internal and external uniform pressure. It is assumed that the material is isotropic heterogeneous with constant Poisson's ratio and radially varying elastic modulus. First, general governing equations of the heterogeneous thick cylinders are derived by virtual work principle, and using FSDT. Then the obtained equations are solved under the generalized plane strain assumptions. The results are compared with the findings of both plane elasticity theory (PET) and finite element method (FEM).

Keywords: thick cylinder; shear deformation theory (SDT); heterogeneous; functionally graded material (FGM); finite element method (FEM)

1 Introduction

Axisymmetric hollow shells are important in industries. In order to optimize the weight, displacement and stress distribution of a shell, one approach is to use shells with Functionally Graded Materials. FGMs or heterogeneous materials are advanced composite materials with microscopically inhomogeneous character that are engineered to have a smooth spatial variation of continuous properties. The concept of FGMs was proposed by material scientists in Japan [1].

1.1 Homogeneous Cylinders

First Lamé (1852) found the stress distribution in an isotropic homogeneous hollow cylinder under uniform pressure.This solution has been extensively used to solve many engineering problems. Naghdi and Cooper [2] started with a Reissner's variational theorem and included the effects of shear deformation. The first order displacement field for thick cylindrical shells was expressed by Mirsky-Hermann [3] which is the extension of the Mindlin plate theory [4] and includes transverses shear deformation. Greenspon [5] compared the results of different theories of thick-walled cylindrical shells. Ziv and Perl [6] obtained the response of vibration analysis of a thick-walled cylindrical shell using FSDT theory and solved by finite difference method. Suzuki et al. [7] used the FSDT for vibration analysis of axisymmetric cylindrical shell with variable thickness. They assumed that the problem is in the state of plane stress and ignored the normal stress in the radial direction. Simkins [8] used the FSDT for determining displacement in a long and thick tube subjected to moving loads. Eipakchi et al. [9] used the FSDT for driving governing equations of thick cylinders with varying thickness and solved the equations with perturbation theory. Using FSDT, Ghannad and Zamani Nejad [10] present the general method for analysis of internally pressurized thick-walled cylindrical shells with clamped-clamped ends.

1.2 Heterogeneous Cylinders

Heterogeneous composite materials are functionally graded materials (FGMs) with gradient compositional variation of the constituents from one surface of the material to the other which results in continuously varying material properties. These materials are advanced, heat resisting, erosion and corrosion resistant, and have high fracture toughness. The FGMs concept is applicable to many industrial fields such as aerospace, nuclear energy, chemical plants, electronics, biomaterials, and so on. Fukui and Yamanaka [11] used the PET for the derivation of the governing equation of a thick-walled FGM tube under internal pressure and solved the obtained equation numerically by means of the Runge-Kutta method. Horgan and Chan [12] analyzed a pressurized hollow cylinder in the state of plane strain. The exact solution for stresses in FGM pressure vessels alone using Lamé's solution was provided by Tutuncu and Ozturk [13]. They assumed material stiffness obeys a simple power law through the wall thickness with Poisson's ratio being constant. In this reference formula and plot for circumferential stress are incorrect. Jabbari et al. [8] have presented a general analysis of one-dimensional steady-state thermal stresses in a hollow thick cylinder made of FGM. Hongjun et al. [14] and Zhifei et al. [15] provided elastic analysis and an exact solution for stresses in FGM hollow cylinders in the state of plane strain with isotropic multi-layers based on Lamé's solution. Thick-walled cylinders with exponentially-varying material properties were solved by Tutuncu [16]. Zamani Nejad et al. [17]

developed 3-D set of field equations of FGM thick shells of revolution in curvilinear coordinate system by tensor calculus. Ghannad et al. [18] provided a general axisymmetric solution of FGM cylinders based on PET in the state of plane stress, plane strain and closed cylinder. Abedi et al. [19] obtained a numerical solution using finite element method and a static analysis for stresses and displacements in FGM parabolic solid cylinder.

The following topics will be described; using FSDT, PET and FEM, the heterogeneous hollow cylinders have been solved and have been compared with homogenous cylinders.

2 Governing Equations

In the Plane Elasticity Theory (PET), axisymmetric thick cylinders with constant thickness and uniform pressure are analyzed by Lamé's solution in cylindrical coordinates. The radial displacement of this cylinder is given by:

$$u_r = C_1 r + \frac{C_2}{r} \tag{1}$$

where C_1 and C_2 are constants and r is the radius of cylinder. Consider FGM circular cylindrical shell shown in Fig. 1.

In this figure, P_i and P_o are internal and external pressures, R is the radius of the middle surface and z is the distance from the middle surface which ranges in such an interval as $\left(-h/2 \le z \le h/2\right)$, so one can write:

$$r = R + z \Rightarrow u_r = C_1(R+z) + \frac{C_2}{R+z} \tag{2}$$

If $\left|z/R\right| < 1$ and by Taylor expansion:

$$u_r = C_1(R+z) + \frac{C_2}{R}\left(1 - \frac{z}{R} + \frac{z^2}{R^2} - \frac{z^3}{R^3} + \cdots\right)$$

$$= \left(C_1 R + \frac{C_2}{R}\right) + \left(C_1 - \frac{C_2}{R^2}\right) z + \frac{C_2}{R^3} z^2 + \cdots \tag{3}$$

The radial displacement is:

$$u_r = u_0 + u_1 z + u_2 z^2 + \cdots \tag{4}$$

This means that the displacement can be written as a polynomial of z, and u_0 is the displacement of the middle surface (if $z=0$). h and L are the thickness and the length of the cylinder, in which r_i and r_o are inner and outer radiuses of the cylinder.

The general axisymmetric displacement field in FSDT can be expressed on the basis of axial displacement and radial displacement, as follows:

$$U_x = u(x) + \phi(x)z \ , \ U_\theta = 0 \ , \ U_z = w(x) + \psi(x)z \tag{5}$$

where $u(x)$ and $w(x)$ are the displacement components of the middle surface. Also, $\phi(x)$ and $\psi(x)$ are the functions used to determine the displacement field. The strain-displacement relations in the cylindrical coordinates system are:

$$\begin{cases} \varepsilon_x = \dfrac{\partial U_x}{\partial x} = \dfrac{du}{dx} + \dfrac{d\phi}{dx}z & \varepsilon_\theta = \dfrac{U_z}{r} = \dfrac{w}{R+z} + \dfrac{\psi}{R+z}z \\ \varepsilon_z = \dfrac{\partial U_z}{\partial z} = \psi & \gamma_{xz} = \dfrac{\partial U_x}{\partial z} + \dfrac{\partial U_z}{\partial x} = \left(\phi + \dfrac{dw}{dx}\right) + \dfrac{d\psi}{dx}z \end{cases} \tag{6}$$

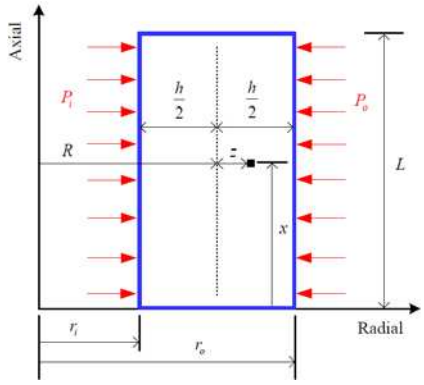


Figure 1
Geometry of the cylinder

The elastic modulus is assumed to vary as follows:

$$E(r) = E_i \overline{r}^n = E_i \left(\frac{r}{r_i}\right)^n \tag{7}$$

By substituting $r = R + z$ into Eq. (7), $E(z)$ is defined as:

$$E(z) = E_i \left(\frac{R+z}{r_i}\right)^n = \frac{E_i}{r_i^n}(R+z)^n \tag{8}$$

Here, E_i is Young's modulus of the inner surface and n is inhomogeneity constant. In the present paper n is assumed to range $-2 \le n \le 2$. Further, the Poisson's ratio υ is assumed a constant. Therefore, the stress-strain relations are:

$$\begin{cases} \begin{Bmatrix} \sigma_x \\ \sigma_\theta \\ \sigma_z \end{Bmatrix} = \dfrac{E(z)}{(1+\upsilon)(1-2\upsilon)} \begin{bmatrix} 1-\upsilon & \upsilon & \upsilon \\ \upsilon & 1-\upsilon & \upsilon \\ \upsilon & \upsilon & 1-\upsilon \end{bmatrix} \begin{Bmatrix} \varepsilon_x \\ \varepsilon_\theta \\ \varepsilon_z \end{Bmatrix} \\ \tau_{xz} = \dfrac{E(z)}{2(1+\upsilon)} \gamma_{xz} \end{cases} \tag{9}$$

Similarly, these can be written as:

$$\begin{cases} \sigma_i = \lambda E(z)\left[(1-\upsilon)\varepsilon_i + \upsilon(\varepsilon_j + \varepsilon_k)\right] \quad i \ne j \ne k \\ \tau_{xz} = \dfrac{1-2\upsilon}{2} \lambda E(z) \gamma_{xz} \quad , \quad \lambda = \dfrac{1}{(1+\upsilon)(1-2\upsilon)} \end{cases} \tag{10}$$

In order to drive the differential equations of equilibrium, the principle of virtual work has been used as:

$$\delta U = \delta W \tag{11}$$

where U is the total strain energy of the elastic body and W is the total external work due to internal pressure. The strain energy is:

$$\left\{ U = \iiint_V U^* dV \ , \ dV = r dr d\theta dx \quad \& \quad U^* = \frac{1}{2}\left(\sigma_x \varepsilon_x + \sigma_\theta \varepsilon_\theta + \sigma_z \varepsilon_z + \tau_{xz} \gamma_{xz}\right) \right. \tag{12}$$

and the external work is:

$$\left\{ W = \iint_S \left(\vec{f} \cdot \vec{u}\right) dS \ , \ dS = r d\theta dx \quad \& \quad \left(\vec{f} \cdot \vec{u}\right) dS = \left(P_i r_i - P_o r_o\right) U_z d\theta dx \right. \tag{13}$$

The variation of the strain energy is:

$$\delta U = R \int_0^{2\pi} \int_0^L \int_{-h/2}^{h/2} \delta U^* \left(1 + z/R\right) dz dx d\theta \tag{14a}$$

$$\Rightarrow \quad \frac{\delta U}{2\pi} = R \int_0^L \int_{-h/2}^{h/2} \left(\sigma_x \delta\varepsilon_x + \sigma_\theta \delta\varepsilon_\theta + \sigma_z \delta\varepsilon_z + \tau_{xz} \delta\gamma_{xz}\right)\left(1 + \frac{z}{R}\right) dz dx \tag{14b}$$

and the variation of the external work is:

$$\begin{cases} \delta W = \int_0^{2\pi}\int_0^{L}\left[P_i r_i - P_o r_o\right]\delta U_z\, dx d\theta \\ \Rightarrow \dfrac{\delta W}{2\pi} = \int_0^{L}\left[P_i\left(R-h/2\right) - P_o\left(R+h/2\right)\right]\delta U_z\, dx \end{cases} \tag{15}$$

By substituting Eqs. (6), (8) and (9) into Eqs. (14) and (15) and by using Eq. (11) and carrying out the integration by parts, the equilibrium equations and the boundary conditions are obtained in the form of:

$$\begin{cases} R\dfrac{dN_x}{dx} = 0 \;\;,\;\; R\dfrac{dM_x}{dx} - RQ_x = 0 \\ R\dfrac{dQ_x}{dx} - N_\theta = -P_i\left(R-h/2\right) + P_o\left(R+h/2\right) \\ R\dfrac{dM_{xz}}{dx} - M_\theta - RN_z = h/2\left[P_i\left(R-h/2\right) + P_o\left(R+h/2\right)\right] \end{cases} \tag{16}$$

and

$$R\left[N_x\delta u + M_x\delta\phi + Q_x\delta w + M_{xz}\delta\psi\right]_0^L = 0 \tag{17}$$

respectively, where the axial force, bending moment and shear force resultants are defined as the shell theory by:

$$\begin{cases} \begin{Bmatrix} N_x \\ N_\theta \\ N_z \end{Bmatrix} = \int_{-h/2}^{h/2} \begin{Bmatrix} \sigma_x\left(1+z/R\right) \\ \sigma_\theta \\ \sigma_z\left(1+z/R\right) \end{Bmatrix} dz \quad , \quad \begin{Bmatrix} M_x \\ M_\theta \end{Bmatrix} = \int_{-h/2}^{h/2} \begin{Bmatrix} \sigma_x\left(1+z/R\right) \\ \sigma_\theta \end{Bmatrix} z dz \\ Q_x = \int_{-h/2}^{h/2} \tau_{xz}\left(1+z/R\right) dz \qquad , \qquad M_{xz} = \int_{-h/2}^{h/2} \tau_{xz}\left(1+z/R\right) z dz \end{cases} \tag{18}$$

Substituting for the stress components into Eqs. (18), the equilibrium equations of the shell can be written in the abbreviated form:

$$\begin{cases} \left[A_1\right]\dfrac{d^2}{dx^2}\{y\} + \left[A_2\right]\dfrac{d}{dx}\{y\} + \left[A_3\right]\{y\} = \{F\} \\ \{y\} = \begin{Bmatrix} u \\ \phi \\ w \\ \psi \end{Bmatrix} \quad \& \quad \{F\} = \dfrac{r_i^{(n+1)}}{\lambda E_i} \begin{Bmatrix} 0 \\ 0 \\ -P_i + kP_o \\ h/2\left(P_i + kP_o\right) \end{Bmatrix} \end{cases} \tag{19}$$

where $k = r_o/r_i$ is the radius ratio. The above equations are a set of inhomogeneous linear differential equations with constant coefficients, solved by

using theory of ordinary differential equations [20]. These equations have the general and particular solutions.

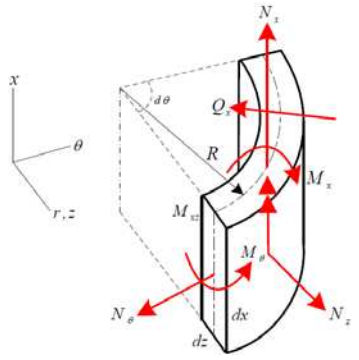


Figure 2
Graphical depiction of force and moment resultants

The general solution is in the form of $\{y\}=\{v\}e^{mx}$ and by substitution in homogeneous equations, one can calculate eigenvalues (m_i) and eigenvectors $(\{v\}_i)$.

$$e^{mx}\left[m^2[A_1]+m[A_2]+[A_3]\right]\{v\}=\{0\} \quad \Rightarrow \quad \left|m^2A_1+mA_2+A_3\right|=0 \tag{20}$$

Consequently the general solution is:

$$\{y\}_g=\sum_{i=1}^{6}C_i\{v\}_i e^{m_i x} \tag{21}$$

Finally, total solution is a summation of the general and particular solution.

$$\{y\}=\{y\}_g+\{y\}_p=\sum_{i=1}^{6}C_i\{v\}_i e^{m_i x}+\{K_0\} \tag{22}$$

In the state of plane strain, the solution of the cylinders in regions away from the boundaries is obtained. It means that the unknown vector $\{y\}$ is constant and all the terms which contain d/dx are removed.

$$[A_3]\{y\}=\{F\} \quad \Rightarrow \quad \{y\}=[A_3]^{-1}\{F\} \tag{23}$$

The solution of Eqs. (23) can be written as follows:

$$\begin{Bmatrix} w \\ \psi \end{Bmatrix}=\frac{r_i^{(n+1)}}{\lambda E_i}[A_3]^{-1}_{2\times 2}\begin{Bmatrix} -P_i+kP_o \\ h/2\left(P_i+kP_o\right) \end{Bmatrix} \tag{24}$$

The radial displacement is:

$$U_z = w + \psi z \quad \Rightarrow \quad u_r = (w - \psi R) + \psi r \tag{25}$$

The strains are:

$$\varepsilon_x = 0 \quad plane\ strain \quad , \quad \varepsilon_r = \frac{du_r}{dr} = \psi \quad , \quad \varepsilon_\theta = \frac{u_r}{r} = \psi + \frac{w - \psi R}{r} \tag{26}$$

The maximum stress in the cylinders is:

$$\sigma_{\max} = \sigma_\theta = \lambda E_i \overline{r}^n \left[\upsilon \varepsilon_r + (1-\upsilon)\varepsilon_\theta \right] \tag{27}$$

3 Solution of the Homogeneous Cylinders

In the isotropic homogeneous cylinders, Young's modulus and Poisson's ratio are constant. By setting $n = 0$ in Eq. (9), the elasticity modulus of homogeneous material is resulted.

$$E = cons \tag{28}$$

By substituting stress components into Eqs. (18), the force and moment resultants have been derived as follows:

$$\begin{cases} N_x = \lambda E h \left[(1-\upsilon)\left(\dfrac{du}{dx} + \dfrac{h^2}{12R}\dfrac{d\phi}{dx} \right) + \upsilon \left(\dfrac{w}{R} + \psi \right) \right] \\ N_\theta = \lambda E \left[\upsilon h \dfrac{du}{dx} + (1-\upsilon)\alpha w + \left(h - (1-\upsilon)R\alpha \right)\psi \right] \\ N_z = \lambda E h \left[\upsilon \left(\dfrac{du}{dx} + \dfrac{h^2}{12R}\dfrac{d\phi}{dx} \right) + \upsilon \dfrac{w}{R} + (1-\upsilon)\psi \right] \end{cases} \tag{29a}$$

$$\begin{cases} M_x = \lambda E \dfrac{h^3}{12R} \left[(1-\upsilon)\left(\dfrac{du}{dx} + R\dfrac{d\phi}{dx} \right) + 2\upsilon\psi \right] \\ M_\theta = \lambda E \left\{ \upsilon \dfrac{h^3}{12}\dfrac{d\phi}{dx} + (1-\upsilon)\left[(h - R\alpha)w + \left(R^2\alpha - Rh \right)\psi \right] \right\} \end{cases} \tag{29b}$$

$$\begin{cases} Q_x = K \dfrac{(1-2\upsilon)}{2} \lambda E h \left[\phi + \dfrac{dw}{dx} + \dfrac{h^2}{12R}\dfrac{d\psi}{dx} \right] \\ M_{xz} = K \dfrac{(1-2\upsilon)}{2} \lambda E \dfrac{h^3}{12R} \left[\phi + \dfrac{dw}{dx} + R\dfrac{d\psi}{dx} \right] \end{cases} \tag{29c}$$

K is the shear correction factor that is embedded in the shear stress term with an analogy to the Timoshenko beam theory. In the static state, for cylinders $K = 5/6$ [21].

By substituting above relations into Eqs. (16), the matrices of coefficient and the vector of force are defined as:

$$[A_1] = \begin{bmatrix} (1-\upsilon)Rh & (1-\upsilon)\frac{h^3}{12} & 0 & 0 \\ (1-\upsilon)\frac{h^3}{12} & (1-\upsilon)\frac{Rh^3}{12} & 0 & 0 \\ 0 & 0 & \mu Rh & \mu\frac{h^3}{12} \\ 0 & 0 & \mu\frac{h^3}{12} & \mu\frac{Rh^3}{12} \end{bmatrix} \tag{30a}$$

$$[A_2] = \begin{bmatrix} 0 & 0 & \upsilon h & \upsilon Rh \\ 0 & 0 & -\mu Rh & -(\mu-2\upsilon)\frac{h^3}{12} \\ -\upsilon h & \mu Rh & 0 & 0 \\ -\upsilon Rh & (\mu-2\upsilon)\frac{h^3}{12} & 0 & 0 \end{bmatrix} \tag{30b}$$

$$[A_3] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -\mu Rh & 0 & 0 \\ 0 & 0 & -(1-\upsilon)\alpha & -\left[h-(1-\upsilon)R\alpha\right] \\ 0 & 0 & -\left[h-(1-\upsilon)R\alpha\right] & -(1-\upsilon)R^2\alpha \end{bmatrix} \tag{30c}$$

$$\{F\} = \frac{r_i}{\lambda E}\left\{0 \quad 0 \quad -P_i + kP_o \quad h/2\left(P_i + kP_o\right)\right\}^T \tag{30d}$$

where, the parameters are as follows:

$$\begin{cases} \alpha = \ln\left(\dfrac{R+h/2}{R-h/2}\right) = \ln k \quad , \quad k = \dfrac{r_o}{r_i} \\ \mu = K\dfrac{(1-2\upsilon)}{2} \end{cases} \tag{31}$$

In the state of plane strain, the solution is obtained by Eq. (24) as follows:

$$\begin{Bmatrix} w \\ \psi \end{Bmatrix} = \frac{-r_i}{\lambda E} \begin{bmatrix} (1-\upsilon)\alpha & h-(1-\upsilon)R\alpha \\ h-(1-\upsilon)R\alpha & (1-\upsilon)R^2\alpha \end{bmatrix}^{-1} \begin{Bmatrix} -P_i + kP_o \\ h/2\left(P_i + kP_o\right) \end{Bmatrix} \tag{32}$$

The radial displacement on the basis of FSDT is:

$$u_r = \frac{\mathrm{r}_i^2}{\lambda E h\left[h-2(1-\upsilon)R\alpha\right]} \Big\{ \Big[h\left(P_i - kP_o\right) - (1-\upsilon)\left(P_i - k^2 P_o\right) r_i \alpha \Big] \bar{r} - kh\left(P_i - P_o\right) \Big\} \tag{33}$$

4 Comparisons between FSDT and PET

The radial displacement in the state of plane strain, on the basis of PET has been obtained [12] as follows

$$u_r = \frac{(1+\upsilon)P_i r_i \bar{r}}{E(k^2-1)} \left[(1-2\upsilon) + \frac{k^2}{\bar{r}^2} \right] \tag{34}$$

For the comparison between FSDT and PET, we assume a cylinder with an inner radius $r_i = 40$ mm, an outer radius $r_o = 60$ mm, Young's modulus $\mathrm{E_i} = 200$ GPa and Poisson's ratio $\upsilon = 0.3$ under an internal pressure of $P_i = 80$ MPa. The radial displacement along the thickness by both FSDT and PET has been calculated and plotted in Fig. 3.

Fig. 3 shows that the radial displacement calculated by the two methods is almost identical at the middle surface domain and it increases at the inner surface; it is less than 4% anyway. In order to evaluate the effect of the wall thickness on the radial displacement, Eqs. (33) and (34) are expressed on the basis of $\bar{h} = h/R$.

$$u_r^F = \frac{-P_i R\left(1-\bar{h}/2\right)}{\lambda E \bar{h}\left[\bar{h}-2(1-\upsilon)\alpha\right]} \left[\bar{h}^2 + (1-\upsilon)\left(1-\bar{h}/2\right)^2 \alpha \right] \ , \ \alpha = \ln\left(\frac{2+\bar{h}}{2-\bar{h}}\right) \ \text{FSDT} \tag{35}$$

$$u_r^P = \frac{(1+\upsilon)P_i R\left(1-\bar{h}/2\right)}{E(k^2-1)} \left[(1-2\upsilon) + k^2 \right] \ , \ k = \frac{2+\bar{h}}{2-\bar{h}} \quad \text{PET} \tag{36}$$

In Fig. 4, the percentage difference between FSDT and PET radial displacements $\left(Diff = \left(\left(u_r^P - u_r^F\right)/u_r^P\right) \times 100\right)$ has been shown. This difference is increased with an increase in the thickness of the cylinder. The maximum difference occurs at $1/20 \le h/R \le 16/20$ and reaches 15%, which is an acceptable value for the

analysis of thick cylinders. If the thickness of the wall equals radius of the middle surface $\left(\bar{h}=1\right)$, the difference reaches 25% .

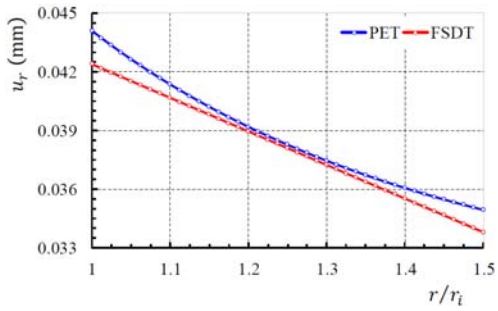


Figure 3
Distribution of radial displacement in homogeneous cylinder

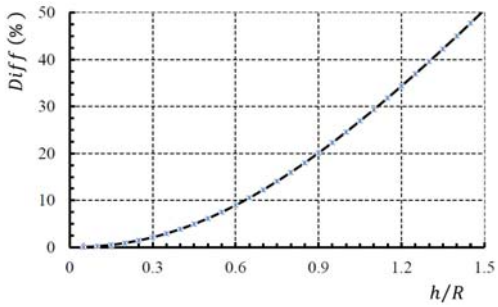


Figure 4
Difference percentages with respect to $\bar{h}=h/R$

5 Solution of the Heterogeneous Cylinders

In the isotropic heterogeneous cylinders, Poisson's ratio is constant and Young's modulus is calculated by inserting n in Eq. (8). In the current study, a range of $-2\le n\le 2$ is employed. By substituting stress components into Eqs. (18), the force and moment resultants are obtained. The matrices of coefficient in Eq. (19) are defined as the following form:

$$\begin{cases}\left[A_1\right]_{4\times4}=\left[a_{ij}\right] & Symmetric \quad , \quad \left[A_3\right]_{4\times4}=\left[c_{ij}\right] \quad Symmetric \\ \left[A_2\right]_{4\times4}=\left[b_{ij}\right] & Antisymmetric\end{cases} \tag{37}$$

Constants in above relations are:

$$\begin{cases} k=\dfrac{r_o}{r_i}\ ,\ \overline{r}=\dfrac{r}{r_i}\ ,\ \mu=K\dfrac{(1-2\upsilon)}{2} \\ \alpha=\ln\left(\dfrac{R+h/2}{R-h/2}\right)=\ln k\ ,\ \beta=\dfrac{h}{(R+h/2)(R-h/2)}=\dfrac{k-1}{kr_i} \end{cases} \tag{38}$$

5.1 Inhomogeneity Constant of $n=-2$

Elasticity modulus on the basis of Eq. (8) is:

$$E(z)=\frac{E_i r_i^2}{(R+z)^2} \tag{39}$$

Following above, nonzero components of the symmetric matrix $[A_1]_{4\times4}$ are:

$$\begin{cases} a_{11}=(1-\upsilon)\alpha & a_{22}=(1-\upsilon)(R^2\alpha-Rh) & a_{33}=\mu\alpha \\ a_{44}=\mu(R^2\alpha-Rh) & a_{12}=a_{21}=(1-\upsilon)(h-R\alpha) & a_{34}=a_{43}=\mu(h-R\alpha) \end{cases} \tag{40a}$$

and nonzero components of the antisymmetric matrix $[A_2]_{4\times4}$ are:

$$\begin{cases} b_{13}=-b_{31}=\upsilon\beta & b_{14}=-b_{41}=\upsilon(2\alpha-R\beta) \\ b_{23}=-b_{32}=-\mu\alpha+\upsilon(\alpha-R\beta) & b_{24}=-b_{42}=-\mu(h-R\alpha)+\upsilon(2h-3R\alpha+R^2\beta) \end{cases} \tag{40b}$$

and nonzero components of the symmetric matrix $[A_3]_{4\times4}$ are:

$$\begin{cases} c_{22}=-\mu\alpha & c_{33}=-(1-\upsilon)\dfrac{R}{h}\beta^2 \\ c_{44}=-2\alpha+(1+\upsilon)R\beta-(1-\upsilon)\dfrac{Rh}{4}\beta^2 & c_{34}=c_{43}=-\upsilon\beta+(1-\upsilon)\dfrac{h}{4}\beta^2 \end{cases} \tag{40c}$$

and the force vector $\{F\}_{4\times1}$ is:

$$\{F\}=\frac{1}{\lambda E_i r_i}\left\{0 \quad 0 \quad -P_i+kP_o \quad h/2\left(P_i+kP_o\right)\right\}^T \tag{40d}$$

Finally, the radial displacement on the basis of Eq. (25) is obtained as follows:

$$u_r=\frac{1}{\lambda E_i\left[1-2(1-\upsilon)\dfrac{R\alpha}{h}\right]\beta^2}\left\{\left[\upsilon\beta\left(P_i-kP_o\right)+(1-\upsilon)\left(P_i+k^2P_o\right)\frac{\beta^2 r_i}{2}\right]\overline{r}\right.$$

$$+\frac{2\alpha}{r_i}\left(-P_i+kP_o\right)+\beta\left(P_i-k^2P_o\right)\Bigg\} \tag{41}$$

5.2 Inhomogeneity Constant of $n=+2$

Elasticity modulus on the basis of Eq. (8) is:

$$E(z)=\frac{E_i}{r_i^2}\left(R+z\right)^2 \tag{42}$$

Following above, the nonzero components of the symmetric matrix $\left[A_1\right]_{4\times4}$ are:

$$\begin{cases} a_{11}=(1-\upsilon)\left(R^3h+\dfrac{Rh^3}{4}\right)\ a_{22}=(1-\upsilon)\left(\dfrac{R^3h^3}{12}+\dfrac{3Rh^5}{80}\right) & a_{33}=\mu\left(R^3h+\dfrac{Rh^3}{4}\right) \\ a_{44}=\mu\left(\dfrac{R^3h^3}{12}+\dfrac{3Rh^5}{80}\right) & a_{12}=a_{21}=(1-\upsilon)\left(\dfrac{R^2h^3}{4}+\dfrac{h^5}{80}\right)\ a_{34}=a_{43}=\mu\left(\dfrac{R^2h^3}{4}+\dfrac{h^5}{80}\right) \end{cases} \tag{43a}$$

and nonzero components of the antisymmetric matrix $\left[A_2\right]_{4\times4}$ are:

$$\begin{cases} b_{13}=-b_{31}=\upsilon\left(R^2h+\dfrac{h^3}{12}\right)\ b_{14}=-b_{41}=\upsilon\left(R^3h+\dfrac{5Rh^3}{12}\right)\ b_{23}=-b_{32}=-\mu \\ \times\left(R^3h+\dfrac{Rh^3}{4}\right)+\upsilon\dfrac{Rh^3}{6}\quad b_{24}=-b_{42}=-\mu\left(\dfrac{R^2h^3}{4}+\dfrac{h^5}{80}\right)+\upsilon\left(\dfrac{R^2h^3}{3}+\dfrac{h^5}{40}\right) \end{cases} \tag{43b}$$

and nonzero components of the symmetric matrix $\left[A_3\right]_{4\times4}$ are:

$$\begin{cases} c_{22}=-\mu\left(R^3h+\dfrac{Rh^3}{4}\right) & c_{33}=-(1-\upsilon)Rh \\ c_{44}=-(1-\upsilon)R^3h-\dfrac{Rh^3}{6} & c_{34}=c_{43}=-\left(\upsilon R^2h+\dfrac{h^3}{12}\right) \end{cases} \tag{43c}$$

and the force vector $\{F\}_{4\times1}$ is:

$$\{F\}=\frac{r_i^3}{\lambda E_i}\left\{0\quad 0\quad -P_i+kP_o\quad h/2\left(P_i+kP_o\right)\right\}^T \tag{43d}$$

Finally, the radial displacement on the basis of Eq. (25) is obtained as follows:

$$u_r = \frac{r_i^4}{\lambda E_i h\left[\frac{h^4}{144}-(1-2\upsilon)\left(R^2+\frac{h^2}{6}\right)R^2\right]}\left\{\left[\frac{Rh}{2}\left(P_i+kP_o\right)+\frac{h^2}{12}\left(P_i-kP_o\right)+\upsilon R r_i\left(P_i-k^2P_o\right)\right]\bar{r}\right.$$
$$\left.-\left[\frac{R}{r_i}\left(R^2+\frac{h^2}{4}\right)\left(P_i-kP_o\right)+\frac{h}{2r_i}\left(R^2+\frac{h^2}{12}\right)\left(P_i+kP_o\right)\right]\right\} \quad (44)$$

6 Numerical Analysis

The Finite Element Method (FEM) is a powerful numerical method in shell analysis. An axisymmetric thick cylindrical shell is studied in the field of the plane elasticity. In this field, it suffices to model only the shell section. An axisymmetric element has been applied for modeling and meshing. The degrees of freedom are two translations in the radial and axial direction for each node.

For the modeling of the FGM hollow cylinders, an innovative application for the multilayering of wall thickness in the radial direction has been performed. In this approach, N homogenous layers which are of identical thickness and step-variable elasticity modulus has been formed. The elasticity modulus of each layer is then calculated by the following relation:

$$E = E_i\left[1+\sum_{j=1}^{N}(j-1)\frac{k^n-1}{N}\right] \quad , \quad k=\frac{r_o}{r_i} \quad (45)$$

where N is the number of layers, n is the inhomogeneity constant and j is the number allocated to each layer. In our study, 20 layers have been used for modeling exercise.

The nodes are free in all the elements. However, in the boundaries of $x=0$ and $x=L$, to create plane strain conditions, nodes are free along the radius and the circumference, but are constrained along the length.

7 Discussions

As a case study, we consider a thick cylinder whose elasticity modulus varies in radial direction and has the following characteristics: $r_i = 40$ mm, $r_o = 60$ mm, Young's modulus of inner surface $E_i = 200$ GPa and Poisson's ratio $\upsilon = 0.3$. Fig. 5 shows the distribution of elasticity modulus with respect to the normalized radius in a heterogeneous cylinder for integer values of n.

7.1 Internal Pressure

In this section, consider a nonhomogeneous thick cylinder in which the inner surface is compressed by uniform pressure $P_i = P = 80$ MPa and the outer surface is traction free. The distribution of the normalized radial displacement of this cylinder is depicted in Fig. 6. It is seen that for negative values of n, the displacements of FGM cylinders are higher than of a homogeneous cylinder. For positive values of n, the situation is reverse, i.e. the displacement is lower. The variation in the displacement of heterogeneous material is similar to that of homogenous material.

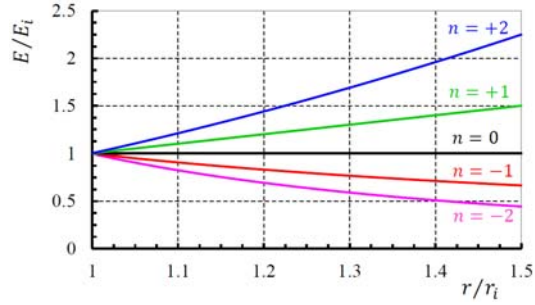


Figure 5
Distribution of elasticity modulus in FGM cylinder

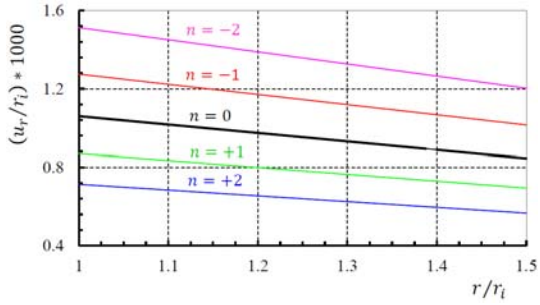


Figure 6
Distribution of radial displacement in FGM cylinder ($P_i = 80$ MPa)

Fig. 7 illustrates the distribution of normalized circumferential stress in a FGM cylinder. It should be pointed out that the equivalent graphs in Ref. (Tutuncu, 2001) are incorrect. For $n < 0$, in the inner half of the cylinder, the amount of circumferential stress is higher than that of the homogeneous cylinder. In contrast, in the outer half, it is lower. For $n > 0$, the situation is reverse. In the inner half of the cylinder, the amount of circumferential stress is lower than that of the homogeneous cylinder. As opposed to this, in the outer half, it is higher. In the

domain of the middle surface, the behavior of a FGM cylinder is similar to homogenous cylinder. The numerical results of this study are presented in Tables 1 and 2.

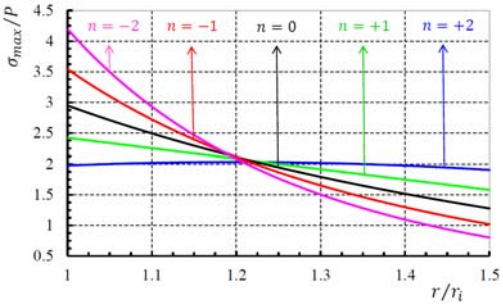


Figure 7
Distribution of circumferential stress in FGM cylinder ($P_i = 80$ MPa)

Table 1
Numerical results of radial displacement (P_i =80 MPa)

Surface	u_r , mm	*n*= -2	*n*= -1	*n*= 0	*n*= +1	*n*= +2
Middle surface	**FSDT**	0.054309	0.045790	0.038096	0.031266	0.025312
	PET	0.054541	0.045989	0.038272	0.031426	0.025458
	FEM	0.054559	0.045997	0.038272	0.031426	0.025458
Inner surface	**FSDT**	0.060512	0.050984	0.042388	0.034764	0.028124
	PET	0.062163	0.052673	0.044096	0.036471	0.029811
	FEM	0.062182	0.052680	0.044096	0.036468	0.029806

Table 2
Numerical results of maximum stress (P_i =80 MPa)

Surface	σ_θ, MPa	*n*= -2	*n*= -1	*n*= 0	*n*= +1	*n*= +2
Middle surface	**FSDT**	141.35	149.30	155.62	160.00	162.36
	PET	148.99	151.08	156.16	159.31	159.81
	FEM	144.29	151.15	156.16	159.12	159.80
Inner surface	**FSDT**	335.72	283.23	235.78	193.63	156.85
	PET	307.27	255.13	208	166.11	129.51
	FEM	299.91	252.09	208	168.18	133.60

7.2 External Pressure

In this section, consider a nonhomogeneous thick cylinder in which the outer surface is compressed by uniform pressure $P_o = P = 80$ MPa and the inner surface is traction free.

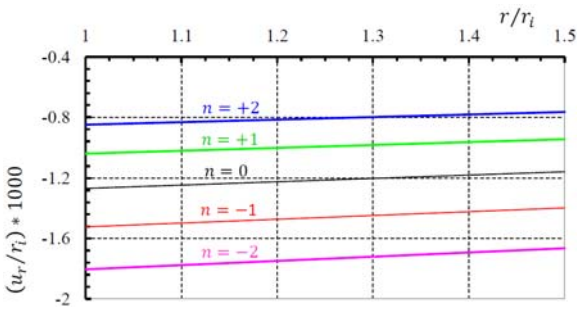


Figure 8
Distribution of radial displacement in FGM cylinder ($P_o = 80$ MPa)

The distribution of the normalized radial displacement of this cylinder is depicted in Fig. 8. It is seen that for negative values of n, the displacements of FGM cylinders are higher than of a homogeneous cylinder. For positive values of n, the situation is reverse, i.e. the displacement is lower. The variation in the displacement of homogenous material is similar to that of heterogeneous material.

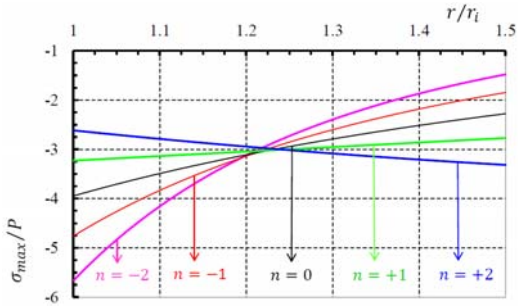


Figure 9
Distribution of circumferential stress in FGM cylinder ($P_o = 80$ MPa)

Fig. 9 illustrates the distribution of normalized circumferential stress in a FGM cylinder. For $n < 0$, in the inner half of the cylinder, the amount of circumferential stress is higher than that of the homogeneous cylinder. In contrast, in the outer half, it is lower. For $n > 0$, the situation is reverse. In the inner half of the cylinder, the amount of circumferential stress is lower than that of the homogeneous cylinder. As opposed to this, in the outer half, it is higher. In the domain of the middle surface, the behavior of a FGM cylinder is similar to homogenous cylinder. The numerical results of this study are presented in Tables 3 and 4.

Table 3
Numerical results of radial displacement (P_o =80 MPa)

Surface	u_r , mm	n= -2	n= -1	n= 0	n= +1	n= +2
Middle surface	**FSDT**	-0.069367	-0.058383	-0.048496	-0.039745	-0.032136
	PET	-0.069714	-0.058630	-0.048672	-0.039872	-0.032228
	FEM	-0.069737	-0.058640	-0.048675	-0.039871	-0.032227
Inner surface	**FSDT**	-0.072159	-0.060894	-0.050707	-0.041653	-0.033749
	PET	-0.074801	-0.063030	-0.052416	-0.043005	-0.034809
	FEM	-0.074821	-0.063037	-0.052416	-0.043002	-0.034806

Table 4
Numerical results of maximum stress (P_o =80 MPa)

Surface	σ_θ, MPa	n= -2	n= -1	n= 0	n= +1	n= +2
Middle surface	**FSDT**	-218.44	-228.32	-235.62	-240	-241.29
	PET	-221.87	-230.20	-236.16	-239.43	-239.81
	FEM	-222.03	-230.28	-236.16	-239.42	-239.80
Inner surface	**FSDT**	-453.47	-380.89	-315.79	-258.34	-208.54
	PET	-411.0	-346.32	-288	-236.29	-191.26
	FEM	-401.0	-342.07	-288	-239.22	-196.04

Conclusions

In this research, the heterogeneous hollow cylinders have been solved by FSDT, PET and FEM, and have been compared with homogenous cylinders. We conclude that for the positive or negative values of n, the maximum stress increases in one half of the cylinder, and it is decreased in the other half. The radial displacement for positive values of n is decreases; it is however increased for negative values. As $|n|$ is increases, the amount of changes in both displacements and stresses increases too. Therefore, positive values of n lead to a decrease in the displacement and stress in the inner surface. This is highly important for a large number of industries.

Acknowledgement

The authors wish to thank Dr. Kazemi for his kind cooperation.

References

[1] Koizumi, M.: The Concept of FGM, Ceramic Transactions Functionally Graded Material, Vol. 34, pp. 3-10, 1993

[2] Naghdi, P. M. and Cooper, R. M.: Propagation of Elastic Waves in Cylindrical Shells Including the Effects of Transverse Shear and Rotary Inertia, Journal of the Acoustical Society of America, Vol. 28, No. 1, pp. 56-63, 1956

[3] Mirsky, I. and Hermann, G.: Axially Motions of Thick Cylindrical Shells, Journal of Applied Mechanics, Vol. 25, pp. 97-102, 1958

[4] Mindlin, R. D.: Influence of Rotary Inertia and Shear on Flexural Motions of Isotropic Elastic Plates, Journal of Applied Mechanics, Vol. 18, pp. 31-38, 1951

[5] Greenspon, J. E.: Vibration of a Thick-walled Cylindrical Shell, Camparison of the Exact Theory with Approximate Theories, Journal of the Acoustical Society of America, Vol. 32, No. 5, pp. 571-578, 1960

[6] Ziv, M. and Perl, M.: Impulsive Deformation of Mirsky-Hermann's Thick Cylindrical Shells by a Numerical Method, Journal of Applied Mechanics, Vol. 40, No. 4, pp. 1009-1016, 1973

[7] Suzuki, K., Konnon, M. and Takahashi, S.: Axisymmetric Vibration of a Cylindrical Shell with Variable Thickness, JSME, Vol. 24, No. 198, pp. 2122-2132, 1981

[8] Simkins, T. E.: Amplifications of Flexural Waves in Gun Tubes, Journal of Sound and Vibration, Vol. 172, No. 2, pp. 145-154, 1994

[9] Eipakchi, H. R., Rahimi, G. H. and Khadem, S. E.: Closed Form Solution for Displacements of Thick Cylinders with Varying Thickness Subjected to Non-Uniform Internal Pressure, Structural Engineering and Mechanics, Vol. 16, No. 6, pp. 731-748, 2003

[10] Ghannad, M. and Nejad, M. Z.: Elastic Analysis of Pressurized Thick Hollow Cylindrical Shells with Clamped-Clamped Ends, Mechanika, Vol. 85, pp. 11-18, 2010

[11] Fukui, Y. and Yamanaka, N.: Elastic Analysis for Thick-walled Tubes of Functionally Graded Materials Subjected to Internal Pressure, JSME Ser. I, Vol. 35, No. 4, pp. 891-900, 1992

[12] Horgan, C. O. and Chan, A. M.: The Pressurized Hollow Cylinder or Disk Problem for Functionally Graded Isotropic Linearly Elastic Materials, Journal of Elasticity, Vol. 55, No. 1, pp. 43-59, 1999

[13] Tutuncu, N. and Ozturk, M.: Exact Solutions for Stresses in Functionally Graded Pressure Vessels, Composites Part B-Engineering, Vol. 32, No. 8, pp. 683-686, 2001

[14] Hongjun, X., Zhifei, S. and Taotao, Z.: Elastic Analyses of Heterogeneous Hollow Cylinders, Mechanics Research Communications, Vol. 33, No. 5, pp. 681-691, 2006

[15] Zhifei, S., Taotao, Z. and Hongjun, X.: Exact Solutions of Heterogeneous Elastic Hollow Cylinders, Composite Structures, Vol. 79, No. 1, pp. 140-147, 2007

[16] Tutuncu, N.: Stresses in Thick-walled FGM Cylinders with Exponentially-Varying Properties, Engineering Structures, Vol. 29, No. 9, pp. 2032-2035, 2007

[17] Nejad, M. Z., Rahimi, G. H. and Ghannad, M.: Set of Field Equations for Thick Shell of Revolution Made of Functionally Graded Materials in Curvilinear Coordinate System, Mechanika, Vol. 77, No. 3, pp. 18-26, 2009

[18] Ghannad, M., Rahimi, G. H. and Khadem, S. E.: General Plane Elasticity Solution of Axisymmetric Functionally Graded Thick Cylindrical Shells, Journal of Modares Technology and Engineering, Vol. 10, pp. 31-43, 2010

[19] Abedi, M., Nejad, M. Z., Lotfian, M. H. and Ghannad, M.: Static Analysis of Parabolic FGM Solid Cylinders, Journal of Basic and Applied Scientific Research, Vol. 1, No. 11, pp. 2339-2345, 2011

[20] Wylie, C. R.: Differential Equations, McGraw-Hill, New York, 1979

[21] Vlachoutsis, S.: Shear Correction Factors for Plates and Shells, International Journal for Numerical Methods in Engineering, Vol. 33, No. 7, pp. 1537-1552, 1992

Advanced Character Collage CAPTCHA

Goran Martinovic, Zdravko Krpic

Faculty of Electrical Engineering, Josip Juraj Strossmayer University of Osijek,
Cara Hadrijana bb, 31000 Osijek, Croatia
goran.martinovic@etfos.hr; zdravko.krpic@etfos.hr

Abstract: Text based CAPTCHA systems are widely used as a security mechanism for web access control. Considering their broad use, many attacks are challenging them every day. Most of the attacks aimed at CAPTCHAs are based on the latest computer vision techniques, AI methods and OCRs, so it is imperative to enhance these methods even more. There are number of proposals for CAPTCHA security, but it is hard to achieve a good balance between CAPTCHA practicality and its security. Advanced Character Collage CAPTCHA is a highly random novel method which uses the strengths of unbroken CAPTCHAs along with the weaknesses of present ones, and relies on imperfection of computer vision techniques. The proposed CAPTCHA is generated through a series of unique creation steps, each of them implementing carefully analyzed features in order to increase human recognition rate, and at the same time, to reduce computer recognition rate. The degree of recognition within the proposed method is evaluated using several tests, while its readability by humans is tested through two surveys.

Keywords: CAPTCHA; recognition rate; security; web access control

1 Introduction

Security is a major concern on web exposed systems holding valuable data or something that can be compromised. There are many types of attacks that can be carried out on these systems. A variety of bots, spiders, DOS attacks, domain hijacking, cache poisoning, worms and spam pose a serious threat to online systems and can cause major losses. Therefore, it is imperative that these systems have the most reliable security systems. Besides encryption, secure connections and protocols, there is one portion of authorization system where the computer has to decide: "Human user or computer bot?" If the user is human, an access is granted, possibly to very important data, money or goods. CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a test which distinguishes whether a user is human or computer bot. There are many different types of CAPTCHAs, most of them including a small image from which the user has to decipher the letters and type them in a small form box in order to identify himself as a human to be granted an admission to a certain part of

the web site. CAPTCHA security is mostly used in web sites which include an e-mail account creation, web rating systems, polls, search entries, forum posts, downloads and many other in order to prevent malicious users from spamming, distributing copyrighted and stolen material, inducing inflation or deflation of rankings and polls, and similar unwanted actions. The fact that CAPTCHAs presently protect many systems makes them a desirable target for everyday attacks by using various machine vision and AI techniques. There are number of specialized attacks which aim at CAPTCHA security, such as OCR and non-OCR based attacks, statistical based methods, structural analysis, wavelet fractal feature extraction, neural network "divide and conquer", and other various AI-based procedures. Even a new threat, called 3rd party attacks, has emerged, which uses cheap human labor to manually solve CAPTCHAs in order to create thousands of various accounts, polls, spam, etc. Furthermore, 3rd party attacks are used to create databases of CAPTCHA-solution pairs for finite-state CAPTCHAs (combinations of which can be exhausted in a feasible amount of time) as an input for brute-force attacks. CAPTCHA security systems include different visual and non-visual CAPTCHAs which are presented to the user, and then he has to identify or compare certain images, retype the presented distorted letters or words, or type the letters heard in a sound CAPTCHA. Even a combination of visual and non-visual CAPTCHA is possible, enabling use by people with disabilities, although CAPTCHAs based on sensory abilities cannot be used on sensory-impaired human beings, as stated in [1].

The rest of the paper is organized as follows: Section 2 enumerates some of the most important related works that have led to many ideas proposed in this paper. Section 3 covers the proposed method in detail, including generation and implementation of the proposed method. Furthermore, in Section 4, readability features of the proposed method are evaluated through a couple of surveys, while the security features are analyzed in Section 5. Section 6 shows plans for future research, upgrades and plans based on this paper, and Section 7 concludes the paper.

2 Related Work

Since this paper analyzes only image-based CAPTCHAs, which are the part of visual CAPTCHA systems, only a fraction of the vast related work from this area will be mentioned.

Authors in [2] use the term Collage CAPTCHA for a three-step process of authorization, in which the user must choose the correct image and the name of the object on the image, and only then he is granted an access to the third step, which is entering the image name into a text box. Collage CAPTCHA is considerably secure, but the major flaw of this system is its usability, considering the length of the process, and the fact that error chance by the user is multiplied.

Also, the CAPTCHA alone is relatively easy to solve. An interesting concept is proposed in [3], in which the discernment between people and bots is done by the means of recognizing strangeness in a machine translation. The differentiation of this method is excellent, but the limited number of sentences and language dependability, as well as exposure to 3rd party solvers, bound this method to limited use. A 3D CAPTCHA [4] is a promising technique, in which the authors propose different implementations of 3D letters to deceive bots, with 80% overall hit-rate by humans and a relatively complicated generation (DirectX). One of the hardest methods for bots to decipher is proposed in [5], which is based on animation with moving letters presented to the human. The method protection is very good but the advancing OCR techniques and slow and complicated implementation limit its practicality. In [6] the authors propose kernels to break different common types of CAPTCHA, accentuating major flaws, such as susceptibility to line removal algorithms, letter pattern matching, dot removal, binarization, etc., all of which are more or less absent from the method proposed in this work – Advanced Character Collage CAPTCHA. They managed to solve EZ-Gimpy with 88% success rate.

Other related papers include various novel approaches, such as [7, 8], and methods of breaking visual CAPTCHAs [9, 10].

3 Proposed Algorithms

Various authors have proposed different CAPTCHA classifications, but the most common one divides CAPTCHA systems in to visual and non-visual. Most of the non-visual CAPTCHAs are based on sound, making them less secure than visual ones due to the high-quality voice recognition and noise removal programs, as mentioned in [7, 11]. Visual CAPTCHAs, on the other hand, can be divided into OCR and non-OCR based ones, as proposed in [2, 4, 5, 12]. Non-OCR CAPTCHAs are mostly image-based, a concept which many authors, e.g. [2, 7, 12], consider to be the future of CAPTCHA protection, or at least an important part of it. The same authors claim that these CAPTCHAs do not cause dissatisfaction to its users, as most of the OCR-based ones do, but they are the most susceptible to 3rd party attacks because of databases with a limited number of images. Even Asirra[1], the most famous example of an image based CAPTCHA, with the largest image database, was broken by the authors in [13] using machine learning techniques. The authors in [1] made a good point when they said that, "There is no way to prove that a program cannot pass a test which a human can

[1] Asirra is an image based CAPTCHA which uses one of the largest lost pets database in the world (http://www.petfinder.com) to generate an image query for a human to solve. A human has to distinguish between cats and dogs. More at http://research.microsoft.com/en-us/um/redmond/projects/asirra/.

pass, since here is a program – the human brain – which passes the test". So, the goal is not to make a computer-unsolvable CAPTCHA (which is impossible), but to create a CAPTCHA system which is difficult to solve for a computer and easy for a human. It is enough to make novel CAPTCHAs better than other CAPTCHAs to divert attacks from the system which it protects. In our proposed work, the goal was to make a simple CAPTCHA that is easy to generate and easy to implement on a variety of platforms, and at the same time, that provides significant resistance to computer vision attacks. However, in order to maintain high recognition rate by humans, and at the same time, to deny computer bots deciphering CAPTCHAs, the best approach is to try to implement as many strengths of the existing strong CAPTCHAs and to avoid as many of their flaws as possible. Therefore, the best method to use is to learn from previous experience in CAPTCHA systems, as well as from machine vision and state-of-the-art AI.

The property of CAPTCHA which enables recognition by humans (RBH) is the distinction between characters, background and the clutter. Moreover, the same property is used by computer vision to decipher CAPTCHA, but in a different way. Human perception is associative, and therefore this fact should be more exploited. This knowledge gives an important but often overlooked postulate: the characters do not need to be entirely visible to facilitate RBH and at the same time deny recognition by a computer vision (RBC) due to the lack of the character integrity. In the following subsections, Advanced Character Collage CAPTCHA creation steps will be analyzed in detail, resulting in the complete CAPTCHA generation algorithm presented at the end of the section.

3.1 Strengths and Weaknesses

The major weaknesses of present OCR-based CAPTCHAs, as pinpointed in [14], can be: constant font, aligned glyphs, constant glyph position, no deformation, constant colors, no perturbation, constant background, non-textured background, weak color variation, etc. So the proposed CAPTCHA should avoid all these flaws as much as possible. The method proposed in [15] has undergone some major changes in order to fulfill security demands as much as possible, while retaining high RBH and easiness of implementation. Most aspects of the proposed method were analyzed and improved, and for every aspect there follows a description. The major change of the proposed CAPTCHA is the use of an edge detection filter, which facilitates two major improvements: resistance to assorted color segmentation attacks and usability by color blind people. Another characteristic which is omitted from our proposed method is that of using a finite set of CAPTCHA images, as they can be easily classified by the 3rd party attacks. A different improvement of the method from [15] is proposed in [16], retaining the color based CAPTCHA.

3.2 Background

The proposed CAPTCHAs were made on a 640x190 pixel white background canvas. Any color can be used for the background, but lighter colors increase RBH. The background is composed of basic geometric shapes (rectangles, circles and semicircles) in order to increase curve similarity with the characters which are going to be placed on the clutter. These shapes are painted with various semitransparent pale colors with reduced contrast, they are randomly sized, their placement is random, and they overlap. Semi transparency ensures better clutter, especially if edge detection is applied to it afterwards. Shape size and a color palette are limited by a certain threshold. Moreover, shapes can be rendered randomly, or a database of these shapes can be used, from which shapes are randomly chosen and copied at various locations on a canvas. We propose random generation of shapes, thus avoiding the need for their external storage. If an external storage is used, it can also be utilized for a CAPTCHA buffer, a concept which is described in subsection 5.4. The number of generated shapes is also bound to a certain threshold based on the canvas and shape size, because there should be enough shapes to saturate the background, but not too many, in order to avoid oversaturation and thus making characters more distinguishable to computer bots, and less visible to humans.

3.3 Character Composition

Allowed characters are random clear type font letters (uppercase and lowercase) and digits 0-9. Bold or very thick fonts should be avoided because they cause readability issues. After the background has been generated, it is split into r vertical regions, $R_{i,vert}$, where r is the number of characters in one CAPTCHA, $4 \leq r \leq 6$. In our experiment, font face and size were constant. An outlined character-shaped mask is placed on each region $R_{i,vert}$, and the masked region of a character is then copied to region $R_{i,vert}$ of another character, and vice versa. That way, characters are composed of the same texture as the background clutter. Furthermore, the character-shaped mask is meshed into regions based on the previously designed texture beforehand, with the intention of avoiding pixel continuity. The mesh texture lines should be sufficiently thick to separate characters into pieces, but not so thick as to reduce RBH. Our research has shown that optimal meshed texture line is approximately 6 pixels thick, and the example of the used mesh texture can be seen in Fig. 1.

Optionally, before placing meshed character-shaped masks on the background, r regions with greatest color difference $R_{i,color}$ could be found on the background. These regions can serve as placeholders for character masks before copying, increasing readability on both the color and grayscale versions of the Advanced Character Collage CAPTCHA.

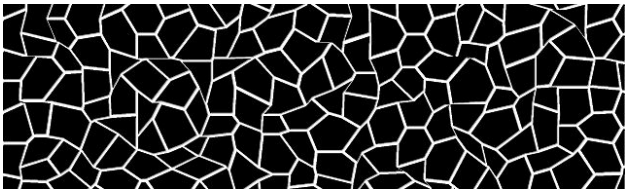


Figure 1
Mesh texture used in the experiment

3.4 Character Placement

A common method to avoid one of the major weaknesses of the existing CAPTCHAs, which is constant glyph alignment and rotation, is to apply mild warp to a glyph-shaped mask, along with slight random rotation (up to 30º in an arbitrary direction), which is different for every glyph, although some authors propose up to 45º [4]. In our work, rotation and warp were neglected, because the background was composed only of straight lines and circles; warped lines would be too prominent after line removal preprocessing methods and would compromise security of the CAPTCHA. However, glyph deformation analysis will be a part of our future work. Letter placement in a region is random, considering that the whole glyph is visible, i.e. is not outside the canvas, by the means of using random offset values from the centre of the region.

3.5 Edge Detection

The major change to the proposed method based on the work done in [15] was to apply an edge detection filter to the image, which also converts image to grayscale. The main reason for such a change was insufficient resistance to color segmentation attacks, which could easily separate glyphs from the background, as seen in Fig. 2. Fig. 2a shows the first implementation of the Advanced Character Collage CATPCHA; there is an exclusion operation applied between glyph layer and the background layer, resulting in a high RBH. However, if the color channel mixing is applied, the glyphs can easily be isolated by computer vision, as shown in Fig. 2b. Another major weakness is brightness and contrast tuning, the implementation of which can lead to an even better glyph segmentation by computer, Fig. 2c.

Edge detection, in addition to providing greater resistance to attacks, allows color blind people to solve the CAPTCHA, thus spreading the pool of potential users.

The illustration of the Advanced Character Collage CAPTCHA creation steps can be seen in Fig. 3. Fig. 3a shows a white background canvas that is saturated with random shapes, shown in Fig. 3b.

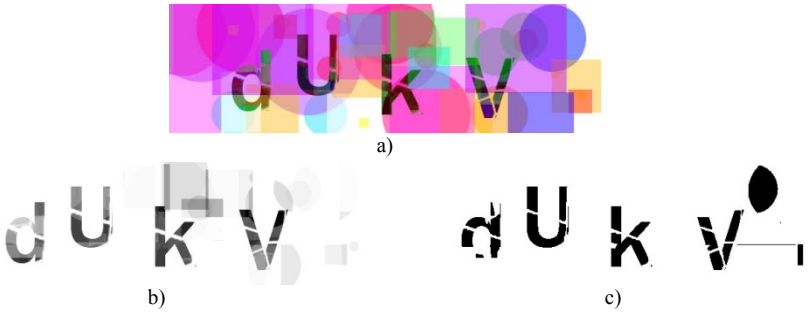


Figure 2
Character Collage CAPTCHA from [15]: a) original image, b) after using color channel mixer, c) after contrast and brightness tuning

The glyph shaped regions are placed onto saturated background and copied, Fig. 3c, put back to the corresponding background regions $R_{i,vert}$, Fig. 3d, and finally by applying the edge detection result in the CAPTCHA image, shown in Fig. 3e.

With everything taken into consideration, a complete Advanced Character Collage CAPTCHA algorithm can be proposed, that is illustrated in Fig. 4.

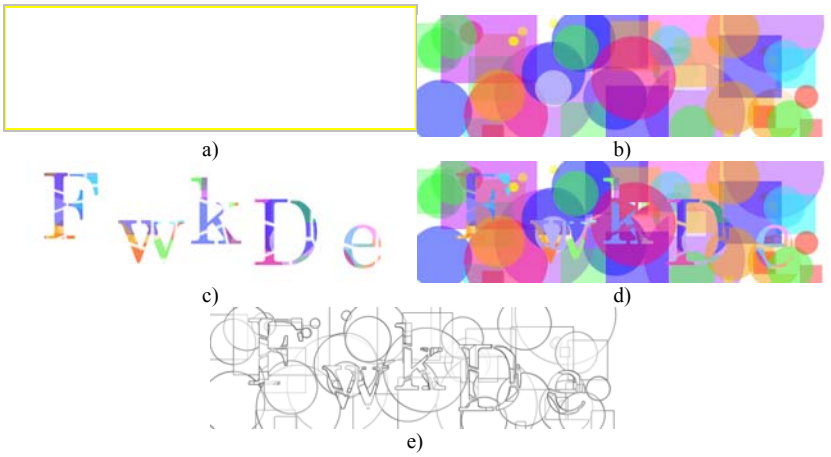


Figure 3
Advanced Character Collage CAPTCHA creation steps: a) white canvas, b) background clutter, c) meshed glyph shaped regions, d) glyph shaped regions put onto the background clutter, e) resulting CAPTCHA

1. *Open canvas.*
2. ***For*** *i=0 to* $\boldsymbol{m}$, $\boldsymbol{m} \in [min\_saturation, max\_saturation]$
   - 2.1. *Generate two numbers* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *such that* $\boldsymbol{x} \in [0, image\_height]$, $\boldsymbol{y} \in [0, image\_width]$.
   - 2.2. *Create randomly sized shape and paint it with randomly generated color.*
   - 2.3. *Place the centre of the shape at coordinates* $(\boldsymbol{x}, \boldsymbol{y})$.
3. *Generate random number* $\boldsymbol{r}$ *such that,* $4 \leq \boldsymbol{r} \leq 6$.
4. *Split the canvas into* $\boldsymbol{r}$ *vertical regions* $\boldsymbol{R_{i,vert}}$.
5. *Calculate the central coordinates* $(\boldsymbol{x_i}, \boldsymbol{y_i})$ *of every region* $\boldsymbol{R_{i,vert}}$.
6. ***For*** *i=0 to* $\boldsymbol{r}$
   - 6.1. *Choose random character* $\boldsymbol{c_i}$.
   - 6.2. *Create character shaped mask.*
   - 6.3. *Load meshed texture from a file.*
   - 6.4. *Apply random slight warp to the loaded texture.*
   - 6.5. *Permeate the character shaped mask with the resulting mesh texture.*
   - 6.6. *Create offset* $\boldsymbol{p}$ *and* $\boldsymbol{q}$.
   - 6.7. *Place the centre of the mask into coordinates* $(\boldsymbol{x_i} + \boldsymbol{p}, \boldsymbol{y_i} + \boldsymbol{q})$.
   - 6.8. *Copy masked area.*
   - 6.9. *Select random region* $\boldsymbol{R_{j,vert}}$, $j \neq i$.
   - 6.10. ***If*** *selected regions flag* $\boldsymbol{s_j} \neq 0$
     - 6.10.1. *Repeat selection*
   - 6.11. ***Else***
     - 6.11.1. *Paste copied masked area to selected region, with its centre aligned to the offset* $(\boldsymbol{x_j} + \boldsymbol{p}, \boldsymbol{y_j} + \boldsymbol{q})$.
     - 6.11.2. *Set selected regions' flag* $\boldsymbol{s_j}$ *to 1.*
     - 6.11.3. *Put character* $\boldsymbol{c_i}$ *into CAPTCHA string at* j^{th} *position.*
7. *Apply edge detection filter on the whole image.*
8. *Save as image.*
9. *Save CAPTCHA string as an image pair for comparison.*

Figure 4
Advanced Character Collage CAPTCHA generation algorithm

4 Readability Survey

RBH is the most important feature of the CAPTCHA system. Consequently, humans have to test the CAPTCHA to facilitate its readability properties. Two surveys have been conducted to eliminate features which reduce human recognition rate. Both surveys were attended by random groups of people in a way that they have been given CAPTCHA tests to solve them.

4.1 First Survey

In the first survey, participants were not exclusively informed that the CAPTCHA was case-sensitive. A survey consisted of 24 CAPTCHAs, which all included letters a-z and A-Z, and were of a length between 4 and 5 characters ($4 \leq r \leq 5$). Digits were not used. A sample of a given CAPTCHA test is given in Fig. 5.

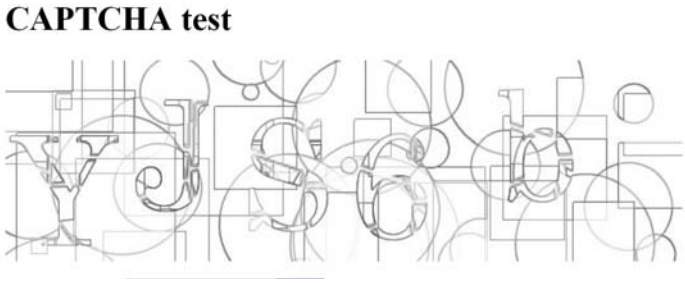


Figure 5
Example of a survey test

110 anonymous random people did the test, so the demographic data is not available/not known. The results of the survey are presented in the Table 1. The overall CAPTCHA hit-rate was 62.7%, which is a good result considering the lack of fine tuning. 65.5% of all errors were caused by inability to determine the glyph case, while the remainder relates to false glyph recognition. There were no obscure glyphs.

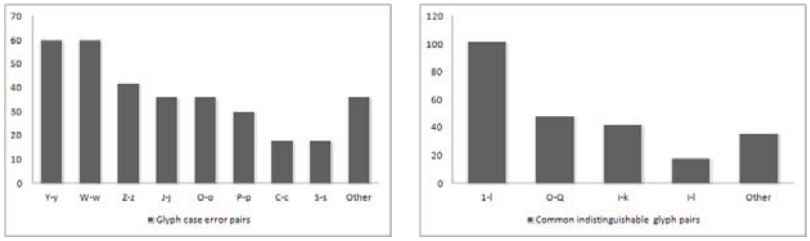
Table 1
Results of a first survey

Overall performance	**Hit rate**	[%]	62.7
	Average solving time	[s]	9.1
No. of unidentified glyphs per CAPTCHA	**1-glyph miss**	[%]	59.8
	2-glyph miss	[%]	24.4
	3-glyph miss	[%]	11.6
	4-glyph miss	[%]	4.3
Glyph error types	**Glyph case miss**	[%]	65.5
	Glyph miss	[%]	34.5
	Obscure glyphs	[%]	0.0

Most of the incorrectly recognized CAPTCHA tests had a single glyph miss (59.8%), and 24.4% of all faulty tests had two-glyph miss. Moreover, multiple glyph misses were mainly caused by the fact that users did not know that tests were case-sensitive, therefore resulting in a 11.6% share of three-glyph misses, and even a 4.3% share of four-glyph misses in the total human recognition error. It is estimated that approximately 13% of all solved CAPTCHAs were falsely recognized because of the above reason.

Other multi-glyph misses were caused by problematic glyph pairs, which can be divided into two groups. The first group consisted of glyphs whose uppercase and lowercase versions are difficult to discern. These glyph pairs are shown in Fig. 6a, with their respective shares in the overall human recognition error. The second group of glyph pairs consisted of similar letters, and their share in the total error is shown in Fig. 6b. The first survey has shown that the hit rate can be improved by

avoiding the mentioned glyphs pairs, which would then improve RBH, but it would also source a smaller CAPTCHA combination space. The negative effect of these improvements can be avoided by increasing the glyph pool, which will be considered in survey 2. Another way to improve RBH is to use fonts which offer greater dissimilarity between letters in the problematic pairs, such as console fonts or old style fonts. Additionally, the hit rate can be further improved by using letter mask region select technique proposed in subsection 3.3, based on $R_{i,\ color}$ regions.



a) b)

Figure 6
Error intensity for the problematic glyph pairs

Analysis of the average time for a human to solve the proposed CAPTCHA shows that it is not time consuming, with an average human solving time of only 9.1 seconds.

4.2 Second Survey

In the second survey, 104 participants were informed that the given CAPTCHA is case-sensitive. In addition to that, the elimination of some problematic glyph pairs was done; for example, uppercase glyphs “I” and “O” and lowercase glyphs “l” and “q” were removed from the glyph pool. A different font (Century) was chosen to accommodate greater difference between glyphs in other problematic glyph pairs. Additionally, to increase the glyph pool, and therefore to reduce the risk of brute force attacks, digits 1-9 were added to the glyph pool. “0” was left out intentionally because of the similarity with the letter “O”. The CAPTCHA length was increased from $4 \leq r \leq 5$ to $4 \leq r \leq 6$. Table 2 shows the performance and error analysis of the second survey. The overall hit rate was improved to 89.9%, which is a very good result. The average solving time was increased by 0.4s, but if the increased number of glyphs is taken into consideration, this increase is negligible. It is worth mentioning that the inability to determine whether the letter is uppercase or lowercase is still a major cause of CAPTCHA recognition errors (40.9%), but this time other letter recognition errors are almost equally present (37.7%). There were a number of situations where one glyph was significantly less visible than the other, which caused 14.3% of all errors. Finally, there were several occasions (with 7.1% share of errors) where participants mistyped the number (0 and 9 most of the time) because of their keyboard location.

Table 2
Results of a second survey

Overall performance	**Hit rate**	[%]	89.9
	Average solving time	[s]	9.5
No. of unidentified glyphs per CAPTCHA	**1-glyph miss**	[%]	96.4
	2-glyph miss	[%]	3.6
	3-glyph miss	[%]	0.4
	4-glyph miss	[%]	0.0
Glyph error types	**Glyph case miss**	[%]	40.9
	Glyph miss	[%]	37.7
	Obscure glyphs	[%]	14.3
	Number glyph mistype	[%]	7.1

5 Security and Performance

Since there are no available tools for testing the CAPTCHA resistance to AI attacks, deciphering steps are analyzed from related and previous work. That way, the most common attack routines can be isolated and simulated in order to apply them to the proposed method. Unfortunately, the best CAPTCHA benchmark is real-world use, i.e., when it draws enough attention, as mentioned in Section 3. Most of the researchers [4, 6, 11] agree that the procedure for deciphering CAPTCHA can be divided into three main steps: Preprocessing, Segmentation and Classification, although some authors [10] include an additional step before the last – Feature extraction. The preprocessing part of the process converts the CAPTCHA to grayscale and removes any noise and background. After the image passes through the preprocessing step, segmentation is applied, which separates regions on the image which (should) contain glyphs. The optional next task is to extract unique features of the characters (number of holes, height of character, etc.) to further enhance the last step, which is character recognition. Finally, OCRs are applied for character recognition.

The CAPTCHA proposed in [15] was susceptible to certain image manipulations which did manage to successfully isolate glyphs in several cases. For example, these were mixing color channels, contour tracing algorithms, brightness and contrast adjustments and custom edge detection tools, which served as the guidelines for enhancements to the Advanced Character Collage CAPTCHA. These tests can be seen in [16]. In the next subsections resistance to RBC in every CAPTCHA deciphering step will be analyzed.

5.1 Preprocessing

Background removal tools are mainly based on distinction between characters and the clutter, such as line, color, discontinuity, dot and mesh removal, and color segmentation. Line removal is not feasible because the characters are mainly composed of the same lines as the clutter, so by removing them, glyph information is also removed. Moreover, dot removal fails for the same reason line removal tool does. Color segmentation, as mentioned in subsection 3.5, does not apply to the proposed CAPTCHA, and the discontinuity removal is dependent on line removal. Also, converting CAPTCHA to binary colors removes a lot of information from the glyphs, making them unusable for recognition. In addition, textured mesh is random and not regular, so applying universal mesh removal tools also does not lead to deciphering improvement. The texture mesh comprises hexagons with randomly sized edges to eliminate the possibility of an attacking party duplicating it and filling in the missing glyph pieces. Moreover, the texture mesh can be generated each time a new CAPTCHA is generated.

5.2 Segmentation

Segmentation is the most important step in deciphering CAPTCHA, because this is the step in which the human outperforms the machine. Therefore, segmentation should be as hard to perform as possible. In the proposed method, if the attacker manages to separate glyph regions from the clutter, the resulting image does not have enough information about the glyphs to successfully implement common OCR. If the glyphs are extracted, however, they still pose a challenge for an OCR or the pixel count methods because of the meshed nature. This challenge would be easier for an attacker if the mesh would have been a simple continuous mesh, but in the proposed CAPTCHA this is not the case, as noted in the previous subsection.

5.3 3rd Party Attacks

The proposed novel CAPTCHA is highly resistant to 3rd party attacks, by the means of exploiting finite CAPTCHA states to create a database that the malicious user can exploit to spam, or to create thousands of fake accounts, polls and so on. The proposed CAPTCHA is highly random, from the background to glyph generation and placement, which also discourages the use of machine learning techniques, neural networks and similar AI attacks. Using letters instead of complete words also helped in improving the resistance from 3rd party attacks because of significant increase in CAPTCHA combination space. Moreover, this feature also helped to eliminate the possibility of dictionary attacks. However, if an attack method is used such as the authors in [8] tried to overcome, the proposed CAPTCHA could not resist. Nevertheless, an additional improvement of the proposed CAPTCHA can be done through interaction similar to [8], or just by using allowed time windows for solving CAPTCHA.

5.4 CAPTCHA Buffer

With the aim of using the proposed CAPTCHA in high traffic web applications, another novel idea is proposed. When issuing concerns about system slowdowns because of a lot of CAPTCHA generation demands, although the generation algorithm is not computationally intensive, a CAPTCHA buffer can be used. A CAPTCHA buffer is a storage which contains a predefined number of pre-generated CAPTCHAs. The concept is based on the idea of generating a certain number of CAPTCHAs in advance, when the system is lightly used, not only by demand, and summoning them as necessary. In this way, the sporadic system slowdowns can be avoided in peak usage periods. This mechanism can be used in conjunction with the clutter shape storage, described in subsection 3.2.

6 Future Work

The future tasks for the proposed method development are simplification of generation and readability improvement. Moreover, some fine tuning characteristics of the proposed method should be evaluated, such as the influence on RBC and RBH of glyph rotation, font face, character shaped mask warping, CAPTCHA size, etc. The authors plan to differ the number of characters more intensively, such as $4 \leq r \leq 7$. Also, another survey is planned that will encompass more people and with many different CAPTCHAs in order to gain better perspective on the practicality of the future proposed CAPTCHA.

Conclusions

CAPTCHA protection is considered by some authors as a weak protection against bots and spammers, but its ubiquitous use and the many researches still pending on it tell the different story. Its main characteristics are simple implementation, high practicality, good acceptance by people and fair security. In order to avoid the necessity of complicated security systems for simple tasks, such as mail registrations, auctions, polls, ballots, forum posts, etc. CAPTCHA is seen as the best balance of all the needed characteristics for these tasks. However, as CAPTCHA security advances, so too do the attack methods, such as various OCR and non-OCR based attacks, 3rd party solving techniques, AI approaches and many others. All these considerations need to be taken into account when creating new CAPTCHA security.

The proposed novel method, Advanced Character Collage CAPTCHA, proved itself superior to the most common types of attacks, along with its simple implementation and good readability. Its basic strengths lay in the strengths of unbroken CAPTCHAs, as well as in the flaws of the machine vision technologies, and AI imperfections. Despite the fact that the proposed CAPTCHA will be probably rendered unusable over time as technology advances, it is still a full of

challenges for an attacker and carries some novel ideas for possible future CAPTCHA implementations.

Acknowledgement

This work was supported by research project grant No. 165-0362980-2002 from the Ministry of Science, Education and Sports of the Republic of Croatia.

References

[1] L. von Ahn, M. Blum, N. J. Hopper, J. Langford: CAPTCHA: Using Hard AI Problems for Security, Proc. of 22nd Int. Conf. on Theory and Applications of Cryptographic Techniques, Warsaw, Poland, 4-8 May 2003, pp. 294-311

[2] R. Soni, D. Tiwari: Improved CAPTCHA Method, International Journal of Computer Applications, Vol. 1, No. 25, pp. 92-94, 2010

[3] T. Yamamoto, J. D. Tygar, M. Nishigaki: CAPTCHA Using Strangeness in Machine Translation, Proc. of 24th IEEE Int. Conf. on Advanced Information Networking and Applications, Hamamatsu, Japan, 20-23 April 2010, pp. 430-437

[4] M. Imsamai, S. Philmoltares: 3D CAPTCHA: A Next Generation of the CAPTCHA, Proc. of 2010 Int. Conf. on Information Science and Applications, Seoul, South Korea, 21-23 April 2010, pp. 1-8

[5] J. Cui, L. Wang, J. Mei, D. Zhang, X. Wang, Y. Peng, W. Zhang: CAPTCHA Design Based on Moving Object Recognition Problem, Proc. of 3rd Int. Conf. on Information Sciences and Interaction Sciences, Wuhan, China, 23-25 June 2010, pp. 158-162

[6] A. A. Chandavale, A. M. Sapkal: Algorithm for Secured Online Authentication Using CAPTCHA, Proc. of 3rd Int. Conf. on Emerging Trends in Engineering and Technology, Goa, India, 19-21 November 2010, pp. 292-297

[7] H. Gao, D. Yao, H. Liu, X. Liu, L. Wang: A Novel Image Based CAPTCHA Using Jigsaw Puzzle, Proc. of 13th IEEE Int. Conf. on Computational Science and Engineering, Xi'an, China, 11-13 December 2010, pp. 351-356

[8] H. D. Truong, C. F. Turner, C. C. Zou: iCAPTCHA: The Next Generation of CAPTCHA Designed to Defend Against 3rd Party Human Attacks, Proc. of 23rd IEEE Int. Conf. on Communications, Kyoto, Japan, 5-9 June 2011, pp. 1-6

[9] J. Yan, A. S. E. Ahmad: Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms, Proc. of 23rd IEEE Computer Security Applications Conference, Miami Beach, FL, USA, 10-14 December 2007, pp. 279-291

[10] A. A. Chandavale, A. M. Sapkal, R. M. Jalnekar: Algorithm to Break Visual CAPTCHA, Proc. of 2nd Int. Conf. on Emerging Trends in Engineering and Technology, Nagpur, India, 16-18 December 2009, pp. 258-262

[11] E. Bursztein, S. Bethard: Decaptcha: Breaking 75% of eBay Audio CAPTCHAs, Proc. of 3rd USENIX Workshop on Offensive Technologies, Montreal, Canada, 10 Aug 2009, pp. 1-7

[12] M. Shirali-Shahreza, S. Shirali-Shahreza: Advanced Collage CAPTCHA, Proc. of 5th In. Conf. on Information Technology: New Generations, Tehran, Iran, 7-9 Apr 2008, pp. 1234-1235

[13] P. Golle: Machine Learning Attacks against the Asirra CAPTCHA, Proc. 15th ACM Conf. on Computer and Communications Security, New York, NY, USA, 17 Oct 2008, pp. 535-542

[14] S. Hocevar, 2004 [Online] Available: http://caca.zoy.org/wiki/PWNtcha. [Accessed: March 2011]

[15] G. Martinovic, A. Attard, Z. Krpic: Proposing a New Type of CAPTCHA: Character Collage, Proc. of the 34th Int. Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija, Croatia, 23-25 May 2011, pp. 1447-1451

[16] G. Martinovic, Z. Krpic: Testing the Reliability of Character Collage CAPTCHA Protection, Proc. of the 27th International Kandó Conference – Science in Practice, Óbuda University, Budapest, 17-18 November 2011

A Hybrid Algorithm for Parameter Tuning in Fuzzy Model Identification

Zsolt Csaba Johanyák, Olga Papp

Department of Information Technologies, Faculty of Mechanical Engineering and Automation, Kecskemét College, Izsáki út 10, H-6000 Kecskemét, Hungary
E-mail: {johanyak.csaba, papp.olga}@gamf.kefo.hu

Abstract: Parameter tuning is an important step in automatic fuzzy model identification from sample data. It aims at the determination of quasi-optimal parameter values for fuzzy inference systems using an adequate search technique. In this paper, we introduce a new hybrid search algorithm that uses a variant of the cross-entropy (CE) method for global search purposes and a hill climbing type approach to improve the intermediate results obtained by CE in each iteration stage. The new algorithm was tested against four data sets for benchmark purposes and ensured promising results.

Keywords: cross-entropy; hill climbing; fuzzy rule interpolation; fuzzy model identification

1 Introduction

Fuzzy systems have been successfully applied in a wide range of areas in this century and the previous. Typical fields are controllers (e.g. [32] [33] [35]), expert systems (e.g. [11] [24]), clustering (e.g. [9] [28] [40]), fuzzy modeling (e.g. [18]), management decision support (e.g. [31] [42]), time series estimation (e.g. [14]), etc. The proper functioning of such systems greatly depends on the underlying rule base. Thus, the methods used for its automatic generation and the determination of the rules' optimal parameters become particularly important.

There are several methods for the automatic generation of the rule base from sample data. Generally, they form two main groups. The methods belonging to the first group (e.g. [6] [8] [41]) create the rule base in two steps. Firstly, they define the structure by creating an initial rule base, and next, they look for an optimal parameter set applying a search algorithm. The methods belonging to the second group (e.g. [19] [39]) differ from this approach only in their second step, when they allow the modification of the structure by creating new rules or deleting old ones.

In our previous work [20], we presented a comparative analysis of a global and a local search algorithm for parameter optimization. They were applied in the second step of a rule base generation conforming to the above mentioned first approach. As a result of the analysis, we found that the local search algorithm ensured a significant improvement of the system performance in case of the used benchmark problems. In comparison, the global search method improved the system performance on three out of four benchmark problems; however its running time was remarkably better than the local heuristic's. This prompted us to implement a hybrid approach, where after enhancing some parts of the two algorithms; we combined the quick run of the global search technique with the increased accuracy of the local heuristic.

In this paper, we present this new hybrid algorithm and the results obtained by its application for finding optimal parameters in the case of the same benchmarking problems as used in [20]. The rest of this paper is organized as follows. Section 2 presents the applied global (sec. 2.1) and local (sec. 2.2) search methods as well as the concept of their combination. Section 3 gives a brief review of the applied fuzzy inference technique. Section 4 reports the results of the tests.

2 Parameter Tuning

The starting point is an initial rule base created with an arbitrary method (e.g. based on fuzzy clustering) automatically from sample data or manually by a human expert. Next, by the help of parameter tuning one tries to find such values for the parameters of the rules that ensure a better performance for the fuzzy system. The performance evaluation method we applied is discussed in sec. 2.4. In the following three subsections we present two search techniques and their proposed integration.

2.1 Cross-Entropy Method

The Cross-Entropy (CE) method is a global search algorithm used for solving continuous multi-extremal and discrete optimization problems, such as buffer allocation [2], static simulation models [12], control and navigation [10], reinforcement learning [27] and others. Its original version was proposed by Rubinstein [34]. The method does not use the local neighborhood structure, instead it works as a black-box and looks for the optimal parameter values using an iterative approach.

Suppose we want to find the best parameter vector p for which our black box yields a performance index $PI(p)$. This parameter (p) should be between a given lower bound (lb) and upper bound (ub). Starting with the first iteration, an initial

probability parameter vector (pr_0) is optimized for each parameter, for example pr_0={0.5, 0.5, ... 0.5}.

In each iteration step i, $S(p_1, p_2,..., p_S)$ samples are generated according to the latest pr_{i-1} probability vector values. Performance index values are calculated for each generated sample, and according to its values the samples are ordered increasingly. After ordering the samples, one of them is chosen according to a parameter q for comparison. The sample with the performance index $g_i = PI_{[1-q]N}$ is chosen. Using g_i the new probability parameter, values are determined by

$$pr_i = \frac{\sum_i I(PI(p_i) \geq g_i) I(p_i \geq lb_i) I(p_i \leq ub_i)}{\sum_i I(PI(p_i) \geq g_i)}, \tag{1}$$

where I is an indicator function which returns *1* if the condition in its parenthesis is true, and *0* otherwise.

The algorithm generates a series of performance index values g_i which get smaller with each iteration, approaching the desired minimum.

The number of the iteration cycles (n_{iCE}), the number of generated samples for each iteration (S), and the optimization parameter q are parameters of the method.

2.2 Hill Climbing Type Local Search

The local search algorithm presented in this subsection is a modified version of the algorithm used by the ACP [16] rule base generation method. It searches for better parameter values through several iterations by applying a hill climbing type approach. The number of iteration cycles (n_{iHC}) is a parameter of the method.

In each cycle all parameters (in all antecedent and consequent dimensions for all fuzzy sets) are examined one-by-one. In the case of each parameter $2 \cdot n_p$ new values are calculated (see Fig. 1) and the fuzzy system is evaluated against the training data set for each new parameter value. Finally, that parameter value is kept from the $2 \cdot n_p + 1$ ($2 \cdot n_p$ new and the original one) candidates that ensures the best system performance. The new parameter values are calculated from the original one by increasing/decreasing its value as follows

$$\begin{aligned} p_i^k &= p_0^k + i \cdot s, i = \overline{1, n_p}, \\ p_i^k &= p_0^k - (i - n_p) \cdot s, i = \overline{n_p + 1, 2 \cdot n_p}, \end{aligned} \tag{2}$$

where p_0^k is the original value of the kth parameter of a fuzzy set, s is the actual step, and n_p is a parameter of the method. Owing to the possible different ranges of the partitions in different dimensions, the step size is calculated by

$$s = c_s \cdot r\,, \tag{3}$$

where r is the range of the actual partition defined by its upper (r_2) and lower (r_1) bounds,

$$r = r_2 - r_1\,, \tag{4}$$

and $c_s \in [0, 1]$ is the step coefficient, which is also a parameter of the method.

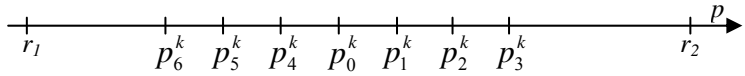


Figure 1
Original and new values of a fuzzy set's kth parameter in case of n_p=3

After calculating the new parameter values, some constraints are applied to preserve the validity and interpretability of the resulting fuzzy sets. These constraints strongly depend on the used membership function types and the parameterization technique. Further on we will present the constraints for the case of piece-wise linear membership functions and break-point type parameterization.

1. The new (ith) parameter value must remain inside its neighbors.
   - If the new value of the actual (kth) parameter is smaller than the previous parameter, it will be increased to that parameter's value

     $$p_i^k = \max\left(p_i^{k-1}, p_i^k\right), k = \overline{2, n_s}, i = \overline{1, 2 \cdot n_p}\,, \tag{5}$$

     where n_s is the number of a fuzzy set's parameters.
   - If the new value is greater than the next parameter it will be reduced to that parameter's value

     $$p_i^k = \min\left(p_i^k, p_i^{k+1}\right), k = \overline{1, n_s - 1}, i = \overline{1, 2 \cdot n_p}\,. \tag{6}$$

2. The set must remain at least partially inside the range.
   - The first parameter must always be smaller or equal to the upper bound of the range of the current linguistic variable (r_2)

     $$p_i^1 = \min\left(p_i^1, r_2\right), i = \overline{1, 2 \cdot n_p}\,. \tag{7}$$
   - The last parameter must always be greater or equal to the lower bound of the range of the current linguistic variable (r_1)

     $$p_i^{n_s} = \max\left(r_1, p_i^{n_s}\right), i = \overline{1, 2 \cdot n_p}\,. \tag{8}$$

Owing to the above-mentioned corrections, two or more new parameter values could result identical. Therefore, the duplicate values are removed from the parameter vector p.

Another feature of the algorithm is that the step coefficient c_s is decreased when the amelioration of the performance index in the course of two consecutive iteration cycles is smaller than the threshold value (dPI_{tr})

$$c_s = c_s \cdot c_d, c_d \in [0,1], \tag{9}$$

where c_d is the decrement coefficient. Its value, as well as the value of dPI_{tr}, are parameters of the algorithm.

2.3 The Hybrid Approach

The basic idea of the hybrid approach is that the local search method is integrated with the global technique as follows. The parameter tuning is started with five steps of the global search method presented in sec. 2.1, where after selection of the samples $\{p_i \mid PI(p_i) \geq g_i\}$ for each selected sample, a local search is launched to find better parameter values in the neighborhood of the initial values determined by the previous step of the CE method. The local search is performed by executing one, two, respectively three steps as indicated in sec. 2.2. The local search results in for each x_i a new p_i^* value set with $PI(p_i^*) \geq PI(p_i)$ performances. Next, the new p_i^* samples are used for the calculation of the probability parameters in (1).

After each parameter modification and system evaluation, the whole parameter set (fuzzy system) and its performance measure against the training data set are saved. After finishing the tuning process, all saved parameter sets are tested against the test data set (PI_{te}) as well. The variation of PI_{tr} and PI_{te} give a good picture about the tuning process, indicating clearly in most of the cases the phenomenon of parameter overfitting to the train data.

For example, supposing an error related performance index which is of type “the smaller the better”, Fig. 2 illustrates the variation of the performance indexes in the function of the number of system evaluations.

In order to minimize the overfitting effect and get a system performing well on the entire input space, an overall system performance (PI_{ov}) is calculated, which takes into consideration both the training and the test data sets. Finally, that parameter set is chosen as the best one that ensures the best PI_{ov} value (indicated by an arrow in Fig. 2).

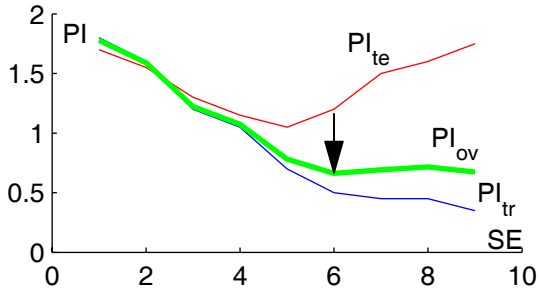


Figure 2
Variation of the performance index in case of the train and test data and the overall performance index in function of the number of system evaluations

2.4 Performance Evaluation

The performance index (*PI*) expresses the quality of the approximation ensured by the fuzzy system using a number that aggregates and evaluates the differences between the prescribed output values and the output values calculated by the fuzzy system. We used as the performance index of the resulting fuzzy systems the root mean squared error, expressed in percentage, of the output variable's range. It was chosen because it facilitates the interpretation of the error and its benchmarking against the width of the variation interval of the output. It is calculated by

$$PI = \frac{1}{r} \cdot \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \cdot 100 \,[\%], \tag{10}$$

where n is the number of data points in the sample, y_i is the ith output value from the sample, and $\hat{y}_i$ is the ith output value calculated by the fuzzy system.

The overall performance indicator (PI_{ov}) of a fuzzy system takes into consideration the performance against both the training (PI_{tr}) and the test (PI_{te}) data sets in a weighted manner, where the weighting expresses the measure of the whole data set's coverage by the two samples. It is calculated by

$$PI_{ov} = \frac{n_{tr} \cdot PI_{tr} + n_{te} \cdot PI_{te}}{n_{tr} + n_{te}} \,[\%], \tag{11}$$

where n_{tr} and n_{te} are the number of data points in the training and test data sets, respectively.

3 Fuzzy Inference by FRISUV

The tuning of the fuzzy sets' parameters can produce a sparse rule base when the modification of the supports is enabled in course of the tuning. A rule base is characterized as sparse when there is at least one possible observation for which none of the rule's activation degree is greater than zero. The activation degree of a rule R_i [37] for an n-dimensional observation A^* is

$$\varpi_{h,t}(R_i) = s\left(t\left(A_{i1}, A_1^*\right), \ldots, t\left(A_{in}, A_n^*\right)\right), i = \overline{1, n_R}, \tag{12}$$

where s is an arbitrary s-norm, t is an arbitrary t-norm, A_{ij} is the antecedent set in the jth dimension of the ith rule, and n_R is the number of rules.

The traditional compositional fuzzy inference methods (e.g. Mamdani [25], Takagi-Sugeno [36], etc.) require a full coverage of the input space by rule antecedents. This demand cannot be fulfilled in sparse rule bases. The recognition of this shortcoming led to the emergence of inference techniques based on fuzzy rule interpolation (e.g. [4] [7] [13] [15] [17] [21] [22] [23] [26] [29]).

In the course of the experiments aimed at testing the new tuning method, the FRISUV [15] inference method was used, owing to its low computational complexity. The key idea of the fuzzy rule interpolation based on subsethood values is that it measures the similarity between the current observation and the rule antecedents, taking into consideration two factors: the shape similarity and the relative distance.

The shape similarity between the observation and the rule antecedent sets is calculated separately in each antecedent dimension by the means of the fuzzy subsethood value. First, the examined antecedent set is shifted into the position of the observation. Here the reference point of the fuzzy set is used for the definition of its position and for the calculation of distances between sets. The fuzzy subsethood value in case of the ith rule and the jth dimension is

$$FSV_{ij} = \frac{\sum_{x \in X_j} \mu_{A_j^* \cap A_{ij}}(x)}{\sum_{x \in X_j} \mu_{A_{ij}}(x)}, \tag{13}$$

where $\cap$ is an arbitrary t-norm, and X_j is the jth dimension of the input universe of discourse. The individual FSVs are aggregated by an average calculation

$$FSV_i = \frac{\sum_{j=1}^{n} FSV_{ij}}{n}. \tag{14}$$

The second aspect of the applied similarity measure is determined based on the Euclidean distance between the two points of the antecedent space defined by the

reference points of the fuzzy sets that describe the current observation and the reference points of the fuzzy sets that form the antecedent part of the current rule. It is a relative distance, defined by

$$d_i = \sqrt{\frac{\sum_{j=1}^{n}\left(RP(A_j^*) - RP\left(A_{ij}\right)\right)^2}{\sum_{j=1}^{n}\left(x_{j\max} - x_{j\min}\right)^2}}, \tag{15}$$

where *RP(.)* denotes the reference point of a fuzzy set, and x_{jmin} and x_{jmax} are the lower and upper bounds in the *j*th antecedent dimension, respectively. Finally, the similarity measure will be

$$S_i = \frac{FSV_i + 1 - d_i}{2}. \tag{16}$$

FRISUV calculates the position of the conclusion adapting the Shepard crisp interpolation [38]

$$RP\left(B^*\right) = \begin{cases} RP\left(B_i\right) & if\ S_i = 1, \\ \dfrac{\sum_{i=1}^{n_R}\dfrac{1}{1-S_i}\cdot RP\left(B_i\right)}{\sum_{i=1}^{n_R}\dfrac{1}{1-S_i}} & otherwise. \end{cases} \tag{17}$$

The method demands that all the sets of the consequent partition have the same shape. Thus the membership function of the conclusion will also share this feature.

4 Results

We performed tests of the hybrid algorithm on four benchmark problems. Three of them were real life problems, namely ground level ozone prediction [30], petrophysical properties prediction [41], yield strength prediction [1] [3], and the fourth was a synthetic function approximation problem. Testing was performed by executing one, two or three local search steps (n_p) after each five global search steps. Table 1 presents the test results. The number of data points (cardinality of the data samples) are summarized in Table 2. The overall performance indicator (PI_{ov}) values are contained in Table 3.

Table 1
Performance of the Systems Tuned by the CE Method compared to the Hybrid Method with one, two respectively three local search steps after each five global search steps

Dataset	CE Method		Hybrid Method with n_p=1		Hybrid Method with n_p=2		Hybrid Method with n_p=3	
	Train	Test	Train	Test	Train	Test	Train	Test
Ozone	14.6531	13.2386	14.5919	13.1057	14.6395	13.1737	14.4234	12.8965
Yield Strength	38.2629	36.1852	26.2513	15.3209	30.0461	22.0258	31.0267	24.3468
Well	27.4533	28.5658	14.8432	13.6063	14.9870	14.0165	14.4390	13.4913
Synthetic	19.6862	18.2116	19.0711	18.2106	17.5059	15.7902	18.8306	18.1833

Table 2
Number of data points in the training (n_{tr}) and test (n_{te}) data sets

Dataset	n_{tr}	n_{te}
Ozone	224	112
Yield Strength	310	90
Well	71	51
Synthetic	196	81

Table 3
Overall performance indicator (PI_{ov}) values

Dataset	CE Method	Hybrid Method		
		n_p=1	n_p=2	n_p=3
Ozone	14.1816	14.0965	14.1509	13.9144
Yield Strength	37.7954	23.7920	28.2415	29.5237
Well	27.9184	14.3261	14.5813	14.0428
Synthetic	19.2550	18.8104	17.0042	18.6413

The application of the Hybrid Method resulted in improvements compared to the usage of the CE method on all datasets. Examining the improvements separately for the case of train and test data samples we can summarize the followings.

In case of the train data samples the least improvement (0.09%) was encountered in case of the ozone data set and n_p=2, while n_p=3 in case of the well data set ensured the best improvement (47.41%). Although in two out of four cases n_p=3 led to a better result, surprisingly the average improvement (20.22%) was observed by n_p=1.

In the case of the test data samples, the improvement varied between 0.01% (synthetic data set and n_p=1) and 57.66% (yield strength data set and n_p=1). In the case of all the samples, the greatest improvement was found by the same local search number as in case of the train data sets. The greatest average improvement (27.76%) was observed by n_p=1.

Evaluating the results based on the overall performance indicator (PI_{ov}), we found a bit narrower variation interval for the improvement([0.60, 49.70]) with an overall average improvement of 20.85%. The greatest variation of PI_{ov}'s improvement due to n_p was 15.17%, in the case of the yield strength sample.

Conclusions

The test results show clearly that the Hybrid Method has great potential in parameter tuning, and the number of local search steps can have a significant influence on the achieved results.

Further research will concentrate on further adjusting the parameters of the presented method and examining the relation between some features of the modeled phenomena and the achieved improvement measure with the help of the Hybrid Method.

Acknowledgment

This research was partly supported by the Hungarian National Scientific Research Fund (Grant No. OTKA K77809) the Normative Application of R & D by Kecskemét College, GAMF Faculty (Grant No. 1KU31).

References

[1] Ádámné, A. M., Belina K.: Effect of Multiwall Nanotube on the Properties of Polypropylenes. Int. J. of Mater. Form., Vol. 1, No 1, 2008, pp. 591-594

[2] Alon, G., Kroese, D. P., Raviv T., Rubinstein, R. Y.: Application of the Cross-Entropy Method to the Buffer Allocation Problem in a Simulation-Based Environment. Ann. of Oper. Res., 2005

[3] Ádámné, A. M., Belina, K.: Investigation of PP and PC Carbon Nanotube Composites. 6th International Conference of PhD Students, Miskolc, 12-18 August 2007, pp. 1-6

[4] Baranyi. P., Kóczy, L. T., Gedeon. T. D.: A Generalized Concept for Fuzzy Rule Interpolation. in IEEE Trans. on Fuzzy Syst., Vol. 12, No. 6, 2004, pp 820-837

[5] Bezdek, J. C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981

[6] Botzheim, J., Hámori, B., Kóczy, L. T.: Extracting Trapezoidal Membership Functions of a Fuzzy Rule System by Bacterial Algorithm, 7th Fuzzy Days, Dortmund, Springer-Verlag, 2001, pp. 218-227

[7] Chen, S. M., Ko, Y. K.: Fuzzy Interpolative Reasoning for Sparse Fuzzy Rule-based Systems Based on α-cuts and Transformations Techniques, IEEE Trans. on Fuzzy Syst, Vol. 16, No. 6, 2008, pp. 1626-1648

[8] Chong, A., Gedeon, T. D., Kóczy L. T.: Projection-based Method for Sparse Fuzzy System Generation. in Proceedings of the 2nd WSEAS International Conference on Scientific Computation and Soft Computing, Crete, Greece, 2002, pp. 321-325

[9] Devasenapati, S. B., Ramachandran, K. I.: Hybrid Fuzzy Model-based Expert System for Misfire Detection in Automobile Engines, International Journal of Artificial Intelligence, Vol. 7, No. A11, 2011, pp. 47-62

[10] Helvik, B. E., Wittner, O.: Using the Cross-Entropy Method to Guide / Govern Mobile Agent's Path Finding in Networks, Lect. Notes in Comp. Sci., 2164/2001, pp. 255-268

[11] Hládek, D., Vaščák, J., Sinčák, P.: Hierarchical Fuzzy Inference System for Robotic Pursuit Evasion Task. in Proceedings of SAMI 2008, 6th International Symposium on Applied Machine Intelligence and Informatics, January 21-22, 2008, Herľany, Slovakia, pp. 273-277

[12] Homem-de-Mello, T., Rubinstein, R. Y.: Estimation of Rare Event Probabilities using Cross-Entropy, WSC 1, 2002, pp. 310-319

[13] Huang, Z. H., Shen, Q.: Fuzzy Interpolation with Generalized Representative Values. In Proceedings of the UK Workshop on Computational Intelligence, 2004, pp. 161-171

[14] Joelianto, E., Widiyantoro, S., Ichsan, M.: Time Series Estimation on Earthquake Events using ANFIS with Mapping Function, International Journal of Artificial Intelligence, Vol. 3, No. A09, 2009, pp. 37-63

[15] Johanyák, Zs. Cs.: Fuzzy Rule Interpolation Based on Subsethood Values. in Proceedings of 2010 IEEE Interenational Conference on Systems Man. and Cybernetics (SMC 2010) 2010, pp. 2387-2393

[16] Johanyák, Zs. Cs.: Sparse Fuzzy Model Identification Matlab Toolbox - RuleMaker Toolbox. IEEE 6th International Conference on Computational Cybernetics, November 27-29, 2008, Stara Lesná, Slovakia, pp. 69-74

[17] Johanyák, Zs. Cs.: Performance Improvement of the Fuzzy Rule Interpolation Method LESFRI, in Proceeding of the 12th IEEE International Symposium on Computational Intelligence and Informatics, Budapest, Hungary, November 21-22, 2011, pp. 271-276

[18] Johanyák, Zs. Cs., Ádámné, A.M.: Mechanical Properties Prediction of Thermoplastic Composites using Fuzzy Models. Scientific Bulletin of "Politehnica" University of Timisoara. Romania. Transactions on Automatic Control and Computer Science, Vol: 54(68), No: 4/2009, pp. 185-190

[19] Johanyák, Zs. Cs., Kovács, S.: Sparse Fuzzy System Generation by Rule Base Extension. in Proceedings of the 11th IEEE International Conference of Intelligent Engineering Systems (IEEE INES 2007) Budapest, Hungary, pp. 99-104

[20] Johanyák, Zs. Cs., Papp, O.: Comparative Analysis of Two Fuzzy Rule Base Optimization Methods. 6th IEEE International Symposium on Applied

Computational Intelligence and Informatics (SACI 2011) 19-21 May 2011, Timisoara, Romania, pp. 235-240

[21] Kóczy, L. T., Hirota, K.: Approximate Reasoning by Linear Rule Interpolation and General Approximation. in International Journal of Approximative Reasoning, Vol. 9, 1993, pp. 197-225

[22] Kovács, L.: Rule Approximation in Metric Spaces. Proceedings of 8th IEEE International Symposium on Applied Machine Intelligence and Informatics SAMI 2010, Herl'any, Slovakia, 2010, pp. 49-52

[23] Kovács, S.: Extending the Fuzzy Rule Interpolation "FIVE" by Fuzzy Observation. Advances in Soft Computing. Computational Intelligence, Theory and Applications, Bernd Reusch (Ed.) Springer Germany, 2006, pp. 485-497

[24] Kovács, S., Kóczy, L. T.: Application of Interpolation-based Fuzzy Logic Reasoning in Behaviour-based Control Structures. Proceedings of the FUZZIEEE. IEEE International Conference on Fuzzy Systems, 25-29 July 2004, Budapest, Hungary, p. 6

[25] Mamdani, E. H., Assilian, S.: An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller, in International Journal of Man Machine Studies, Vol. 7, 1975, pp. 1-13

[26] Detyniecki, M., Marsala, C., Rifqi, M.: Double-Linear Fuzzy Interpolation Method, IEEE International Conference on Fuzzy Systems (FUZZ 2011), Taipei, Taiwan, Jun 27-30, 2011, pp. 455-462

[27] Menache, I., Mannor, S., Shimkin, N.: Basis Function Adaptation in Temporal Difference Reinforcement Learning, Ann. of Oper. Res., 134, 2005, pp. 215-238

[28] Palanisamy, C., Selvan, S.: Wavelet-based Fuzzy Clustering of Higher Dimensional Data, International Journal of Artificial Intelligence, Vol. 2, No. S09, 2009, pp: 27-36

[29] Perfilieva, I., Wrublova, M. Hodakova, P.: Fuzzy Interpolation According to Fuzzy and Classical Conditions, Acta Polytechnica Hungarica, Vol. 7, Issue 4, Special Issue: SI, 2010, pp. 39-55

[30] Pires, J. C. M., Martins F. G., Pereira M. C., Alvim-Ferraz M. C. M.: Prediction of Ground-Level Ozone Concentrations through Statistical Models. in Proceedings of IJCCI 2009 - International Joint Conference on Computational Intelligence, 5-7 October 2009 Funchal-Madeira, Portugal, pp. 551-554

[31] Portik, T., Pokorádi, L.: Possibility of Use of Fuzzy Logic in Management. 16th Building Services. Mechanical and Building Industry days" International Conference, 14-15 October 2010, Debrecen, Hungary, pp. 353-360

[32] Precup, R. E., Doboli, S., Preitl, S.: Stability Analysis and Development of a Class of Fuzzy Systems, Eng. Appl. of Artif. Int., Vol. 13, No. 3, June 2000, pp. 237-247

[33] Precup, R.-E., Preitl, S.: Optimisation Criteria in Development of Fuzzy Controllers with Dynamics. Eng. Appl. of Artif. Intell., Vol. 17, No. 6, 2004, pp. 661-674

[34] Rubinstein, R. Y.: The Cross-Entropy Method for Combinatorial and Continuous Optimization, Methodol. and Comput. in Appl. Probab, 1999

[35] Škrjanc, I., Blažič, S., Agamennoni, O. E.: Identification of Dynamical Systems with a Robust Interval Fuzzy Model, Automatica, 2005, Vol. 41, No. 2, pp. 327-332

[36] Takagi, T. and Sugeno, M.: Fuzzy Identification of Systems and its Applications to Modeling and Control, in IEEE Transactions on System, Man and Cybernetics, Vol. 15, 1985, pp. 116-132

[37] Tikk, D., Johanyák, Zs. Cs., Kovács, S., Wong, K. W.: Fuzzy Rule Interpolation and Extrapolation Techniques: Criteria and Evaluation Guidelines, Journal of Advanced Computational Intelligence and Intelligent Informatics, ISSN 1343-0130, Vol. 15, No. 3, 2011, pp. 254-263

[38] Shepard, D.: A Two Dimensional Interpolation Function for Irregularly Spaced Data. In Proceedings of the 23rd ACM International Conference, 1968, New York, USA, pp. 517-524

[39] Vincze, D., Kovács, S.: Incremental Rule Base Creation with Fuzzy Rule Interpolation-based Q-Learning. Studies in Computational Intelligence - Computational Intelligence in Engineering, Vol. 313, 2010, pp. 191-203

[40] Wang, W., Zhang, Y.: On Cluster Validity Indices. Fuzzy Sets and Systems, 158, 2007, pp. 2095-2117

[41] Wong, K. W., Gedeon, T. D.: Petrophysical Properties Prediction Using Self-generating Fuzzy Rules Inference System with Modified Alpha-Cut-based Fuzzy Interpolation. Proceedings of the Seventh International Conference of Neural Information Processing ICONIP 2000, November 2000, Korea, pp. 1088-1092

[42] Zemková, B., Talašová, J.: Fuzzy Sets in HR Management. Acta Polytechnica Hungarica, Vol. 8, No. 3, 2011, pp. 113-124

The Analysis of Two-Phase Condensation Heat Transfer Models Based on the Comparison of the Boundary Condition

Róbert Sánta

College of Applied Sciences – Subotica Tech
16 Marka Oreskovica, 24000 Subotica, Serbia
santa@vts.su.ac.rs

Abstract: The article aims to present, analyze and select two-phase condensation heat transfer coefficients for the refrigerant-side in the condenser of the heating and cooling system of a heat pump. The heat transfer models are analyzed for condensation in horizontal smooth tubes. The mathematical models published by various authors refer to the range of heterogeneous condensation. The final aim of the analysis is to select the optimal model from the examined two-phase condensation heat transfer models. The selection of the condensing refrigerant two-phase heat transfer model is based on the boundary condition. The applied method of analysis is numerical-graphical.

Keywords: heat pump; condenser; R134a; heat transfer; boundary condition

Nomenclature

D	inside diameter of tube [m]	*Greek Symbols*	
x	vapor quality [-]	α	heat transfer coefficient[W/m^{2}K]
G	mass velocity [kg/m^{2}s]	λ	thermal conductivity [W/mK]
T	temperature [K]	ρ	density [Ns/m^2]
p	pressure [Pa]	ε	void fraction [-]
p*	reduced pressure [-]	η	dynamic viscosity [Ns/m^2]
Re	Reynold`s number [-]		
Pr	Prandt number [-]	*Subscript*	
F	two phase multiplier	g	vapor
Fr	Froude number [-]	f	liquid
Z	Shah`s correlating parameter [-]	kf	two phase
X_{tt}	Martinelli`s correlating parameter [-]		
C	Constant [-]		

1 Introduction

The condenser heat exchanger plays a significant role in the structure and operation of the heat pump as it affects the system's coefficient of performance (COP).

The motivation for the composition of this article was the selection of the optimal two-phase heat transfer mathematical model of refrigerant among the analyzed mathematical models in the horizontal tube of the condenser. The structure and dimensions of the condenser have a significant impact on the heat transfer intensity. The parameters that influence heat transfer are the heat transfer coefficients of the refrigerant and of the heated water.

The earliest determination of the condensation heat transfer coefficient of the two-phase refrigerant flowing in horizontal smooth tubes was carried out by Boyko and Kruzhilin [1] and Akers et al. [2]. For the stationery condensation, there are a large number of mathematical models for different conditions and refrigerants, authored by Cavallini and Zecchin [3, 4], Shah [5], M. K. Dobson, J. C. Chato [6], and J. R. Thome et al. [7]. The common characteristics of these heat transfer models are that under the same conditions they give a high difference in values of heat transfer coefficient. M. M. Awad et al. [8] in their mathematical model used the heat transfer correlations.

A number of researchers have dealt with this field of science and implemented heat transfer coefficients in numerous mathematical models of heat pump; see J. Nyers et al. [9-12].

2 The Physical Model of Condenser with Horizontal Smooth Tubes

The condenser is a heat exchanger where the higher temperature refrigerant gives heat to the lower temperature heating medium. The analyzed condenser has parallel, straight and smooth tube bundles with baffles. The refrigerant flows inside tubes, while the heated water flows in the shell tube.

The process in the condenser on the refrigerant side is made up of three sections:

A the Superheated vapor section

B the Condensation section

C the Subcooled liquid section

After compression, the vapor is single-phase and superheated. In the superheated section, the value of vapor quality equals 1.0.

The vapor is in contact with the tube wall, whose temperature is lower than the saturation temperature, and therefore the heat transfers to the wall and vapor condenses. Heat is then transferred from the tube wall to the water by conduction.

In the condensation section, the condensate and the vapor have a heterogeneous flow, which is characterized by intensive turbulence.

At the beginning of the condensation section, the flow pattern is annular, because in the core the velocity of the vapor is much higher than the velocity of the liquid. In the annular flow regime at the liquid-vapor interface, the dominant force is tangential stress, with the gravitational force playing a less important role. As condensation continues, the velocity of the vapor phase reduces and the dominant force shifts from tangential force to gravitational force. The liquid phase accumulates at the bottom of the tube. Condensation takes place mainly at the top of the tube because the liquid layer is thin, and therefore heat resistance is smaller.

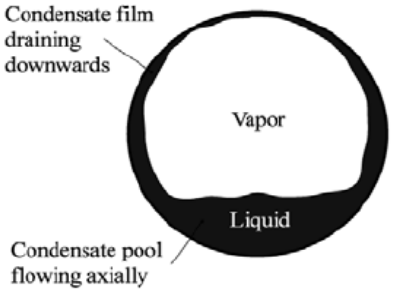


Figure 1
Condensation in fully annular flow in horizontal tube

In the continuation of the condensation, the vapor continually condenses, and in the cross section area the surface of liquid increases, and the flow pattern changes to slug and plug pattern.

The size of the vapor slug reduces further and a bubbly flow pattern develops. At the end of condensation section, the vapor quality reduces to zero and the flow in the tube becomes a single-phase flow (liquid flow).

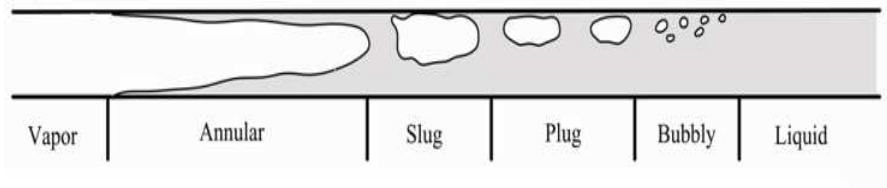


Figure 2
Flow pattern map for condensation in a horizontal tube

A sub-cooling section is only formed in exceptional circum stances, maybe in the cross-flow condensers. In the case of shell-tube condensers, the sub-cooling is realized in a separate heat exchanger (subcooler). In the subcooler, the refrigerant is in the single-phase (liquid phase).

3 Annular Flow Condensation

One of the most important flow regimes is annular flow, which is characterized by a phase interface separating a thin liquid film from the gas flow in the core region. This flow regime is the most investigated one both analytically and experimentally because of its practical significance.

Much of the condensation process occurs in the annular flow regime. Therefore, many of the existing in-tube condensation correlations are based on the annular flow regime. These correlations are classified into three categories, that is, shear-based correlations, boundary layer-based correlations, and two-phase multiplier-based correlations. The majority of smooth-tube heat transfer correlations are of the two-phase multiplier-based variety.

4 The Two-Phase Multiplier-based Heat Transfer Correlations in Condensation Section

The simplest method of heat transfer prediction in the annular flow regime is the two-phase multiplier approach. Two-phase multiplier-based correlations were pioneered for predicting convective evaporation data (Dengler and Addoms 1956) and were adapted for condensation by Shah (1979).

The two-phase multiplier based heat transfer correlations typically result in the following form:

$$\alpha_{kf} = \alpha_f \cdot F_{kf}$$

Where:

$$\alpha_f = f \quad (C, \mathrm{Re}^n, \mathrm{Pr}^m) \text{ and } F_{kf} = f \quad \left(x, \frac{\rho_f}{\rho_g}, \frac{\eta_f}{\eta_g}, Fr_f \right) \tag{1}$$

The two-phase multiplier can depend on more dimensionless groups than those indicated in Eq. 1; the shown groups are the most prevalent.

Several examples of two-phase multiplier-based condensing correlations are available, including those of Akers et al. (1959), Boyko and Kruzhilin (1967), Cavallini and Zecchin (1974), Tang (1998) and Dobson and Chato (1998).

4.1 The Cavallini and Zecchin Correlation

Cavallini and Zecchin developed a semi-empirical equation that has a simple form.

The mathematical model for heat transfer by Cavallini and Zecchin correlation is:

$$\alpha_{kf} = \frac{\lambda_f}{D} \cdot 0.05 \cdot R_e^{0.8} \cdot \mathrm{Pr}_f^{0.33} \tag{2}$$

The equivalent Reynolds number is:

$$\mathrm{Re} = \mathrm{Re}_g \cdot \left(\frac{\rho_f}{\rho_g}\right)^{0.5} \cdot \left(\frac{\eta_g}{\eta_f}\right) + \mathrm{Re}_f \tag{3}$$

where $\mathrm{Re_v}$ and $\mathrm{Re_l}$ are the Reynolds number of the liquid and vapor phase, respectively, which can be calculated by Eq. 4.

$$\mathrm{Re}_l = \frac{G \cdot (1-x) \cdot d_i}{\mu_l}, \ \mathrm{Re}_g = \frac{G \cdot x \cdot d_i}{\mu_l} \tag{4}$$

In Eq. 2 the application range of Cavallini and Zecchin`s correlation was summarized as follows:

$$d_i = 8mm, 30 < T_{sat} < 50^{\circ}C,$$

$$10 < \rho_l / \rho_v < 2 \cdot 10^3, \ 10 < \mu_l / \mu_v < 2 \cdot 10^3,$$

$$0.8 < \mathrm{Pr}_l < 20, \ 1.2 \cdot 10^3 < \mathrm{Re}_l$$

Flow regime: Annular flow

Refrigerants tested: R-11, R-12, R-21, R-22, R-113, R-114, R-134a, R-410A, R-407C

4.2 Shah Correlation

The Shah correlation takes into account the pressure of the refrigerant, in addition to the quality of the mixture. This can also be used to find the local condensation heat transfer coefficient. The heat transfer coefficient is a product of liquid heat transfer coefficient given by the Dittus-Boelter equation and an additional term.

In 1979, Shah presented the following correlation:

$$\frac{\alpha_{kf}}{\alpha_f} = 1 + \frac{3.8}{Z^{0.95}} \tag{5}$$

The liquid heat transfer coefficient is calculated by using the Dittus-Boelter equation:

$$\alpha_f = 0.023 \cdot \mathrm{Re}_f^{0.8} \cdot \mathrm{Pr}^{0.4} \cdot \frac{\lambda_f}{D} \tag{6}$$

The mathematical model for heat transfer by the Shah correlation is:

$$\alpha_{kf} = \alpha_f \cdot \left[(1-x)^{0.8} + \frac{3.8 \cdot x^{0.76} \cdot (1-x)^{0.04}}{p^{*0.38}} \right] \tag{7}$$

The application range of Shah's correlation was summarized as follows:

$$7 < d_i < 40,\ 10 < P_{sat} < 9.87 MPa,$$

$$21 < T_{sat} < 310^{\circ} C,\ 10.8 < G_{cr} < 1.599 \, {kg}/{m^2 s},$$

$$0.5 < \mathrm{Pr}_l,\ 350 < \mathrm{Re}_l,$$

Flow regime: Annular flow

Refrigerants tested: R-718, R-11, R-12, R-22, R-113, methanol, ethanol, benzene, toluene, and ethylene

4.3 The Boyko and Kruzhilin Correlation

The Boyko and Kruzhilin correlation is an adaptation of the Mikheev correlation. The correlation is simple to use, generally conservative, and sufficiently accurate. This correlation takes into account the heat transfer coefficient in single-phase flow, the density of the two-phase flow and vapor quality.

The mathematical model for heat transfer by the Boyko and Kruzhilin correlation is:

$$\alpha_{kf} = \alpha_f \cdot \left(1 + x \cdot \left(\frac{\rho_l}{\rho_g} - 1 \right) \right)^{0.5} \tag{8}$$

The liquid heat transfer coefficient is obtained as follows:

$$\alpha_f = 0.021 \cdot \mathrm{Re}_f^{0.8} \cdot \mathrm{Pr}^{0.43} \cdot \frac{\lambda_f}{D} \tag{9}$$

The application range of Boyko and Kruzhilin's correlation was summarized as follows:

$$1500 < \mathrm{Re} < 15000$$

Flow regime: Annular flow

Refrigerants tested: steam, R22

4.4 Akers Correlation

Akers et al. (1959) developed a two-phase multiplier-based correlation that became known as the "equivalent Reynolds number" model. This model defines the all-liquid mass flow rate that provides the same heat transfer coefficient as an annular condensing flow.

The mathematics model for heat transfer by the Akers correlation is:

$$\alpha_{kf} = \frac{\lambda_f}{D} \cdot 0.05 \cdot R_e^{0.8} \cdot \mathrm{Pr}_f^{0.33} \tag{10}$$

where:

the equivalent mass velocity is: $G_{ekv.} = G \cdot \left[(1-x) + x \cdot \left(\frac{\rho_f}{\rho_g} \right)^{0.5} \right]$ (11)

The multiplier factors function of the Reynolds number is:

$C = 0.0265$ and $n = 0.8$ if $\mathrm{Re}_f \rangle 50000$

$C = 5.03$ and $n = \frac{1}{3}$ if $\mathrm{Re}_f \langle 50000$

Flow regime: Annular flow

Refrigerants tested: R-12, Propane, Methanol

4.5 Dobson et al.

Dobson et al. (1998) proposed separate correlations for the wavy and annular flow regimes. For the annular flow regime, they correlated condensation data by assuming that the ratio of the two-phase Nusselt number to the Nusselt number predicted by a single-phase correlation is exclusively a function of the Martinelli-parameter. Utilizing the Dittus-Boelter correlation to predict the single-phase Nusselt number, regression analysis of annular flow data yielded the following form:

$$Nu = 0.023 \cdot \mathrm{Re}_l^{0.8} \cdot \mathrm{Pr}_l^{0.4} \cdot g(X_{tt}) \tag{12}$$

Where: $g(X_{tt}) = 1 + \frac{2.22}{X_{tt}^{0.889}}$

The function g at the end of equation 12 represents the two-phase multiplier, and the values of the constant and the exponent are similar to the values from convective evaporation correlations [Wattelet et al., 1992].

The mathematical model for heat transfer by the Dobson et al. correlation for annular flow is:

$$\alpha_{kf} = \frac{\lambda_f}{D} \cdot 0.023 \cdot \mathrm{Re}_l^{0.8} \cdot \mathrm{Pr}_l^{0.4} \cdot \left[1 + \frac{2.22}{X_{tt}^{0.89}} \right] \tag{13}$$

where the application range of Dobson et al.'s correlation was summarized as follows:

$$d_i = 4.57mm,\ 35 < T_{sat} < 60^{\circ}C, \quad 75 < G_{cr} < 500\ {}^{kg}\!/_{m^2 s}, \quad 0.1 < x < 0.9,$$

Flow regime: Annular flow

Refrigerants tested: R-22, R-134a, R-410A

5 Comparison of Heat Transfer Correlations

Many correlations that are available come with no explicit range of parameters over which they can be expected to give accurate results. Therefore, a design engineer requiring a suitable heat transfer correlation typically encounters a series of seemingly contradictory reports about which correlation is "best".

The heat transfer process in annular two-phase flow is similar to that in single-phase flow of the liquid, and thus their ratio may be characterized by a two-phase multiplier. This concept uses the same rationale as the Lockhart-Martinelli (1949) two-phase multiplier, developed for the prediction of two-phase frictional pressure drop.

This work investigated the comparison of the two-phase correlations on the basis of the above-mentioned similarities. The deviation of values of the two phase heat transfer correlation was examined in the values of boundary condition, i.e. single-phase heat transfer coefficient values of liquid. The liquid phase heat transfer coefficient was examined the reference value.

The single-phase heat transfer coefficients are typically predicted by the Dittus and Boelter (1930) [13] correlation, which results in the following form:

$$\alpha_{kf} = \frac{\lambda_f}{D} \cdot 0.023 \cdot R_e^{0.8} \cdot \mathrm{Pr}_f^{0.33} \tag{14}$$

Shah [14] also compared the condensation heat transfer correlations. Shah's basis of comparison was his own model, namely the two-phase heat transfer correlation.

6 Initial Condition and Values

The mathematical models are simulated by the use of the software tool MathCAD. The initial conditions and values for the simulation of the stationary condensation are:

Refrigerant: R134a

Mass velocity: $G = 100 - 500 \quad \left[\frac{kg}{m^2 s}\right]$

Vapor quality ranged: x = 1 - 0 [-]

Pressure in the inlet of condenser: p_k= 15 [bar]

Inner diameter of tube: d= 6 [mm]

Below, the reader can find graphs drawn for the considered condensation heat transfer correlations, showing the deviation of the condensation heat transfer coefficients from the reference value and average of local condensation heat transfer in function of vapor quality and different mass velocity.

7 Results and Discussions

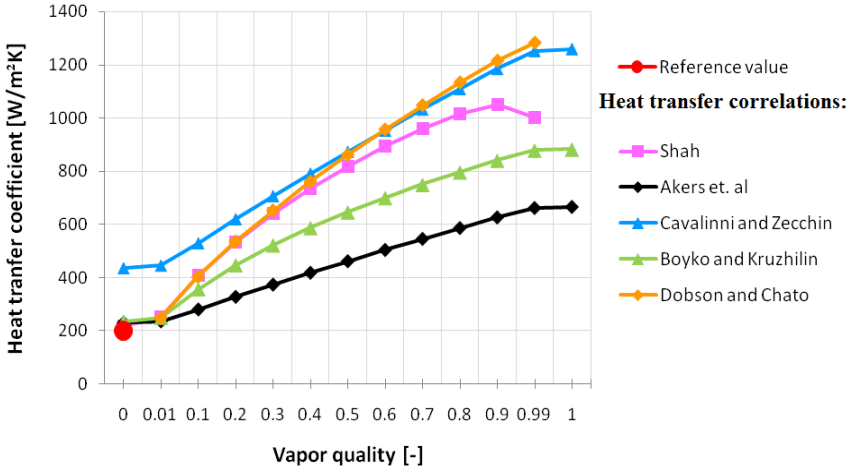


Figure 3

Condensation heat transfer coefficients flow in tube with reference value shown, $G = 100 \quad \left[\frac{kg}{m^2 s}\right]$

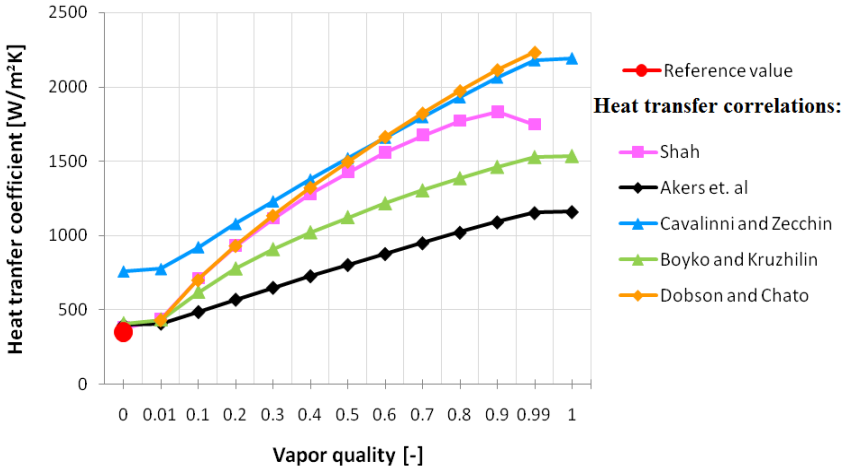


Figure 4

Condensation heat transfer coefficients flow in tube with reference value shown, $G = 200 \left[\frac{kg}{m^2 s}\right]$

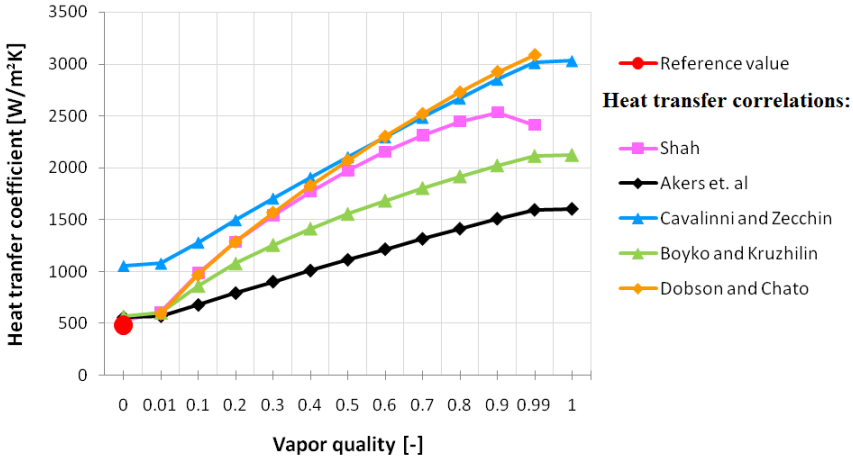


Figure 5

Condensation heat transfer coefficients flow in tube with reference value shown, $G = 300 \left[\frac{kg}{m^2 s}\right]$

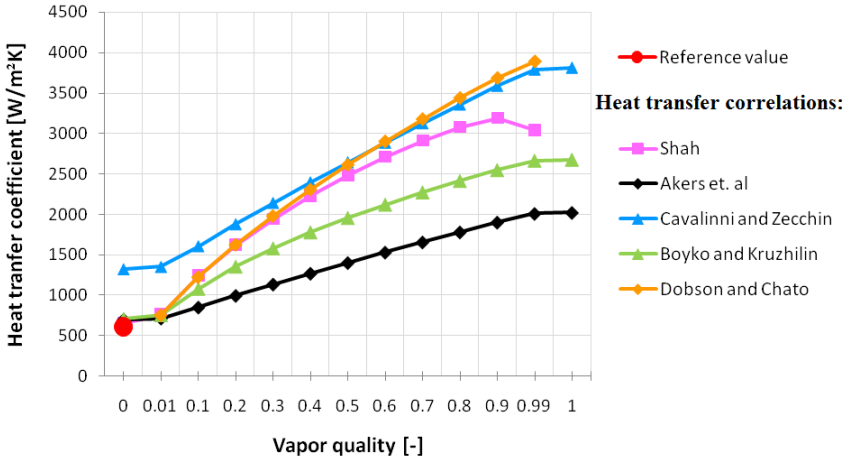


Figure 6

Condensation heat transfer coefficients flow in tube with reference value shown, $G = 400 \left[\frac{kg}{m^2 s}\right]$

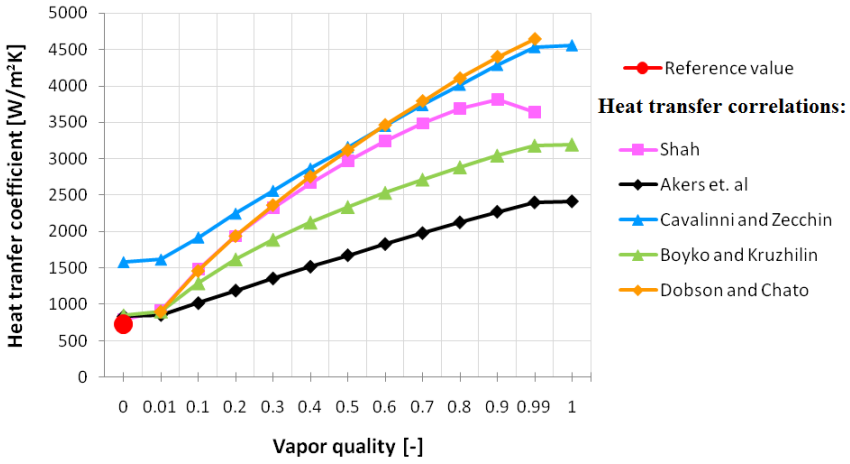


Figure 7

Condensation heat transfer coefficients flow in tube with reference value shown, $G = 500 \left[\frac{kg}{m^2 s}\right]$

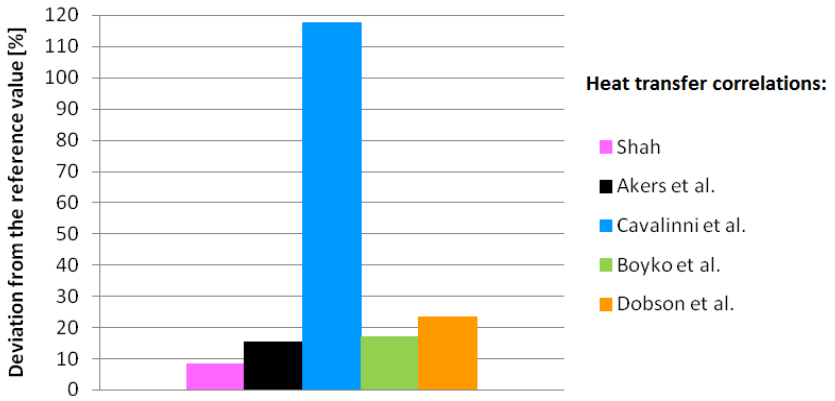


Figure 8
The deviation of the condensation heat transfer coefficients from the reference value, when the vapor quality x=0

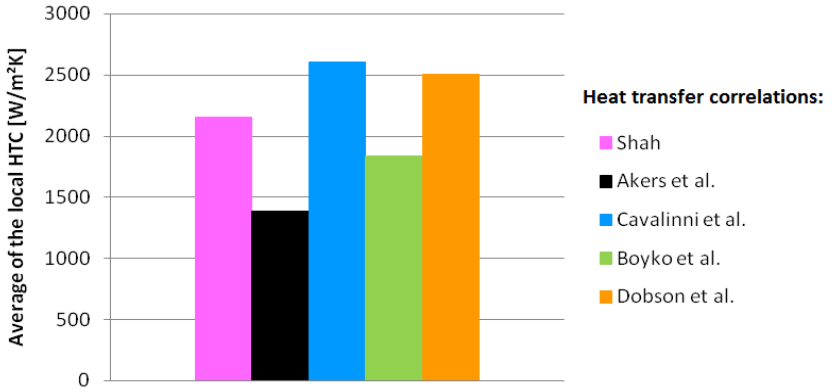


Figure 9
The average condensation heat transfer coefficient, $G = 400 \left[\frac{kg}{m^2 s}\right]$

Conclusions

The condensation heat transfer coefficients linearly increases in the function of the vapor quality of the refrigerant.

For vapor quality x=0.01 (x approximately 0), the values of the condensation heat transfer coefficient obtained from the analyzed models are approximately equal to the values of the models which have been used exclusively for the liquid phase x=0, published by several authors. The Cavalinni et al. condensation heat transfer coefficient is an exception to this rule.

In the case of pure liquid phase, i.e. x=0, the examined condensation heat transfer correlation provide real values, whereas the Dobson et al. correlation is an exception. The Dobson et al. condensation heat transfer correlation is not suitable for determining the single phase heat transfer coefficient.

The heat transfer coefficient of the Shah condensation correlation differs in the smallest amount from the reference value out of the examined correlations. This deviation is only 8.36%, while the heat transfer coefficient of the Cavalinni et al. correlation provides the highest level of deviation. The author examined the heat transfer coefficients deviation from the reference value at x=0 vapor quality and different mass velocity of refrigerant. The Dobson correlation values are an exception to this, where the vapor quality was x=0.01.

In the case of the pure vapor phase, i.e. x=1 vapor quality, the examined condensation heat transfer correlation presents real values of heat transfer coefficient; however, the Shah and Dobson correlation values are exceptions to this rule.

The majority of the condensation heat transfer correlations provide the highest condensation heat transfer coefficient near x = 0.99, Shah's correlation being an exception.

For vapor quality x=0.99, (x approximately 1),the calculated values of the condensation heat transfer coefficients by using the investigated models are much higher than by the value of single heat transfer coefficients suggested by other authors.

The deviation between reference value and calculated heat transfer coefficients is constant with an increase in the mass velocity of the refrigerant.

The average values of the local heat transfer coefficients also provide significant dispersion.

Based on the above analysis and facts, Shah's model is considered as an optimum.

References

[1] Boyko, Kruzhilin: Heat Transfer and Hydraulic Resistance during Condensation of Steam in a Horizontal Tube and in a Bundle of Tubes, International Journal Heat and Mass Transfer 10, 1967, 361-73

[2] Akers, Deans, Crosser: Condensing Heat Transfer within Horizontal Tubes, Chem. Eng. Prog. Symp. Series 55, 1959, 171-6

[3] Cavallini, Zecchin: High Velocity Condensation of Organic Refrigerants Inside Tubes, 8th International Congress of Refrigeration, Vol. 2, Washington DC, 1971, pp. 193-200

[4] Cavallini, Zecchin: A Dimensionless Correlation for Heat Transfer in Forced Convection Condensation. 6th Int. Heat Transfer Conf. Tokyo, 1974, pp. 309-313

[5] Shah: A General Correlation for Heat Transfer During Flow Condensation Inside Pipes, Journal of Heat and Mass Transfer 22, 1979, 547-56

[6] M. K. Dobson, J. C. Chato: Condensation in Smooth Horizontal Tubes, Journal of Heat Transfer 120, 1998, 193-213

[7] J. R. Thome, J. El Hajal, A. Cavallini: Condensation in Horizontal Tubes, Part 2: New Heat Transfer Model Based on Flow Regimes, International Journal of Heat and Mass Transfer 46, 2003, 3365-3387

[8] M. M. Awad, H. M. Mostafa, G. I. Sultan, A. Elbooz, A. M. K. El-ghonemy: Performance Enhancement of Air-cooled Condensers, Acta Polytechnica Hungarica, Volume 4, No. 2, pp. 125-142, 2007

[9] J. Nyers, A. Nyers: COP of Heating-Cooling System with Heat Pump, 3th International Symposium Express 2011, Proceedings 17-21, Subotica, Serbia

[10] J. Nyers et al.: Energy Optimum of Heating System with Heat Pump, 6thInternational Multidisciplinary Conference, Proceedings 545-550, Baia Mare-Nagy Bánya, Romania 2005

[11] J. Nyers, G. Stoyan: A Dynamical Model Adequate for Controlling the Evaporator of a Heat Pump, International Journal of Refrigeration, 1994, Vol. 17, No. 2, pp. 101-108

[12] J. Nyers et al.: Csőköteges elpárologtató hőátadási tényezőjének matematikai modelljei kétfázisú hűtőközegre, Magyar Épületgépészet, LIX. évfolyam, 2010/6. szám, Budapest, Hungary

[13] Dittus, Boelter: Heat Transfer in Automobile Radiators of the Tubular Type. Publications in Engineering, Vol. 2, University of California, Berkeley, p. 443 (1930)

[14] Shah: An Improved and Extended General Correlation for Heat Transfer during Condensation in Plain Tubes, HVAC&R Research, pp. 889-913

A Comparative Study of Offset Plate Quality Parameters using Image Processing and Analytical Methods

Živko Pavlović[1], Tadeja Muck[2], Aleš Hladnik[2], Igor Karlović[1]

[1] University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
zivkopvl@uns.ac.rs; karlovic@uns.ac.rs

[2] University of Ljubljana, Faculty of Natural Sciences and Engineering, Department of Textiles, Chair of Information- and Graphic Arts Technology, Snežniška 5,1000 Ljubljana, Slovenia
tadeja.muck@ntf.uni-lj.si; ales.hladnik@ntf.uni-lj.si

Abstract: Printing plate performance is one of the crucial factors affecting the printed product quality. In particular, when producing printing plates for larger print runs, a rigorous quality control has to be performed, since any bulk and surface imperfection can have a large detrimental effect on print sharpness, contrast, non-uniformity, colour gamut, and other print properties. In our research two characteristics of an aluminium-based offset printing plate after 0, 123,000, 177,000 and 300,000 runs were studied: surface topography and contact angle. Each was determined using two assessment methods. For the surface roughness determination, the conventional mechanical profilometry and scanning electron microscope (SEM) imaging followed by an implementation of a texture analysis method Gray level co-occurrence matrix (GLCM) were applied. The contact angle of a printing plate with water was acquired by the sessile drop method using a DataPhysics OCA30 instrument and, alternatively, by implementing image analysis routines on a sequence of images captured by a simple image acquisition system. Image analysis for both surface topography and contact angle assessment was accomplished using ImageJ, a public domain Java image processing program. Both image analysis-based evaluation methods proved to be a viable alternative to the two established ones, providing a reliable tool for the monitoring of the wearing and other surface changes of a printing plate during the print run.

Keywords: surface roughness; contact angle; scanning electron microscopy; gray level co-occurrence matrix

1 Introduction

The surface quality of materials includes many engineering performance factors, such as mechanical function, wear, lubrication, and appearance. Several surface topography assessment techniques are currently used, and it is desirable that the characteristics of the material under investigation be studied comparatively in view of the practical applications [1]. The surface properties of a particular material are generally described in terms of its chemical composition, morphology and topography. Although all three surface aspects are important for the quality and functionality of materials and products, the influence of topography is frequently underestimated [2]. Surfaces with a non-periodic roughness pattern, such as aluminium printing plates used for lithographic applications, require accurate topography characterization, since this is one of the most important engineering factors that determine product characteristics.

The production of aluminium plates for offset printing involves roughening of the aluminium substrate in order to increase its surface area, which is necessary to improve the adhesion of the photosensitive coating and to enhance the water repellent properties of the aluminium surface [3]. Stability and surface definition are crucial parameters during the production and processing of printing forms. During the printing process, these properties of an aluminium offset plate play a key role in adsorption – wetting process as a function of contact angle between a solid surface (aluminium oxide) and a liquid (fountain solution with water as the basic component). Wetting is a physical phenomenon conditioned by a decrease in the surface tension of a solid – liquid system. A liquid will wet a solid surface only if it has lower surface energy than the solid surface [4]. One can determine the degree of wetting by monitoring a liquid drop shape. Contact angle is an angle between two tangents, one touching the solid surface and second touching a drop of liquid at the intersection point of three phases (solid, liquid and vapour). By measuring contact angle, one can determine the wetting degree of the investigated system [5]. The aim of our study was to compare the performance of the conventional surface topography and contact angle assessment methods with those using image analysis routines implemented through ImageJ, a public domain Java image processing program. The comparison was made in order to evaluate the possibility of utilizing the latter two methods for characterization of surface topography and wettability of printing plates and to determine their usefulness in quality control in the graphic arts technology. The applied methodology is presented in Figure 1.

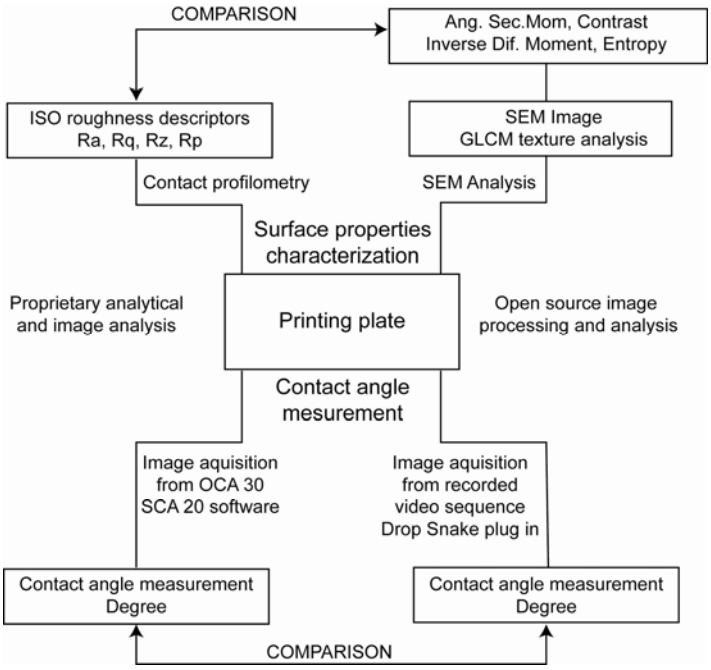


Figure 1
Applied methodology for surface roughness and contact angle characterization

2 Materials and Methods

In this research, we used one particular type of lithographic printing plate with a rather uniform surface structure and roughness of the non-image (aluminium oxide) areas. This selection was made for the following two reasons: first, the printing plates are manufactured in accordance with stringent, standardized procedures [6] assuring that the quality and the size of the grained surface microstructure will influence the printing performance and durability of the printing plate in a suitable way [7]; and second, the printing plate surface characterization during the production stage is of major technological importance [8, 9]. In addition, such a choice was also motivated by the fact that the surface characterization methods examined in this study and the results obtained could be of interest for the whole nanotechnology community, as the aluminium oxide nano-templates play a significant role in the template-based approach to nanotechnology [10-12].

Material used in this study was the thermal positive Kodak Sword Ultra T98 printing plate with a 0.3 mm thick AA1050 aluminium foil electrochemically roughened and anodized. The imaging was made on circular sample areas, of the

radius R = 1.5 cm, from the same non-image (non-printing) region of the printing plate sample, positioned along the line of printing pressure in the printing sets (Figure 2).

Figure 2
Locations of the samples on the printing plate

2.1 Surface Roughness Measurements

Profilometric measurements were made on the samples from an unused (reference) printing plate and on the samples taken after print runs of 123,000, 177,000 and 300,000 impressions. The printing process was performed on the four colour web offset printing press Komori 38 D, which has the ability to print with heat-set printing inks, a maximal printing area of 1250×960 mm, and a top speed 36,000 prints per hour. For each colour, a single printing plate was used and the measurements were done on each printing plate. For this investigation we utilized the data obtained from measurements on the samples from the plates used in the second printing unit, assuming that this printing unit has contact with the paper which has a small amount of dust and other substances left after the first printing unit: ink, fountain solution, etc. It should be emphasized that the printing ink, i.e. the colour, itself had no influence on the measurement results, as the measurements were accomplished on the samples from the non-imaging (non-printing) area of the printing plate.

To make aluminum suitable for producing printing forms, the plate is processed by a rolling procedure, which results in the characteristic structure of the surface in the direction of rolling. Lines that occur on the surface are not desirable in the further preparation of aluminum and require special treatment to reduce their negative impact on the surface roughness. The processing of aluminum includes the processes of electrochemical surface roughening and anodic oxidation, as was mentioned earlier, which produces aluminum surface microstructure of porous aluminum oxide [13]. Therefore, the measurements of surface roughness of the printing forms are carried in x and y direction, i.e. in the direction of aluminum rolling and perpendicularly to it.

Since previous investigations [14] showed that a high depth of focus SEM can provide detailed topographical information about the surface but cannot yield quantitative topographical information, we analysed the printing plates before and after print runs by a roughness meter (Time Group TR200) and by a SEM, and we

thus combined the quantitative topographical information with that contained in the SEM micrographs.

The profilometric measurements were performed with the Portable Surface Roughness Tester TR200 from Micro Photonics, Inc. using a diamond tip with 2 µm radius. The TR200 generates a number of roughness parameters: Ra, Rz, Ry, Rq, Rt, Rp, Rmax, Rm, R3z, S, Sm, Sk, tp, and hybrid parameters: primary profile (P), roughness profile (R), and tp curve (material ratio Mr), all defined according to the pertinent ISO standards [15, 16]. The relevant setting sof the device are presented in Table 1.

Table 1
Portable Surface Roughness Tester TR200measurement settings

Sampling length	Traversing speed	Measuring range	Resolution
0.25 mm	0.135 mm/s	± 20 µm	0.01 µm

The measured surface roughness parameters are compliant to the geometric product specification standards [6, 15, 16] and listed below:

- Ra - average surface roughness:

$$\mathrm{Ra} = \frac{1}{l}\int_0^l |y(x)|dx \tag{1}$$

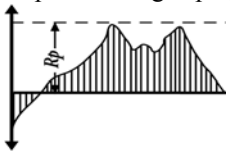
- Rq – root-mean-square roughness(R_{rms}):

$$Rq = \sqrt{\frac{1}{l}\int_0^l y^2(x)dx} \tag{2}$$

R_{zDIN} – mean value of the single roughness depth Z_i:

$$RzDIN = \frac{1}{n}(Z_1 + Z_2 + ... + Z_n) \tag{3}$$

- Rp – levelling depth, distance between the highest peak and the reference line



The micrographs of the samples used for gathering topography information via image analysis procedures were made by a JEOL JSM 6460 LV scanning electron microscope (SEM). To assure the uniform electrical properties and to avoid the charging/discharging of the aluminium oxide surfaces, the printing plate samples were gold coated by ion sputtering (thickness 15.0 nm, density 19.32 g/cm^3). SEM recording parameters are presented in Table 2.

Table 2
SEM recording parameters

Working distance	Voltage	Tilt angles	Magnification	Image size
15 mm	20kV	0 ±5 deg.	500x	128 μm x 96μm

After acquiring the SEM images of the printing plates, ImageJ 1.44 software was used to analyze the images and to calculate the relevant profilometric parameters by texture analysis as explained below. These parameters were subsequently compared with the ones obtained from TR 200 contact profilometer.

2.1.1 Texture Analysis with GLCM

The roughness of a substrate can also be investigated using different texture analysis tools [17]. In this study we focused on the parameters derived from the grey level co-occurrence matrices (GLCM) that were computed from the corresponding images using ImageJ *GLCM Texture* plugin. Image texture is a substrate's appearance related term which can be regarded as a descriptor of local brightness variation from pixel to pixel in a small neighbourhood through an image [18]. Digital images can represent various texture attributes, such as graininess, periodicity, directionality and also smoothness/roughness. As for the latter, topography measurements are usually performed with an either mechanical or laser profilometric device generating range images – 2D images whose pixel values correspond to the distance to points in a scene from a specific point [19]. Our idea was to use SEM images of printing plates to see whether image intensity variations correspond to the expected smoothening of the plates due to the friction/wearing. One of the most frequently used texture analysis methods is based on the computation of GLCM. GLCM is a matrix that keeps track of how often different combinations – pairs – of pixel intensity (gray level) values in a specific spatial relationship and distance occur in an image. From this matrix it is possible to compute various first and second order statistical parameters or texture measures. Details of this procedure can be found in [20]. Each measure describes one aspect of the image texture and does not necessarily correlate to the other measures.

From each of the four SEM images (1280 x 960 pixels) of the cyan printing plate surface – the reference and plates after 123,000, 177,000 and 300,000 impressions, respectively – three smaller 500 x 500 pxl 8-bit grayscale images were obtained. They were after performing Gaussian filtering (r = 1 pxl) subject to image processing in ImageJ. Its plugin GLCM Texture can generate four different texture measures: angular second moment, contrast, inverse difference moment and entropy. Their values depend on the way the GLCM is calculated. In our case, these settings were as follows: step size: 1 pxl, step direction: 0 deg, meaning that the calculations were based on the horizontally adjacent image pixels.

2.2 Contact Angle Measurements

The contact angle measurements of the applied fountain solution (H_2O) on the non-printing areas of lithographic printing plate were performed by a Dataphysics SCA 20 instrument. The computer-controlled unit can operate in several modes: Spinning drop, Needle in, and Sessile drop method. The latter one was used in our investigations [21]. Accurate drop position and volume – precision up to 0.1 µl – are possible. Recording measurement with CCD camera makes it possible to determine static or dynamic contact angles [22]. In our study the liquid drop volume was 1.5 µl. In Fig. 3 one can see two significant phases of the contact angle measurement.

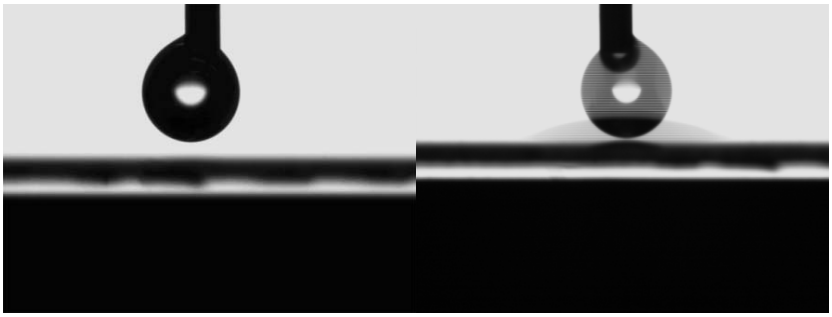


Figure 3
a) Drop formation; b) First contact between a liquid and a solid surface

The unit is controlled thorough a computer program, OCA SCA20, which in addition to measuring contact angle enables calculation of surface free energy of a substrate providing dispersive and polar components. The program also provides the possibility to use various fitting curves (circle, ellipse, tangent and Laplace-Young fitting) depending on the shape of the drop. Fig. 4 shows a software screenshot during the measurement.

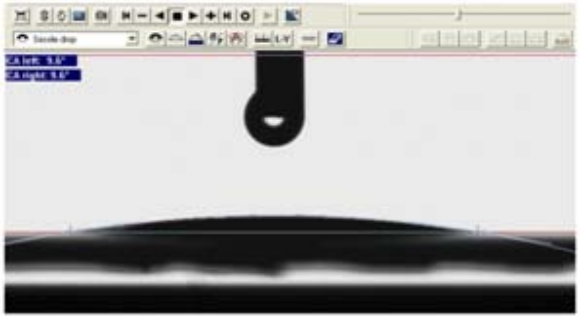


Figure 4
Measurement of contact angle by means ofSCA 20 software

The contact angle was also determined on the basis of images captured by a video-based optical angle measurement device followed by an ImageJ implementation of the DropSnake plugin developed by the Biomedical Imaging Group, École polytechnique fédérale de Lausanne – EPFL [23]. The plugin uses a general method to measure contact angle and is suited for non-asymmetric or general drops. The contact angle is obtained by a piecewise polynomial fit on B-spline snakes (active contours). The drop reflection may be used to detect the interface and a small tilt in the image [24, 25]. The images of the three drops were assessed for each plate. All images (768x574 pixels) were first transformed to8-bit ones and saved as uncompressed TIFF files. On each drop 10 to 12 knots were manually added, starting at the left interface point and following the drop contour until the right interface point (Figure 5). When the last knot was placed, the spline was closed and the contact angles separately for the left and right drop side were calculated. Finally the average value for all three drops for both left and right contact angle was calculated.

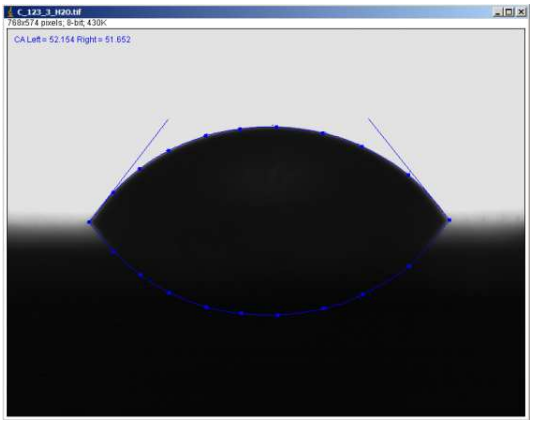


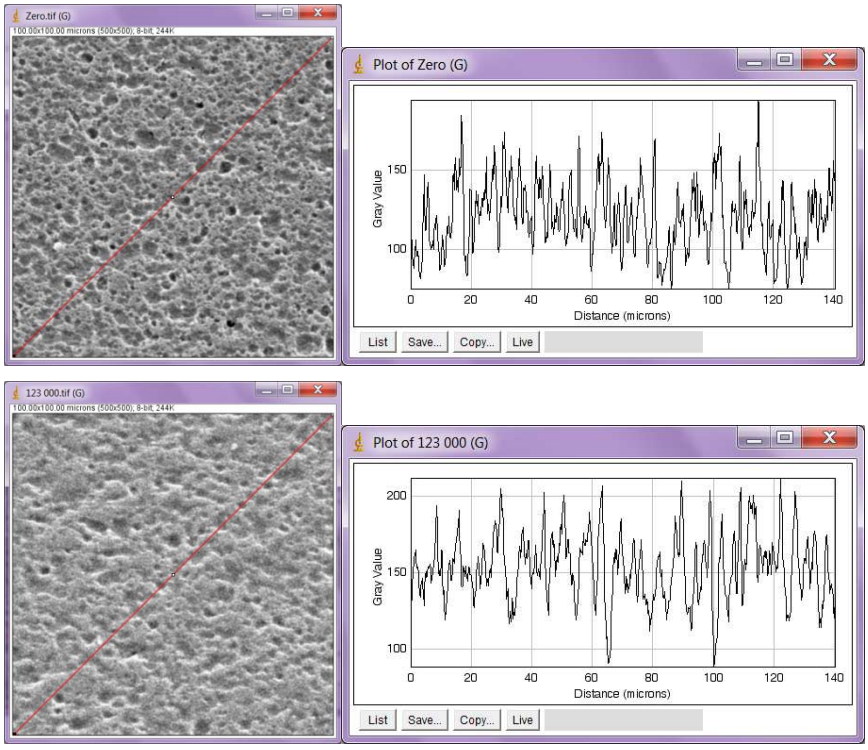
Figure 5
DropSnake measurement

3 Results and Discussion

The values of standard roughness descriptors obtained by the TR 200 measuring instrument together with GLCM derived texture measures for the investigated samples/images are displayed in Table 3. Fig. 6 shows their 1D intensity profiles along the depicted diagonal lines.

Table 3
Roughness descriptors and GLCM based texture measures for investigated sample images

No. of print runs	Standard topography descriptors (μm)				GLCM-based texture measures			
	Ra	Rq	Rp	Rz	Angular Second Moment	Contrast	Inverse Difference Moment	Entropy
0	0.39	0.52	1.17	3.23	3.79E-04	93.848	0.165	8.196
123000	0.38	0.48	1.04	2.87	4.51E-04	96.778	0.171	8.014
177000	0.33	0.42	0.90	2.40	5.51E-04	86.812	0.198	7.897
300000	0.30	0.39	0.81	2.28	5.49E-04	88.006	0.18	7.821



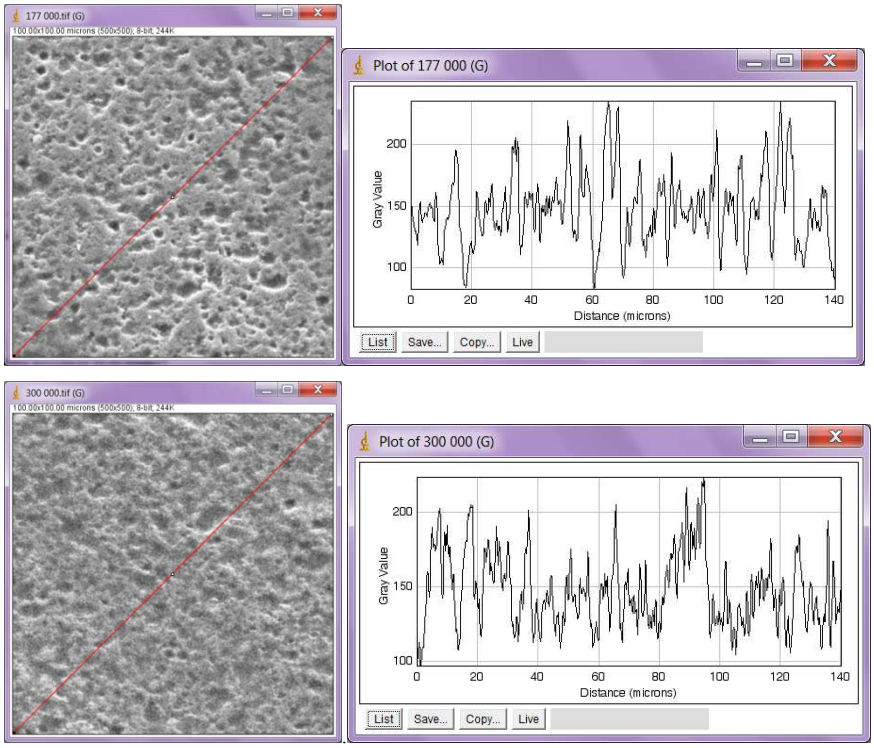


Figure 6

Linear intensity profiles of SEM images

Of the four GLCM-based texture measures, the angular second moment, inverse difference moment and entropy show good correlation with the values obtained by the stylus-based topography assessment method: with an increasing number of impressions, the values of the first two parameters monotonically increase – with a single exception of the 300,000 impressions' plate – while the value of the third parameter decreases. As angular second moment corresponds to the uniformity of pixel values in the grayscale image and inverse difference moment to their similarity and since entropy is a measure of disorder or randomness, these findings are in agreement with an increasing smoothness of the plate surface. The contact angle was determined with a DataPhysics OCA30 instrument and, alternatively, using digital images processed by ImageJ plugin DropSnake. The results of both methods are displayed in Table 4.

Table 4
Results of contact angle measurements obtained by OCA 30 and ImageJ plugin DropSnake

No. of print runs	Optical angle measurement DataPhysics OCA30	ImageJ - DropSnake
	Contact angle [°]	Contact angle [°]
0	48.85 ± 0.86	49.67 ± 1.94
123000	61.87 ± 0.64	61.79 ± 0.30
177000	64.55 ± 0.55	62.98 ± 3.21
300000	67.32 ± 0.30	68.88 ± 2.74

From the results presented in Table 4, we can see that with a growing number of impressions, the contact angle increases, thus influencing the wettability of the investigated offset aluminium plates. This trend of increasing contact angle is most probably due to the applied printing pressure. With larger print runs the non-printing areas of the plate are exposed to a stronger continuous pressure and the pre-roughed surfaces are smoothed, leading to higher contact angles. This change of the surface properties from rougher to smoother surfaces can be observed through the ISO surface roughness parameters and through some of the investigated image analysis parameters. With a decrease in the surface roughness there is a decrease in the surface free energy, which directly influences the amount and angle of the fountain solution. This effect can have negative results on the print quality due to paper stretching or lower ink saturation, and therefore should be controlled during the print run. A comparison of the average contact angle values shows that the two methods produce very similar results. The difference lies in the precision of the methods, which is evidently higher for the OCA 30 instrument.

Conclusions

By comparing different methods – analytical instrumental and image processing based – to characterize surface roughness and wettability of offset printing plates, we have demonstrated the possibilities of quality parameters quantification. The wearing of the aluminium plate surfaces caused by paper dust, pressure and other factors lead to the degradation of several parameters, mostly surface roughness and contact angle, thus reducing the quality of the printed product. Of the four surface descriptors derived from the image analysis-based surface assessment method (GLCM), three of them – angular second moment, inverse difference moment and entropy – were found to correlate well with the values obtained with the stylus profilometer. The fourth descriptor – contrast – did not exhibit linear relationships, possibly because the SEM technique – unlike e.g. confocal laser scanning microscopy (CLSM) – does not generate actual surface images. Both methods for contact angle determination yielded very similar results.

The presented findings indicate that ImageJ can be used as a convenient tool for the inspection of the surface roughness and contact angle assessment. Further investigations will be conducted to test a larger number of important quality

parameters. This can lead to an easier quality control via machine vision and image processing systems that have a potential to replace conventional analytical methods which are sometimes time consuming and cannot be as easily automated.

Acknowledgement

On the Serbian side, the work on this paper was supported by the Serbian Ministry of Science and Technological Development, Grant No.: 35027 "The development of software model for improvement of knowledge and production in graphic arts industry".

References

[1] Pahk H. J., Stout K., Blunt L., A Comparative Study on the Three-Dimensional Surface Topography for the Polished Surface of Femoral Head, Int. J. Adv. Manuf. Technol. 16 (2000) 564-570

[2] Wieland M., Hanggi P., Hotz W., Textor M., Keller B. A., Spencer N. D., Wavelengthdependentmeasurement and Evaluation of Surface Topographies: Application of a New Concept ofwindow Roughness and Surface Transfer Function, Wear 237 (2000) 231-252

[3] Brinkman, H. J., Kernig, B., 2003. *ATB* Metallurgie, Aluminium for Lithographic Applications, R&D Hydro Aluminium,. *Vermeersch* et al., *43(*1-2), pp. *130*-135

[4] Atkins, P. W., (1998) Physical Chemistry, 6th Ed., Oxford University Press

[5] Hiemenz, P. C., Rajagopalan R., (1997) Principles of Colloid and Surface Chemistry, 3rd Ed., Marcel Dekker, New York, ISBN 0-8247-9397-8

[6] ISO 12218:1997, Graphic technology – Process control – Offset plate making; ISO 12647-2:2004. Graphic technology – Process control for the production of halftone color separations, proof and production prints. Part 2. Offset lithographic processes

[7] Hutchinson R (2001) Trans Inst Met Finish 79:B57

[8] Rivett B., Koroleva EV., Garcia-Garcia FJ, Armstrong J, Thompson GE, Skeldon P (2011) Wear 270: 204-217

[9] Noble J. W. III, Leidheiser H. Jr (1981) Ind Eng Chem Prod Res Dev 20(2):344–350. DOI: 10.1021/i300002a022

[10] Li J, Papadopoulos C, Xu JM, Moskovits M (1999) Appl Phys Lett 75:367-369

[11] Tae-Yong K, Jeong SK (2008) Korean J Chem Eng 25(3):609-611

[12] Dong Hyuk Park, Mikyung Kim, Mi Suk Kim, Dae-Chul Kim, Hugeun Song, Jeongyong Kim, and Jinsoo Jooa,; Electrochemical Synthesis and Nanoscale Photoluminescence of Poly(3-butylthiophene) Nanowire; Electrochemical and Solid-State Letters, 11 (7) K69-K72 (2008)

[13] Mahovic Poljacek, S., Gojo, M., Raos, P., Stoic, A., Different Approach to the Aluminium Oxide Topography Characterisation, 10th ESAFORM Conference on Material Forming, In: *AIP* Conference Proceedings Vol. 907, Zaragoza (2007) 64-69

[14] Mahovic Poljacek, S., Risovic, D., Furic, K., Gojo, M., 2008. Comparison of Fractal and profilometric Methods for Surface Topography Characterization, Applied Surface Sciense, 254 (11) pp. 3449-3458

[15] ISO 4287:1997 Geometric Product Specification (GPS). Surface texture: profile method—terms, definitions and surface texture parameters

[16] ISO 4288:1996 Geometric Product Specification (GPS). Surface texture: profile method—rules and procedures for the assessment of surface texture

[17] Singh, S. P. (2008): Paper smoothness evaluation methods. Bioresources 3(2), 503

[18] Russ, J. C. 1999. The Image Processing Handbook, 3rd edition. CRC Press, Florida

[19] Hladnik A., Lazar M.: Paper and Board Surface Roughness Characterization using Laser Profilometry and Gray Level Cooccurrence Matrix. Nordic Pulp and Paper Research Journal, 26(1), 2011, 99-105

[20] Hall-Beyer, M.: The GLCM Tutorial. Web: http://www.fp.ucalgary.ca/mhallbey/tutorial.htm (Accessed on 10 July 2011)

[21] Lander L. M., Siewierski L. M., Brittain W. J., Vogler E. A., "A Systematic Comparison of Contact Angle Methods", Langmuir (1993) pp. 2237-2239

[22] Data Physics Instruments GmbH, Operating manual OCA, 2006

[23] http://bigwww.epfl.ch/demo/dropanalysis/

[24] Stalder A. F., Melchior T., Müller M., Sage D., Blu T., Unser M., "Low-Bond Axisymmetric Drop Shape Analysis for Surface Tension and Contact Angle Measurements of Sessile Drops," Colloids and Surfaces A: Physicochemical and Engineering Aspects, Vol. 364, No. 1-3, pp. 72-81, July 20, 2010

[25] Stalder A. F., Kulik G., Sage D., Barbieri L., Hoffmann P., "A Snake-Based Approach to Accurate Determination of Both Contact Points and Contact Angles," Colloids and Surfaces A: Physicochemical and Engineering Aspects, Vol. 286, No. 1-3, pp. 92-103, September 2006

Measuring Hungarian and Slovakian Students' IT Skills and Programming Knowledge

Gábor Kiss

Óbuda University, Budapest, Hungary, kiss.gabor@bgk.uni-obuda.hu

Abstract: An analysis of Information Technology knowledge of Hungarian and Slovakian students was made using a web based Informatics Test. After the evaluation of the test results, there were found some significant differences in IT skills and programming knowledge of the students from different countries, but these differences do not depend on a different way of teaching. In the following statistical analysis Levene's test, T-test and Z-test was used. The monitoring was held on the p=5% significance level throughout the analysis. The underlying causes are discussed. Survey results are traced back to differences in the educational systems of the two countries.

Keywords: comparative analysis; measuring; knowledge level; IT skills; programming; Hungary; Slovakia

1 Introduction

The aim of this research is to analyze the efficiency of different teaching methods in Information Technology education in Slovakia and Hungary. Some research was done comparing the school systems [1] and the role of Information and Communication Technology in the education of some European countries [2] [3] [4] [5] [6]. The goal of this research is an analysis of the IT skills and programming knowledge of students from different countries.

The National Basic Curriculum of Hungary describes to teachers the learning material grade by grade and subject.

The National Educational Program of Slovakia does not assign precisely what teachers must teach in the various grades but announces the standards to be reached at the end of the senior section; the aim is to reach preset school leaving standards.

The Education System [7] and the Information Technology education in Slovakia bears a close resemblance to that of Hungary from the point of view how various topics are discussed [8]. Theoretical knowledge, word processing, spreadsheet calculation, database management and programming are parts of the curriculum in both countries.

The topics are the same, but the way they are taught is different. This research analyses whether differences in the IT skills of students depends on the methods with which they are taught or not.

In order to compare students' IT knowledge in different countries, a detailed analysis was needed: checking the various curricula of different grades, the number of weekly Informatics classes and whether Information Technology was compulsory or only an optional subject. Still, it was not enough to carry out the examination.

It was also necessary to check the students' knowledge in various grades in both countries. To make comparisons, a uniform questionnaire was built with questions on different subject matters of Information Technology. Only after sending the questionnaire to the students of both countries could the survey be carried out on the basis of their answers.

In Table I the number of weekly Informatics classes are shown in different grades of the respective countries. Note that Informatics is a selectable course only in the last two years of secondary school in Hungary.

Table I

Number of Information Technology classes in the different grades

Country	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Hungary	0,5	0,5	0,5	0,5	2	2	2	2	2	2	3*	3*
Slovakia	0	1	1	1	0,5	0,5	0,5	0,5	0,5	0,5*	0,5*	0,5*

* selectable

Before making any comparison, the first starting hypothesis was that Hungarian students have better IT skills than their Slovakian peers because more Informatics classes are from the 5th grade in Hungary

The second hypothesis was that Hungarian students choosing Information Technology as an optional subject have better programming skills than their Slovakian peers, because the accent is on programming in this level in Hungary.

2 Information Technology Teaching in Hungary

IT education is based on a national curriculum in Hungary [9]. According to the National Basic Curriculum (NBC) of Hungary, the use of IT is to be demonstrated in the first four school grades since 2003 (e.g. search on the Internet, painting with computers, etc.) and is taught in 1 class weekly. According to the Information Technology curriculum, the following subjects are taught from the 5th grade to the 12th grade at the schools of Hungary in 2 classes weekly:

- Word processing
- Spreadsheet calculation
- Presentation
- Algorithm and programming
- Database management

Generally the Microsoft Office suite is taught, and it can be seen that teaching Word processing takes 4 years in Hungary (Table II). Basic algorithms, or rather programming, appears in Information Technology sooner, but recursion, list and tree data structures are only an elective part of the curriculum. Database management begins in the 9th grade. In grades 11-12, CS is just optional. At basic level it is taught 2 hours weekly, on a higher level 3 hours weekly, and a final exam can be taken.

Table II
The subjects of IT by grades in Hungary

Subject	**Grade**							
	5	6	7	8	9	10	11	12
Word processing		✔	✔	✔	✔			
Spreadsheet calculation				✔	✔			
Presentation						✔		
Algorithm and programming			✔	✔	✔			
Database management					✔			

3 Teaching Information Technology in Slovakia

The subject Informatics has been compulsory from the 2nd grade since the school year of 2008/2009 (at the time of the introduction of school reform), but it also appears in the 1st grade, as well as in the nursery school curriculum, though not as a compulsory subject. One Informatics class weekly is compulsory in the junior section and 1/2 class weekly in the senior section, which can be raised by the schools' own program. Some schools took the opportunity and increased it to 1 class weekly. The National Educational Program does not assign precisely what teachers must teach in the various grades but announces the standards to be reached at the end of the senior section. So it does not matter if programming is taught in the 6th grade in one school while in the 8th grade in another; the aim is to reach the school leaving standards. It is part of the own educational program of each school how students should reach these standards, how many classes they should have weekly and at what pace they learn the material; this program is accepted by the management and the teachers of the school together.

The National Educational Program divides Informatics into 5 topics (not specify ing the number of classes):

1 Information around us
2 Communication by the means of the ICT
3 Problem solving, thinking with the help of algorithms
4 Basic principles of the operation of ICT tools
5 IS society

These 5 topics are then to be fitted into the school leaving standards.

4 The Method of the Comparison

It is possible to compare the Hungarian and Slovakian Information Technology education via examining the students' knowledge. Since it is quite difficult to send out questionnaires physically to the various schools, and the order of the questions cannot be changed in that case, and also because students sitting close to each other in the classroom can see the other's answers, the most effective solution seemed to be a web-based Informatics test.

Some research was done towards developing the test in order to standardize the IT education in Switzerland [10] and Austria [11].

Questions of varying difficulty were formed on the main topics of Information Technology in this research. The topics chosen were part of Informatics education in almost every country: theoretical knowledge, word processing, spreadsheet calculation, database management and programming. There can be significant deviations in the curricula of some countries; therefore the test was expanded with questions on cryptographical knowledge as well as formal languages and automats since in certain German provinces these are also part of the Information Technology curriculum [12].

The database structure for the test had to be planned in a suitable way so that the data could be obtained later on. The personal data of the students filling in the test were put in a separate table as well, as were their answers to the questions. When filling in the test, the students first had to give their actual grade and some other data. If students gave the username of their teacher then the teacher also could see how they succeeded and would get a feedback on their progress. *Grade* was important because the student would get a question sheet depending on the grade given. Students could mark topics not taught to them (except basic information technology and office packages). If they marked one, the system would not ask questions dealing with the topic but saved it with the answer "I have never learned that". With this option students got fewer questions, and answers would flow in at a quicker pace. Next, students could begin to fill in the test.

Every test question has 6 possible answers, only one of which is correct, 3 of which are bad; the 5th choice is: "I have never learned that", and the 6th: "I have forgotten it". The answers "I have never learned that" and "I have forgotten it" show which part of the curriculum the students have not learned in that grade and if they could remember it or not. Every question has two time limits given in seconds. The first is the minimum time to read, understand and answer the question; the second is the maximum answering time. The software saves the total time used by the student. These time limits are not seen or known by the students. These are used during the evaluation, so a correct answer is accepted only if it arrived in the available time interval. Teachers can register on this site too if they are willing to give some of their data. The system is protected by registration code, and registered teachers can log in with their username and password. If a student filling in the test also gives the username of the teacher, then the teacher can later see his/her answers and the results. Some reports can be generated, helping the work of the teacher. The test records whether the students have given the teacher the right to inspect. It also indicates if they have marked a question as not learned or if the topic of the question is familiar to them but they have forgotten the right answer.

The evaluation of the answers is only possible after processing the saved data. The first step is to check whether the students of the given country have learned the given topic. If they have not, the comparison with the data of the students of other countries is impossible to make.

If the students knew the topic because they had learned it, it had to be checked if the time spent answering the questions was within the limits given. If so, the answer could be accepted as right.

The mean and the standard deviation of the right answers had to be calculated in the various grades and countries and comparisons made with the help of statistical means. In order to be able to do this, enough students filling in the test were needed in each grade. When comparing the IT skills of students in two countries, the Independent Samples T-test was taken.

5 Number of Participiants

Students filled in the web-test from the 5th grade in Hungary and in Slovakia but so few of the Slovakian students did that in the 6th and 7th grades that the comparison could not be made with them. It was possible only in the 5th and the 8th grades, as well as in the first three years of the secondary school, since the number of Slovakian students in the 12th grade was low (Table III).

Table III
Number of participiants

Grade	Hungarian		Slovakian
	Basic education	Informatics course	Basic education
5	79	0	126
6	14	0	114
7	18	0	108
8	169	0	50
9	552	0	111
10	302	0	97
11	104	69	102
12	212	91	21

6 Survey Results

The web-test on Informatics was filled in by altogether 729 students from 22 Slovak cities. After the examination of the number of filled in tests, it looked possible to compare the Informatics knowledge of the Hungarian and Slovakian students in the 5th and 8th grades and in the first 3 grades of secondary school.

Let us look first at the results of students learning basic Informatics (Table IV).

6.1 Result by Subjects

The following table shows the results by countries and subject (Tables IV and V). The mean shows how many questions the students could answer, the next column shows the ratio in percentage and the following one shows the standard deviations.

Table IV
The mean and the standard deviation of the right answers of the Hungarian and the Slovakian students

Grade	Subject	Nationality of students	Mean	percent of correct answers	Std. Deviation
5	Theoretical knowledge	Hungarian	3,04	75,95%	1,16
		Slovakian	1,91	47,67%	1,29
5	Word processing	Hungarian	2,46	49,11%	1,49
		Slovakian	1,21	24,19%	1,08
5	Spreadsheet calculation	Hungarian	0,95	47,47%	0,81
		Slovakian	0,40	19,77%	0,49
8	Theoretical knowledge	Hungarian	9,20	19,99%	4,94
		Slovakian	7,53	16,37%	5,61

8	Word processing	Hungarian	4,33	30,90%	2,15
		Slovakian	3,41	24,37%	1,87
8	Spreadsheet calculation	Hungarian	2,54	13,36%	2,11
		Slovakian	1,76	9,29%	2,02

Table V

The mean and the standard deviation of the right answers of the Hungarian and the Slovakian grammar school students

Grade	**Subject**	**Nationality of students**	**Mean**	**percent of the correct answers**	**Std. Deviation**
9	Theoretical knowledge	Hungarian	10,87	23,63%	6,16
		Slovakian	8,61	18,71%	6,44
9	Word processing	Hungarian	5,46	38,99%	2,66
		Slovakian	3,11	22,18%	3,64
9	Spreadsheet calculation	Hungarian	2,18	11,50%	1,94
		Slovakian	2,84	14,96%	2,85
9	Database management	Hungarian	0,63	3,47%	1,68
		Slovakian	0,16	0,88%	0,72
9	SQL	Hungarian	0,02	0,25%	0,32
		Slovakian	0,05	0,66%	0,32
9	Programming	Hungarian	0,00	0,00%	0,00
		Slovakian	1,42	5,68%	2,92
10	Theoretical knowledge	Hungarian	10,61	23,07%	6,25
		Slovakian	9,39	20,42%	5,22
10	Word processing	Hungarian	4,96	35,43%	2,76
		Slovakian	3,67	26,19%	2,53
10	Spreadsheet calculation	Hungarian	2,63	13,82%	1,94
		Slovakian	2,58	13,56%	2,31
10	Database management	Hungarian	1,20	6,68%	2,24
		Slovakian	0,18	1,01%	1,04
10	SQL	Hungarian	0,06	0,79%	0,51
		Slovakian	0,00	0,00%	0,00
10	Programming	Hungarian	0,00	0,00%	0,00
		Slovakian	2,12	8,16%	3,79
11	Theoretical knowledge	Hungarian	11,82	25,69%	6,31
		Slovakian	11,23	24,41%	4,68
11	Word processing	Hungarian	5,88	41,96%	2,88
		Slovakian	3,43	24,49%	2,15
11	Spreadsheet calculation	Hungarian	3,27	17,21%	1,88
		Slovakian	4,17	21,95%	2,32
11	Database management	Hungarian	3,17	17,63%	2,73
		Slovakian	1,14	6,35%	2,33

11	SQL	Hungarian	0,21	2,64%	1,36
		Slovakian	0,00	0,00%	0,00
11	Programming	Hungarian	0,00	0,00%	0,00
		Slovakian	4,69	18,02%	3,88

The data in the table show that Hungarian students in the 5th grade performed better since they gave the right answers for more questions. This seems to change in the 8th grade. The topic of database management is not included in the table because no students in this grade learned it in either country.

The Hungarians seemed to be better at word processing in the 9th and 10th grades. Hungarian students do not learn database management until they start their secondary education, though according to the curriculum they should already in the 8th grade. This topic is taught in the secondary school, but in the 9th grade only 10%, in the 10th grade only one third, and in the 11th grade two thirds of the students told so; and most of the good answers (17.6%) were also given by them.

Approximately 5% of the Slovakian students learned database management in the 8th grade, 20% of them in the first part of secondary education and half of them in the 11th grade, but they could not give as many right answers as their Hungarian peers, who did not perform outstandingly either.

The topic of SQL is not discussed in either of the countries, though it should be a significant part of database management. Hungarian students start learning database management in the 10th grade, but few gave correct answers to the 18 questions.

Apparently Slovakian students get to know the topic of programming in the 9th grade and go on with it in the 10th grade, but only a few proper answers were given to the questions. Their Hungarian peers do not learn programming at all despite the regulations of the National Core Curriculum that makes it compulsory from the 7th grade. Hungarian students achieved better results in word processing throughout the test and gave proper answers to approximately half of the questions, while Slovakians knew the right answer only in a quarter of the questions. Hungarian students do not learn programming at all, only if they choose Informatics as an optional subject in the second half of the secondary school years. The National Core Curriculum assigns programming to students from the 7th grade in vain. In Slovakia, one third of the students in the 8th grade have already learned programming, and at least half of them have met the topic of algorithmical thinking by the end of their secondary school years. Their accomplishment is not outstanding in the first half of their secondary school education since they knew the correct answer to only 8% of the questions, but a significant improvement can be observed in the 11th grade. Here they scored 18%.

In many cases it is not enough to examine the results in percentage in order to compare the students' achievements. It can only be stated unambiguously after the statistical analysis regarding in which grades and in which topics are significant differences between the students from different countries.

6.2 Analysis of the Means by Subjects

The next step in the analysis was to inspect whether the means by subject would differ if using the Independent samples test. The null hypothesis was that no significant difference would exist between the means of all subjects by countries. Because of having two independent samples, it was possible to use the T-test to decide whether the hypothesis was true or not (Table VI). If the analysis of the results (*Levene test*) showed the variance of the two groups different ($p<0.05$) [13], in this case the means could be compared with *Welch's t test* ($p<0.05$) [14] otherwise, the means could be compared with *T-test* ($p<0.05$) [15].

Table VI
Independent sample test of the Hungarian and the Slovakian students

Grade	Subject	Levene's test for Equality of variances		T-test for equality of means		means are different
		F	Sig.	t	Sig. (2-tailed	
5	Theoretical knowledge	1,52	0,22	4,95	0,00	yes
5	Word processing	7,39	0,01	5,30	0,00	yes
5	Spreadsheet calculation	0,03	0,86	3,15	0,00	yes
8	Theoretical knowledge	0,30	0,58	1,31	0,19	no
8	Word processing	0,52	0,47	1,69	0,09	no
8	Spreadsheet calculation	2,49	0,12	2,21	0,03	yes
9	Theoretical knowledge	1,95	0,16	2,18	0,03	yes
9	Word processing	3,19	0,07	5,13	0,00	yes
9	Spreadsheet calculation	1,10	0,30	0,10	0,92	no
9	Database management	11,70	0,00	3,42	0,00	yes
9	SQL	1,44	0,23	-0,62	0,54	no
9	Programming	645,44	0,00	-3,00	0,00	yes
10	Theoretical knowledge	0,02	0,90	1,08	0,28	no
10	Word processing	0,03	0,86	2,58	0,01	yes
10	Spreadsheet calculation	0,10	0,75	1,40	0,17	no
10	Database management	26,87	0,00	4,58	0,00	yes
10	SQL	410,24	0,00	2,12	0,03	yes
10	Programming	2,02	0,16	-9,85	0,00	yes

11	Theoretical knowledge	1,42	0,24	0,51	0,61	no
11	Word processing	4,15	0,04	4,61	0,00	yes
11	Spreadsheet calculation	1,45	0,23	-0,58	-0,33	no
11	Database management	1,53	0,22	3,94	0,00	yes
11	SQL	3,34	0,07	0,92	0,36	no
11	Programming	340,64	0,00	-12,41	0,00	yes

The table above contains the results of the Independent Samples T-test, and in the last column it can be seen if there are any significant differences between the knowledge of the Hungarian and the Slovakian students in the respective grades concerning the given topics. On the basis of this, it could be confirmed that Hungarian students in the 5th grade were better in all of the three topics. In the 8th grade they achieved better results only in spreadsheet calculation. In the 9th grade they scored higher marks than their Slovakian peers in both theoretical knowledge and word processing. The differences in database management show that Slovakian students had not learned the subject yet, while in programming it is the other way around; the lack of knowledge of the Hungarian students causes the differences. In the 10th grade the Hungarians achieved better results in word processing; they had already learned database management while their Slovak peers had not learned it yet. The Hungarians had not really learned SQL and this makes a difference. Slovakian students go on learning programming while Hungarian students have to give it up in the basic Informatics training. The 11th grade was the last one under examination. The Hungarians were far better at word processing than their Slovakian peers; more and more of them have learned database management while in Slovakia this topic has been left out of the education. The Hungarians did not learn programming in this grade either while the Slovakian students developed their knowledge in this field.

7 Comparing the Knowledge of Informatics of Regular Slovakian Students with those of Hungarians Specialized in Informatics

In the previous section it was seen that Hungarian students did not learn programming in the basic Informatics classes; they just started to learn database management and did not really get involved in it, and instead, their teachers taught them word processing and spreadsheet calculation. Now let us see the scores of the students who have chosen Informatics as an optional subject and who still learned Informatics in the last two years of secondary school. Since the Slovakian students in the 11th grade were the ones who filled in the test in an adequate number, we can only take this grade into consideration.

7.1 Result by Subjects

The following table shows the results of the proper answers of the Slovakian students and Hungarian students having chosen Informatics as an optional subject in the 11th grade on the basis of topics (Table VI).

Table VII

The mean and the standard deviation of the right answers of the Hungarian students having chosen Informatics as an optional subject and Slovakians in the 11th grade

Grade	Subject	Nationality of students	Mean	percent of the correct answers	Std. Deviation
11	Theoretical knowledge	Hungarian	13,43	29,21%	7,61
		Slovakian	11,23	24,41%	4,68
11	Word processing	Hungarian	5,48	39,13%	3,30
		Slovakian	3,43	24,49%	2,15
11	Spreadsheet calculation	Hungarian	3,16	16,63%	1,88
		Slovakian	4,17	21,95%	2,32
11	Database management	Hungarian	3,09	17,15%	2,29
		Slovakian	1,14	6,35%	2,33
11	SQL	Hungarian	0,45	5,62%	1,55
		Slovakian	0,00	0,00%	0,00
11	Programming	Hungarian	7,09	27,26%	4,67
		Slovakian	4,69	18,02%	3,88

Looking at table VII you see that Hungarian students specialized in Informatics have better skills in database management and programming taught in the 11th grade than Slovakian students, while Slovakian students achieve better results in spreadsheet calculation. But the results given in percentage are not enough; to examine the difference in skills of students from different countries a deeper analysis is needed (Levete's test, T-test).

7.2 Analysis of the Means by Subjects

The following table shows the results of the Independent Samples T-test of the Slovakian students and Hungarian students specialized in Informatics in the 11th grade on the basis of topics (Table VIII)

This table shows that the statements made on the basis of the previous percentile values are confirmed; the differences are also verified by the statistical examination.

Table VIII
Independent sample test of the Slovakian students and Hungarian students specialized in Informatics in the 11th grade

Grade	Subject	Levene's test for Equality of variances		T-test for equality of means		means are different
		F	Sig.	t	Sig. (2-tailed	
11	Theoretical knowledge	3,61	0,06	1,57	0,12	no
11	Word processing	6,66	0,01	3,33	0,00	yes
11	Spreadsheet calculation	0,00	0,95	-2,12	0,04	yes
11	Database management	4,18	0,04	3,66	0,00	yes
11	SQL	12,50	0,00	1,71	0,02	yes
11	Programming	0,44	0,51	2,62	0,01	yes

Hungarian students specialized in Informatics do better in the 11th grade in database management than the Slovakian students, as students getting only basic education. This is not surprising since these students have gained very similar basic knowledge in the first half of secondary education.

The Slovakian students achieve better results in spreadsheet calculation as we have also seen in the case of students learning basic Informatics. SQL is still a neglected area; Hungarian students can hardly give any correct answers, but since Slovakian students do not learn it at all, the difference between the two countries can easily be detected.

It is in the field of programming that we can see the advantages of being on an Informatics course compared to getting a basic Informatics education. Here teachers obviously have the time to help the students know this area. That is why the Hungarian students have scored higher points than their Slovakian peers.

8 Comparing Programming Knowledge of Regular Slovakian Students with that of Hungarian Students Specialized in Informatics

Hungarian students do not learn programming in a basic Informatics course [16, 17]. This means only scores of those students can be analyzed who have chosen Informatics as an optional subject and learn Informatics in the last two years of secondary school [18]. Since Slovakian students in the 11th grade were the ones who filled in the test in adequate numbers, only this grade could be considered in the following analysis.

8.1 Differences by Nationality in Programming

It seems clear from this survey that the results of the independent sample test of regular Slovakian students as well as their Hungarian counterparts specialized in Informatics in the 11th grade show different means of programming marks (Table VIII). Hungarian students scored more correct answers (~27%) than their Slovakian peers (~18%) (Table VII). A deeper analysis of the different programming topics was needed to make a decision about the second hypothesis, as it was in the case when knowledge by genders was discussed in Hungarian grammar schools [19] and higher education [20].

The following table is a break-down by programming topics showing what percent of students in this grade answered correctly the questions put to them (Table IX).

In order to know whether the means by Programming topic are equal, a Z-test [21] was accomplished.

The null hypothesis of Z-test was that no significant difference existed between the means by nationality. The monitoring was held on the p=5% significance level. The critical value of Z-test was between -1.96 and 1.96 at p=5% significance level. If the calculated value of Z-test was in this range, the null hypothesis could be kept. The table shows the calculated values of the Z-test by nationality and the decision on keeping the null hypothesis or not.

Table IX

How many percent of students answered successfully the questions - grouping by nationality

	Hungarian		**Slovakian**			
Programming topic	**%**	**Std. dev.**	**%**	**Std. dev.**	**Value of Z-test**	**Decision**
Flowchart	35,2%	0,48	28,6%	0,46	-2,74	The means are not equal
Structogram	22,0%	0,42	11,4%	0,32	-1,42	The means are equal
FOR cycle	43,4%	0,50	27,1%	0,45	-2,67	The means are not equal
Repeat-Until cycle	19,2%	0,40	14,3%	0,35	-1,78	The means are equal
Do-While cycle	26,4%	0,44	12,9%	0,34	-1,49	The means are equal
Parameter passing	16,5%	0,37	15,7%	0,37	-1,97	The means are not equal
Sort algorithm	14,3%	0,37	24,3%	0,43	-2,80	The means are not equal
Array management	9,9%	0,30	17,1%	0,38	-2,25	The means are not equal
Subroutine	42,9%	0,50	23,8%	0,43	-2,24	The means are not equal
Stack management	12,1%	0,33	14,3%	0,36	-1,94	The means are equal
Binary three knowledge	6,0%	0,24	17,1%	0,44	-2,13	The means are not equal

List knowledge	19,2%	0,40	27,5%	0,41	-3,19	The means are not equal
Recursion	15,0%	0,48	8,6%	0,28	-1,09	The means are equal
Binary search algorithm	11,0%	0,31	20,0%	0,41	-2,52	The means are not equal
The Eight Queens Problem	25,0%	0,28	14,3%	0,36	-2,07	The means are not equal

According to the table it could be asserted at p=5% significance level that Hungarian and Slovakian students were not on the same knowledge level regarding different programming topics in the 11th grade (Table IX).

Hungarian students answered the FOR cycle question more successfully, but no difference could be found in answers dealing with Repeat-Until and Do-While cycles. The questions of the Sort and the binary search algorithm proved easier to answer for Slovakian students in connection with array management. On the other hand, Hungarian students were more experienced in subroutine and backtracking algorithms.

More Slovakian students gave correct answers than Hungarians in the topics related to dynamic memory management (binary trees, lists). It looks like Hungarian and Slovakian students learn the programming part of Information Technology in a different way. Slovakian teachers spend more time on array management, sort, search algorithm and dynamic memory management. The emphasis is on subroutine and backtracking algorithms in Hungary. While Slovakian students acquire a well-founded basic knowledge, Hungarian counterparts rather learn high level programming topics (like backtracking algorithms). My opinion is that a deeper basic knowledge for Hungarian students should have a priority over high level programming topics (which can be taught later anyhow).

Conclusions

The first starting hypothesis according to which the Hungarian students have better IT skills than their Slovakian peers was only partly justified. Topics were found not taught in either of the countries. Examining the efficiency of teaching Informatics in Slovakia and in Hungary it can be said that in the beginning Hungarian students do better concerning theoretical knowledge and they receive a better basic education, but later this advantage disappears. Hungarian students performed better in word processing during the whole test; teachers seem to put the emphasis on this topic in our country. The starting advantage in spreadsheet calculation disappears by the end of the secondary school; Slovak students provided the same results. Hungarian students are more likely to have learned database management then Slovakians but a lot later than assigned in the curriculum and efficiency is not satisfactory. As for programming, it is just the other way around. Slovakian students learn algorithms already in the 8th grade; their Hungarian peers do not meet this topic until they finish secondary school, in

spite of the fact that students should start dealing with it in the 7th grade according to the regulations of the National Core Curriculum. If specialized in Informatics there is enough time to learn it in the second half of secondary school.

The second starting hypothesis was that the Hungarian students specialized in Information Technology would reach higher scores in programming than Slovakian students. This assumption turned out to be correct in the 11th grade. Students choosing this get to know the beauties of programming and produce higher scores than the Slovakians. Hungarian students were more experienced in subroutine and backtracking algorithms, but more Slovakian students gave correct answers in the topics related to dynamic memory management.

Now a conclusion can be made: the IT skills level of students does not depend on the different way of teaching. Teachers in Slovakia have freedom of action as to how to order the topics, but outgoing standards are firmly set by the National Educational Program. On the other hand, teachers in Hungary have to stick strictly the National Basic Curriculum grade by grade, subject by subject. However one can see what really counts is the time and efforts invested by the teachers.

References

[1] Döbert, H., Hörner, W.: Die Schulsysteme Europas. Hohengehren, 2002

[2] Europäische Kommission: Benchmarking Access and Use of ICT in European Schools 2006, Final Report. Bonn, Juni 2006

[3] Dagienė, V., Mittermeir, R.: Information Technologies at School. Vilnius, 2006

[4] Tóth Péter (2002): Az információs és kommunikációs technológiák szerepének vizsgálata néhány európai ország oktatási rendszerében I. Nagy-Britannia. In: Kadocsa, L. – Ludik, P., szerk.: „Multimédia az oktatásban" konferencia kiadványában, Dunaújváros, 2002, pp. 215-225

[5] Tóth Péter (2002): Az információs és kommunikációs technológiák szerepének vizsgálata néhány európai ország oktatási rendszerében II. Olaszország. In: Computer Panoráma XIII. évf. 2002/12 CD mellékletén: „Multimédia az oktatásban"

[6] Schweizerische Fachstelle für Informationstechnologien im Bildungswesen (SFIB): ICT und Bildung in der Schweiz. Bern, 2004

[7] Organizácia vzdelávacieho systému na Slovensku 2009/2010

[8] Vyhlášky MŠ SR č. 282/2009 Z. z. o stredných školách

[9] A Kormány 243/2003. (XII.17.) Kormányrendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról

[10] Bucher, P., Wirthensohn, M. – Test Your IT-Knowledge, Expertenbericht ICTStandardentwicklung. Zürich, 2004

[11] Peter Micheuz - Auf dem Weg zu Standards. Artikel in LOG IN, Heft 135. Berlin, 2005

[12] Gábor Kiss - The Concept to Measure and Compare Students Knowledge Level in Computer Science in Germany and in Hungary, in Acta Polytechnica Hungarica, Volume 5, pp. 145-158, 2008, ISSN: 1785-8860

[13] Levene, Howard (1960). "Robust tests for equality of variances". In Ingram Olkin, Harold Hotelling, et alia. Stanford University Press. pp. 278-292

[14] Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved, Biometrika 34 (1-2), 1947, pp. 28-35

[15] Nahalka István (1993): A változók rendszerének struktúrája. In: Falus Iván (szerk.): *Bevezetés a pedagógiai kutatás módszereibe*. Keraban Kiadó, Budapest

[16] Gábor Kiss - Measuring Student's Computer Science Knowledge at the End of the primary stage in Hungary / 9th IEEE International Symposium on Applied Machine Intelligence and Informatics, Smolenice, Slovakia, ISBN: 978-1-4244-7428-8, pp. 19-22, 2011, IEEE Catalog Number: CFP1108E-CDR, IEEE Xplore digital library Digital Object Identifier: 10.1109/SAMI.2011.5738880; Perspective Tudományos és Kulturális Folyóirat XV. évfolyam, különszám, 2011, ISSN 1454-9921, pp. 156-164

[17] Gábor Kiss - A Comparison of Informatics Skills by schooltypes in the 9-10th grades in Hungary, pp. 417-428 / International Journal of Advanced Research in computer science, Volume 2, No. 2, pp. 279-284, 2011, ISSN: 0976-5697

[18] Gábor Kiss - Measuring Computer Science knowledge at the end of secondary grammar school in Hungary, pp. 839-842 / 10th International Educational Technology Conference (IETC 2010), 2010, Istanbul

[19] Gábor Kiss - A Comparison of Programming Skills by Genders of Hungarian Grammar School Students / Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, Xi'An, China, 2010, ISBN: 978-0-7695-4272-0, pp 24-30, IEEE Catalog Number: CFP1075H-CDR, IEEE Xplore digital library Digital Object Identifier: 10.1109/UIC-ATC.2010.83

[20] Gábor Kiss **-** A Comparison of Informatics Skills by Genders when entering Higher Education in Hungary / 8th IEEE International Symposium on Intelligent System and Informatics, Subotica, Serbia, 2010, ISBN: 978-1-4244-7395-3, pp 179-182, IEEE Catalog Number: CFP1084C-CDR, IEEE Xplore digital library Digital Object Identifier: 10.1109/SISY.2010.5647280

[21] Korpás Attiláné dr. - Általános statisztika II. 95-99. old.

Applying Model-Driven Paradigm for the Improvement of Web Requirement Validation

Gustavo Aragon, M. J. Escalona

IWT2 Group, University of Seville
ETS Ingeniería Informática, Av. Reina Mercedes S/N, 41012 Seville, Spain
gustavo.aragon@iwt2.org; mjescalona@us.es

Jose R. Hilera, Luis Fernandez-Sanz

University of Alcala, ETS Ingeniería Informática, Campus Universitario
Crta. Barcelona KM. 33.6, 28871 Alcalá de Henares, Madrid, Spain
jose.hilera@uah.es; luis.fernandezs@uah.es

Sanjay Misra

Department of Computer Engineering, Faculty of Engineering, Atilim University
Kızılcaşar Mh., 06836 Incek, Ankara, Turkey, smisra@atilim.edu.tr

Abstract: This paper proposes an approach for Web requirements validation by applying the model-driven paradigm in classical requirements validation techniques. In particular, we present how the Navigational Development Techniques (NDT) approach exploits the model-driven paradigm to improve its requirements validation task by exploring tool cases that systematize or even automate the application of requirements validation techniques. Our solution is validated by applying it in a real industrial environment. The results and the learned lessons are presented accordingly.

Keywords: Requirement; validation techniques; Web-application; NDT; Model-driven paradigm

1 Introduction

The requirement phase is the most critical phase of the software development process. The requirements phase consists of several different types of activities, starting from the requirement elicitations to the validation and management of the requirements. Web-engineering follows the principles and concepts of software

engineering in developing the Web-applications. Further changes in requirements and short time lines [4] are inherent features of web applications. It makes the task of requirement engineering for Web engineering more difficult. Escalona and Koch [2] found that the requirement phase is poorly managed in Web engineering methodologies [16]. Although the observation [2] is approximately seven years old, today's situation is not significantly different.

In recent years, although some attempts have been made in developing requirement phases for web-applications, they still need some improvements [3]. Further, requirements must go through the validation process. The requirements validation is defined as the part of the software engineering activity where requirements are valued, analyzed and reviewed with end-users and clients [4], [5]. In this task, the development team should guarantee that requirements are correctly structured and defined and all of them are detected. This task is usually carried out by analysts, end-users, clients and the remaining team members, who work in conjunction to assure consistent requirements. A variety of proposals for requirements validation (in general) [7-16], for Web approaches [17-21] and for using testing as a validation technique [22-24] can be found in the literature. Further, reviews, audits or prototypes are some of the techniques commonly used for this task. However, they are difficult to apply due to development time constraints, communication problems or lack of suitable tools, among other reasons. As previously stated, in the Web Engineering environment the situation of requirement validation is even more complicated [3]. In Web Engineering, end-users and clients are usually unknown, and some characteristics of the Web environment, for instance complex navigation systems, complex interface, security aspects or a heterogeneous development team, significantly complicate the task. This paper sets out some solutions to Web requirements validation through a model-driven approach. In particular, this paper analyzes how the model-driven paradigm helps to reduce time and cost on Web requirements and illustrates this idea with the case of the Navigational Development Techniques (NDT) methodology [6]. NDT is an approach supporting the requirements and analysis phases in Web engineering using the model-driven paradigm. In the recent years, NDT have improved some aspects, mainly focused on the use of the model-driven paradigm, to support the development. In this line of action, this article analyzes how requirements validation can be improved and how NDT can be enriched to make more systematic this complex task. It deeply presents techniques supported by NDT and the solution offered in its case, as well as tools using the model-driven paradigm. It also outlines some conclusions obtained from practical experiences.

This article is organized in seven sections. Section 2 presents related work. It studies the importance of requirements validation and introduces the current situation. Section 3 provides a global vision of NDT and its evolution in the last years and Section 4 states how the application of model-driven Engineering (MDE) can improve validation requirements on NDT and how this methodology

exploits MDE principles to make easier this task. Then, in Section 5, the paper includes some relevant experiences when applying these solutions in the business context. Finally, Sections 6 and 7 summarize some conclusions obtained and present future work.

2 Related Work: Requirements Validation in Web Engineering

The validation of the requirements for a Web application is an important but not a common topic of research. It was a difficult for us to locate the relevant literature on this topic. As a consequence, we focus our search on more general topics, e.g. requirement validation, web applications, and web engineering.

In 2004, Escalona & Koch [2] presents a survey that analyzes how Web requirements are covered by Web approaches. They showed how Web approaches use classical techniques for requirements treatment. In requirements validation, they numbered four techniques: review or walk-throughs, audits, traceability matrixes and prototypes. In this study, ten Web approaches are analyzed and the paper presents how they use these techniques for requirements validation. If compared with requirements specification and capture, requirements validation is the less considered phase. Although this paper was written in 2004, the situation has not recently achieved a significant change. Some Web approaches, like WebML [17], have incorporated requirements phase in their life cycle. However, very few improvements for requirements validation were proposed. Robles et al. [18] propose mockups as techniques to represent requirements, since they help to report results to users. Dargham and Semaan [19] propose a requirement validation technique based on validation through visualization and animation to verify completion, correctness and consistency of Web navigations. This approach is mainly aimed at verifying navigational requirements. In [20], Garrigós et al. proposes the adaptation of the i* modeling framework [21], an approach for analyzing stakeholders' goals and how the intended system would meet them.

Another trend refers to requirements testing. Sommerville and other authors propose requirements testing as a validation technique. Recently, some Web approaches have included this tendency in their life cycles. Thus, WebML [22] includes BPMN (Business Process Management Notation) [230] as Computation Independent Model and proposes the systematic generation of test cases by means of model-driven paradigm. Robles et al. [24] are carrying out something similar. They are working to include testing requirements in their approaches. NDT, as it is presented in the following sections, also includes this possibility in the life cycle. From the previous paragraphs it can be summarized that although requirements validation is a very critical task in requirements engineering, it is poorly covered by Web approaches [16]. There is poor support of concrete techniques and tools,

even though techniques used in Web requirements validation are the same as those applied in classical approaches, e.g reviews, prototypes, traceability matrix, etc. Reviews and prototypes validation are considered "psychological" techniques [16] because they depend on the stakeholders' background and their objective point of view [9]. Development teams have to decide either if a guided or a free revision should be better. This aspect is more complex in the Web environment because end-users are frequently unknown, and assessing requirements with them is not possible. Further, the generation of prototypes, traceability matrixes and requirements testing are usually quite expensive for projects, and this cost is only assumed if duly justified. Besides, the maintenance of traceability matrixes and early tests could be also quite expensive if not supported by a tool case. On Web systems, this problem could becomes even worse since maintenance on Web systems is usually more complex than in classical projects: they must run 24 hours a day, 7 days a week, 365 days a year.

In fact, there are several commercial tools that contribute to the application of these techniques. For instance, IBM Rational Dynamic Object Oriented Requirements System (Doors) [25], HP Requirements Management [26], Blueprint Requirements Center [27], IRQ-A [28] and Polarion Requirements [29] are some examples of generic and commercial tools for requirements management. Each of these offers suitable solutions for the general management of requirements as well as for requirements validation: they execute reviews or support traceability matrices, among others. In addition, they offer a way to continue with the lifecycle. As an example, HP Requirements Management is integrated with the HP Application Lifecycle Management tools. IBM Rational Doors enables the generation of UML 2.0 models that can be exported to UML tool cases, and IRQ-a connection with Enterprise Architect [30].

However, these tools mark a higher distance between requirements and the remaining lifecycle. For instance, if we used IRQ-A or Doors for requirements management and, later, we exported them to another tool, a change in requirements would imply the development team would have to manage this change manually[1].

In addition, they do not offer a concrete solution for requirements validation in the Web environment. All of them offer some mechanisms to define and create different categories of requirements, but the result can be too general without a specific solution.

In conclusion, techniques for Web requirements validation seem to be clear, although their application must improve. Katasonov and Sakkinen [15] highlight that the main problem of requirements validation is communicating requirements because customers and end-users most likely do not have any technical expertise.

[1] In Section 5.1 of this paper there is a reference about this aspect in the Mosaico Project where, initially, Doors was used.

The next section explains how to apply these recommendations and the classical techniques in the Web environment followed by NDT. Furthermore, NDT shows the latest trend to use the model-driven paradigm for this aim and suggests some suitable tools to support the application of these techniques at a lower cost.

3 An Overview of NDT

NDT (Navigational Development Techniques) is a model-driven Web methodology that was initially defined to deal with requirements in Web development. NDT has evolved in the last years and offers a complete support for the whole life cycle. NDT is completely supported by a set of free tools, grouped in the NDT-Suite [31]. It selects a set of metamodels for each development phase. All concepts in every phase of NDT are metamodeled and formally related to other concepts by means of associations and/or OCL constraints [32].

In order to offer suitable support for NDT, we studied a set of different possibilities before starting to develop the NDT-Suite.

The first proposal was to use UML as the basis for NDT models. We defined a set of UML profiles to offer a suitable syntaxis for the use of the NDT metamodel. We selected UML profiles because, after some empirical studies, we realized that it was the easiest device for people in companies; on one hand, UML is commonly used in software companies, and thus development teams already know its notation. On the other hand, they usually work with UML-based tools.

After this consideration, we studied the possibility of developing our own tool for NDT or to use an existing UML based tool where we could define our UML profiles. In this sense, after studying some possibilities, we chose to use Enteprise Architect as the UML tool for NDT. The decision required a comparative study carried out together by our research group and the Andalusian Government. It determined the tool which offered the best position in price/quality ratio[2].

Another suitable possibility was to use Eclipse and EMF technologies [35]. However, as NDT is mainly oriented to requirements and end users' work, it does not offer as good interfaces as UML models, for instance, with use cases [36].

To conclude the sort presentation of NDT, we summarize the following points.

1 NDT is a MDE methodology that covers the whole life cycle. However, it is mainly focused on the requirements phase. In this phase, NDT offers a set of techniques to capture, define and validate requirements of different kinds.

[2] This study was written in Spanish. It was not published, but it can be consulted in www.iwt2.org.

These requirements are formally defined by a metamodel and they can be traced to the remaining artifacts of the life cycle by managing them in a suitable manner.

2 Despite its application in classical environments, NDT is developed in relation to the Web. In this sense, it supports special characteristics like navigation, complex interfaces or RIA [33]. In the requirements validation, NDT is oriented to cover classical techniques like traceability, prototypes, etc. but enriched to support these special web characteristics.

3 The degree of automation of NDT is one of its more relevant qualities. NDT is a theoretical approach, based on metamodels, transformations, etc. However, it is also an approach often used in companies[3].

4 Techniques and Tools for Requirements Validation in NDT

NDT supports different requirements validation techniques and offers several tools to automate their applications. This section presents the solution offered by NDT for requirements validation.

NDT supports the following requirements validation techniques:

- ***Requirements reviews:*** This technique checks requirements in detail. In reviews, clients and analysts need to check the consistency of requirements and whether they were correctly defined. Reviews are sometimes quite difficult to implement because they have a relevant psychological component [15]. Flaws in understanding requirements imply a false acceptance of requirements, which may lead to important errors in the system [2].
- ***Glossaries***: In several studies, such as [2], glossaries are not described as a validation technique. Nevertheless, NDT proposes using them in order to check the terminology consistency in requirements definition as an auxiliary technique. A glossary is a dictionary of terms for the system [5]. It is quite useful in systems with a heterogeneous development team to achieve lexical consistency during the process.
- ***Prototypes***: A prototype is partial implementation of a system that helps to carry out the performance and assessment of the future system with users. Prototypes are useful tools to work with users because they enhance interaction and ease communication. They can be classified in different ways, either upwards or downwards or according to high fidelity or low fidelity. All offer different possibilities to validate systems [36] [37].

[3] In www.iwt2.org in the Project section a detailed list of projects where NDT is used can be checked.

- ***Matrix of traceability***: This technique consists of the use of matrixes to establish correspondence among different artifacts in the system's development. In requirements, this technique allows, for instance, knowing how objectives are satisfied with a set of requirements [7].
- ***Requirements testing***: The requirements testing generates test cases to enable requirements testing [5]. Tests are not executed till the system is implemented. However, the early generation of tests validates the requirements definition by analyzing, in collaboration with end-users, functional paths that in the future should be tested. This technique is frequently named *early testing* [12].

NDT bears out all these techniques and offers tools to support their applications. As an introduction, Table 1 represents a matrix with each technique and the tool that supports it.

In the next section, each tool is presented in order to explain how they support every technique by means of the model-driven paradigm. Screens and examples offered in test introduction were obtained from an example, named Hotel Ambassador, which can be downloaded from [31]. It is an example fully developed with NDT-Profile to test our tools.

Table 1
Tool supplied for each technique

	Reviews	Glossaries	Prototypes	Traceability Matrix	Requirements Testing
NDT-Driver					X
NDT-Quality				X	
NDT-Glossary		X			
NDT-Report	X	X		X	X
NDT-Prototypes	X		X		
NDT-Checked	X				

4.1 NDT-Driver

NDT-Driver is a tool that supports each transformation defined in NDT. It implements transformations from requirements to analysis, analysis to design and requirements to test, as illustrated in Figure 1. The last transformation (Design to Code) is not executed by NDT-Driver in NDT, as it is supported by a plug-in of Enterprise Architect. They define a generation process based on QVT Transformations [34]. Figure 1 presents the idea more specifically.

NDT defines two transformations T1 and T2 in QVT Operational, based on the NDT functional requirements metamodel. T1 generates possible test scenarios from these functional requirements. The method used is the Path Analysis method

[39]. T2 is another transformation that, through the Category-Partition method [40], generates operational values for these test scenarios. A new transformation T3 is defined from both metamodels. T3 is thought to generate the test case metamodel. This process is fully presented in [41].

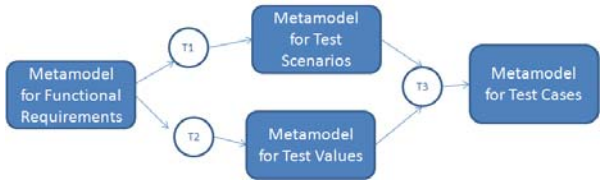


Figure 1
NDT- Driver process to generate requirements testing

In order to use this approach in NDT-Suite, three UML profiles for these new metamodels are defined and included in Enterprise Architect and within the NDT-Profile: test values, test scenarios and test cases profiles. Thus, a concrete syntax to define their test cases is offered. Transformations are translated in Java and they appear in the set of transformations of NDT-Driver.

Through NDT-Report, as the next sections state, the development team can generate functional test documents. They can execute two different scenarios with this approach, as shown in Table 2. Both scenarios start with the definition of functional requirements with NDT-Profile. NDT-Profile proposes an extension of use cases, activity diagrams and some specific patterns to represent them. After checking their quality by means of NDT-Quality, with the possibilities that are presented in next sections, scenarios offer two paths.

Table 2
Scenarios provided by using the NDT-Driver for functional test generation

Scenario 1	**Scenario 2**
1 The development team defines functional requirements in NDT-Profile.	1 The development team defines functional requirements in NDT-Profile.
2 They check requirements quality with NDT-Quality.	2 They check requirements quality with NDT-Quality.
3 They generate the requirements catalogue, using NDT-Report, and validate it with users.	3 They use NDT-Driver and execute Req2Test transformations to generate the functional test catalogue in NDT-Profile.
4 They use NDT-Driver and execute Req2Test transformations to generate the functional test catalogue in NDT-Profile.	4 They use NDT-Report to generate a printable version of the functional test catalogue.
5 They use NDT-Report to generate a printable version of the functional test catalogue.	5 They validate requirements with users through this functional test catalogue.
6 When the test phase is executed, they use this functional test catalogue generated in the requirements phase.	6 When the test phase is executed, they use this functional test catalogue generated in the requirements phase.

In Scenario 1, the development team validates functional requirements by means of other techniques (reviews, prototypes, etc.) supported by NDT with some alternative tools, as is presented in following sections. Once errors and mistakes

have been amended, the development team can execute Requirements to Test transformations in order to generate the functional test cases that will be used in the future when the system may be implemented and the test phase starts.

However, Scenario 2 presents an option oriented to validate requirements through testing. In this situation, once NDT-Quality has checked the requirements quality, functional test cases are generated before the official requirements validation takes place, and they are used to validate requirements with users.

This idea provides NDT with the possibility of validating requirements from tests based on the model-driven paradigm. As Section 5 presents, this event adds some important advantages in the enterprise environment as the process is automatic and offers a powerful mechanism to facilitate the communication with users. For them, analysis of a functional circuit modeled as a test case is frequently easy and clear.

Obviously, this possibility is only a part of the process because it only considers functional requirements. Now, we are working in delivering this test generation from other kinds of requirements, such as navigation requirements.

4.2 NDT-Quality

In NDT, the application of a set of transformations executed by NDT-Driver supports each step in the life cycle. However, NDT-Driver not only carries out these transformations, but goes further; it saves the relation between the source artifact and the target artifact when a transformation is executed. For instance, Figure 1 shows how NDT-Driver saves this relation when a test case X is generated from a functional requirements Y with the transformations Req2Test. The storage of this relation offers several advantages, mainly oriented to the system traceability and maintenance.

In the future, if the functional requirements Y changes, the test case X must be automatically reviewed, as it will probably change. The tool to check quality and traceability in the system is NDT-Quality. It controls three aspects in a system:

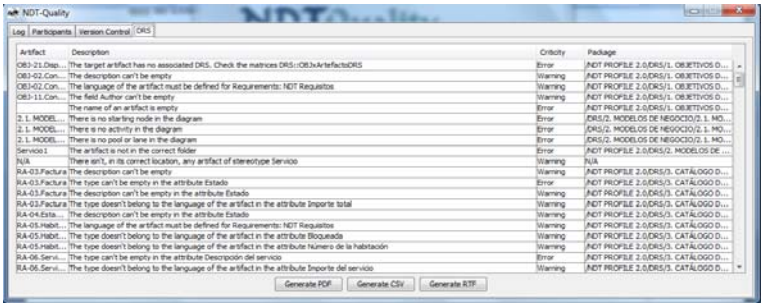
1 It implements a set of rules, defined by NDT as OCL constraints or invariants in their metamodels. Once a development team finishes each phase in the life cycle of NDT, they are expected to execute NDT-Quality to check that each rule or constraint defined by NDT is followed, e.g. it ensures that each requirement is defined by a unique identification code and a short description.

2 It implements a set of rules, defined by UML or general rules. It guarantees that an activity diagram is well-defined, e.g. without independent activities.

3 It implements a set of rules to check the traceability of the system. Following the previous example, it ensures that a change in requirements Y implies a review in test case Y.

NDT-Quality reports the detected errors and recommendations when executed. In Figure 2, Section a. presents the interface of NDT-Quality, which shows aspects that can be checked with NDT-Quality in different phases and traceability aspects whereas Section b. shows an example of a report.

a. Main interface of NDT-Quality



b. Example of report in NDT-Quality

Figure 2
Interface of NDT-Quality

In addition, NDT-Quality manages the traceability matrix of the system generated for relations created from transformations and NDT rules. Figure 3 offers an example of traceability matrix. It is automatically generated and manages how objectives are partially covered by functional requirements. NDT-Quality controls a high number of traceability matrixes: objectives-requirements, storage requirements-analysis classes, functional requirements or test cases, among others. All of them are automatically generated and can be required by the development team, if needed.

The automatic generation of a traceability matrix is a powerful tool in NDT derived from the application of the model-driven paradigm. This is one of the most frequent techniques to validate requirements and it is compulsory in relevant good practices and quality standards, like CMMi (Capability Maturity Model Integration) [42]. Section 4 explains its relevance in empirical experiences.

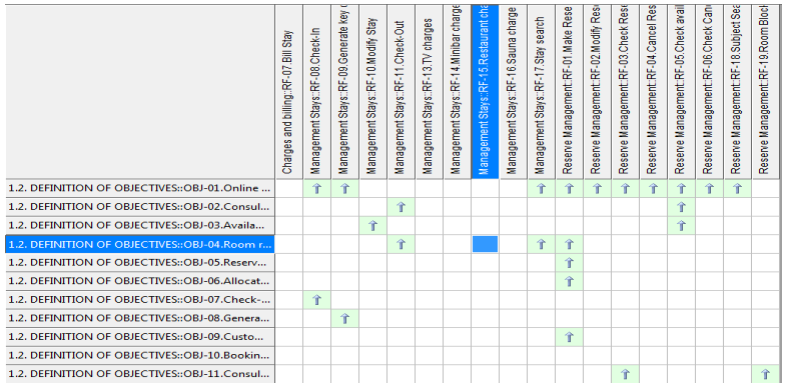


Figure 3
Example of traceability matrixes

4.3 NDT-Glossary

The glossary of terms in software projects allows the development team to store and exchange the knowledge acquired in the system domain. Basically, it is a dictionary that defines the most important concepts used during the development process of a software project. It is oriented to unify the vocabulary and control inconsistencies and ambiguities of concepts within the system domain in the life cycle. Every term is represented in the glossary as a couple, such as name and description, and through a set of relations with other terms; synonyms, related, etc. To keep the integrity of the glossary, each name must be unique. Glossary elaboration is not a requirements validation technique itself, but it results in quite an efficient way to find lexical inconsistencies during the requirements phase. Each glossary must verify two principles [8]: the Principle of Circularity and the Principle of Minimum Vocabulary. On one hand, the Principle of Circularity establishes that a glossary should be as self-content as possible. In this way, it ensures that all terms are related. At the same time every term and the relation with the remaining terms are included in the glossary. On the other hand, the Principle of Minimum Vocabulary states that requirements should be mainly expressed by concepts in the glossary, and thus it will be as understandable as possible.

In conclusion, engineers need to gather and define the most relevant and critical concepts for the system. Furthermore, a common language reduces the risk of misunderstandings and facilitates communication between users and analysts. NDT offers a tool in its suite, named NDT-Glossary, which uses model-driven paradigm to generate a glossary from the requirements model. Figure 4 represents this idea, where NDT defines a metamodel to represent a glossary, and later, a set of QVT transformations from the requirements metamodel to the glossary

metamodel is defined. Thus, these transformations start from the storage information requirements. These types of requirements in NDT define which information the system must manage and they are described with the user's vocabulary. This information is transformed to a glossary model. Both metamodels have an associated profile implemented in NDT-Profile whereas NDT-Glossary implements transformations in Java. Thus, a development team can get a first instance of the glossary from the requirements metamodel, or more precisely, from the storage information requirements metamodel.

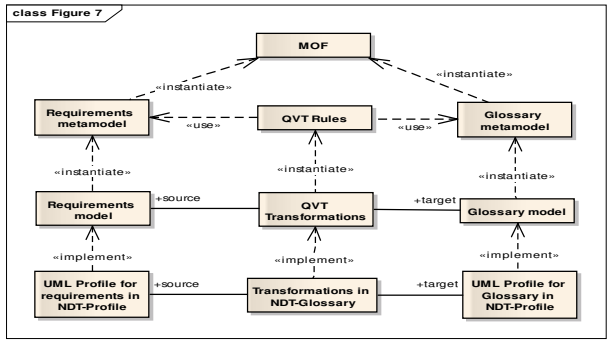


Figure 4
Glossary model transformation pattern

4.4 NDT-Report

When NDT started working in the enterprise environment, we noticed that it had an important bug. It offered a set of powerful tools like NDT-Driver or NDT-Quality quite oriented to make easier the use of the methodology for the development team. However, results given by NDT-Profile were not the best for the end-users' review. NDT-Profiles stores NDT models (mainly represented as UML models and patterns) whose interface is quite useful for development teams, but too complicated for end-users.

For this reason, a new tool named NDT-Report was developed. This tool defines a set of patterns to present NDT results to users. It can generate output in word files, pdf files or html files, as well as implement a set of transformations from NDT metamodels and generate the output to these patterns.

NDT-Report is essential for end-user participation in the NDT development cycle. It can prepare suitable outputs for each phase (requirements, analysis, testing, and so on) but it can also offer an external view of the traceability matrix of NDT-Quality and the glossaries generated by NDT-Glossary.

This tool is not a model-driven case tool. However, it offers suitable outputs to apply classical requirements validation techniques.

NDT-Report is the tool which carries out the revision in liaison with users. In fact, if a suitable and comprehensive output of the requirements is not offered, users cannot assess them [13] [14]. NDT-Report is the tool to get a printable version of requirements, glossaries, traceability matrices and functional tests.

4.5 NDT-Prototypes

Using prototypes to validate requirements is one of the most used techniques in the enterprise environment [43]. Prototypes normally assure that the end-user can easily understand the future system. It is also considered a very useful technique since it involves users in the requirements phase.

However, prototypes can have some disadvantages in use. They generally increase the development time. So they can delay the project due to the extra time needed for development. Normally, this time is paid back by the detection of errors and inconsistences in the first phases of the life cycle, which improves the final system quality. In addition, the granularity or the degree of development of the prototype may be a problem. If it is very detailed, users could consider that the prototype is the final version of the system, while if it is very general, it could not be relevant for evaluation [36].

In this sense, a new tool for the NDT-Suite was developed in order to get prototypes advantages and reduce the elaboration cost. This tool, named NDT-Prototypes, generates a set of prototypes from the requirements model of a system and uses the same ideas of NDT-Glossary. It implements a set of QVT transformations from the requirements model to the prototypes model in JAVA. However, in this case, the source is the interaction requirement. NDT supports interaction requirements presented in its metamodel as Visualization Prototypes. A Visualization Prototype instance represents how users process the information and how they can execute functional requirements and navigate through the system. This information described in the requirements model is transformed to a prototype model executing this set of transformations, which are implemented in Java as in the previous examples.

The interface of NDT-Prototype is quite simple and generates a set of html and css Websites. Figure 5 presents a screen generated with NDT-Prototype for the Ambassador example.

Each specific field is translated into a text field in the prototype. Depending on the values of their attributes (name and type) one specific user interface element is used. As an example, text boxes are used if type is “String”. Buttons offer the possibility of executing functional requirements. Thus, the model includes a relation between the visualization prototype “Create Reserve” and the functional requirements “carry out a reserve”. This functional aspect can be executed from the screen derived from “Create Reserve”.

Figure 5
Screen generated by NDT-Prototype from interaction requirements for the Hotel Ambassador

This model is not complete, as relations with activities, navigation and others are not presented in the figure. Nevertheless, it illustrates how NDT-Prototypes can help to understand the model. The use of prototypes is an essential technique for users' validation [44]. The full example is available in [31] in the Hotel Ambassador example.

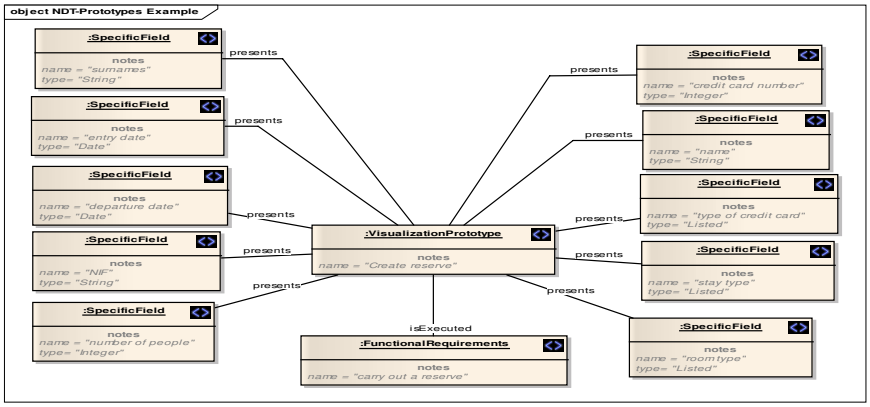


Figure 6
Original model for example in Figure 8

4.6 NDT-Checked

NDT-Checked and NDT-Report are the only tools in NDT-Suite that are not based on the model-driven paradigm. The NDT-Checked tool includes different sheets for each NDT product. These sheets give a set of check lists that should be manually reviewed with users in requirements reviews.

As it is presented in the Related Work section, requirements review is one of the most used techniques. It is essential for validation. However, it is sometimes difficult to consider which aspect must be reviewed with end-users [8]. There are

two main options: *free revision*, in which case the user reviews the requirements catalogue alone and freely; and *guided revision*, where the user reviews the requirements catalogue with the help of a development team member.

In NDT, the first one can be executed generating the requirements catalogue with NDT-Report and presenting the results to the user for revision. In the second option, NDT-Checked was developed. NDT-Checked, which is mainly based on enterprise experience, offers this set of checklists to assess and review each product generated in the life cycle of NDT.

5 Learned Lessons from Industry Experiences

In the last ten years, NDT and NDT-Suite were used in a high number of real projects. In fact, they are currently used in several projects carried out by different companies from public to private ones and from big to small ones.

Some specific projects have been selected in order to put forward some learned lessons from the empirical experience achieved when using the presented tools for requirements validation. All of them were developed with NDT, and its tools were used during their life cycles.

5.1 Projects for Cultural Heritage Management

The Andalusian Regional Cultural Ministry [45] has been applying NDT since 2004. Over the past several years, more than 90 projects of Web systems with different providers, users or development teams have worked with our approach. This experience is quite relevant, mainly in the use of NDT-Report and NDT-Quality. The Cultural Ministry does not accept any results or documents of a project if they are not checked with NDT-Quality.

The use of NDT-Report is also quite relevant. This set of users extends CSS and NDT-Report output design with their own patterns. In addition, they notice that when the same user participates in two or more projects, he or she can easily use this notation. In fact, Mosaico users [46] directly need NDT-Profile to validate its requirements. Mosaico is a big project that started in 2004, but it is continually being improved. These tools are essential to validate and trace requirements. In fact, at the beginning, for the requirements phase of Mosaico, the company that developed it suggested using a profile of Doors. However, in the first iteration, the development team noticed that the traceability of requirements with the rest of the life cycle was quite difficult, due to a disconnection between requirements tools and analysis (in this case, Enterprise Architect).This enhanced our interest in improving the traceability from requirements to analysis in NDT-Suite and, more specifically in NDT-Quality (see Figure 2a where NDT-Quality supports requirements-analysis traceability).

One of our future lines of work in this environment consists in applying NDT-Prototypes. This is one of our youngest tools, only applied in some projects. According to previous experiences in this environment, we conclude that requirements validation can be improved by adapting this tool to this environment.

5.2 AQUA-WS Project

The AQUA-WS (AQUA-WebServices) project [07] is a very important project carried out in Emasesa [08] over three years and finished in 2011. AQUA-WS is very relevant for the application of NDT-Driver in the test phase.

The AQUA-WS project included the development and implementation of an integrated business system for customer management, interventions in water distribution and clean-up, and management work or projects.

This project was launched when Emasesa needed to integrate the existing systems into a single one along and to upgrade the technological platform of the system. The existing systems were the customer management system (AQUA-SiC), network management system (AQUA-ReD) and the work and projects management system (AQUA-SigO). Thus, as the project was a technological migration of old systems, users only took part in the testing phase.

The project followed an iterative life cycle mainly based in RUP [09]. In each iteration, the development team, composed of more than 20 analysts from two companies, defined requirements, after studying the previous systems, and introduced them into NDT-Profile. Then, they were checked with NDT-Quality and NDT-Checker and, later, functional test cases were generated. NDT-Report had these functional test cases presented as functional paths reviewed with users.

The systematic way of generating test cases from functional offers a suitable and quite agile support for validating these functional requirements with users. However, a relevant conclusion is obtained from this experience. The quality and the suitability of derived test cases depend on the quality of the requirements. In some functional iterations, requirements have to be reviewed and written again, even before test generation, since they are poorly described.

5.3 Projects for e-Health Systems

NDT was also widely applied in the e-health environment. In 2006, Alcer Foundation [50] used it within the system to manage the degree of handicap for disabled people. In this project, NDT-Suite was not fully developed and we used a previous tool, named NDT-Tool [51]. However, this project is mentioned as it was the seed for detecting the need for NDT-Glossary. The medical systems environment works with very specific terminology: the project caused a high number of inconsistences only solved by elaborating a glossary manually.

Some years later, NDT-Suite was used in another e-health system, named Diraya [02], which is a very complex system. The requirements phase was developed by a group of six companies with a high number of analysts. Each company was expert in a concrete aspect of Diraya. The use of NDT-Profile and NDT-Glossary was essential to guarantee the unification of criteria in this multidisciplinary development team.

Conclusions and Future Work

This paper analyses the importance of requirements validation in Web Engineering. It presents an overview of today's situation in this research line and concludes with the need for offering systematic mechanisms to improve and even automatize this task. The article defends the idea of applying the model-driven paradigm to these aims. NDT-Suite is presented to validate this idea and more specifically, the tools that support requirements validation techniques.

We included some references to real projects which used these tools to support the requirements validation task.

As lessons learned from our experience with the model-driven paradigm in Web treatment, we could state that the use of this paradigm in this environment can improve the project results. However, development teams do not find the model-driven paradigm too intuitive in practical environments. The concepts of metamodels and transformations, among others, are not common notations for daily practice in industry, as they seem too abstract.

Nevertheless, we conclude that the use of UML profiles and UML-based tools offer an interface to deal with instances of metamodels suitable for analysts, designers and even for expert users.

In the same way, transformations in QVT do not appear easy to understand. However, our users do not work with QVT, but with a very easy interface, like the NDT-Quality interface in Figure 2a, to benefit the power of transformations.

As a summary, our experience has confirmed that requirements validation involves one of the most difficult and critical tasks. The project success heavily depends on the results of this phase; therefore, it is essential to manage it correctly. The lack of systematic or automatic techniques that help to support requirements validation represents an important limitation for software development. They frequently depend on psychological aspects and they are rarely based on tools. Our experience demonstrates that the model-driven paradigm can help to systematize and even automate the most classical requirements validation techniques. It offers cost and time reduction in this phase and helps to increase the quality of results.

The main advantage of our approach is that it uses a model-driven mechanism when offering the requirements output, which reduces the cost of their generation. The division into different types of requirements described in NDT metamodel

implies that each group of stakeholders works in an established group or sub-group: it reinforces the need of giving them an explicit task to accomplish requirements validation, if possible, based on systematic and guided techniques. In our approach, the development team can support several techniques to manage this task, e.g. reviewing a part of the glossary developed by NDT-Glossary; reviewing a set of requirements, such as a functional module in HTML obtained with NDT-Report; or reviewing the coverage and the requirements traceability with traceability matrixes generated by NDT-Quality. Once again, the cost reduction of a model-driven paradigm supports automatically these options and enables improving the reviews.

Finally, it is again easy to divide the review process into problems and small steps to be reviewed using this environment because outputs can be obtained in separate ways without added costs.

As an added value, although it is not the aim of this article, we want to remark that NDT is not only a requirements environment, but offers a connection with the remaining life cycle, even with other activities such as quality assurance or project management. In this sense, requirements information is available for connection with automated methods for test case generation [53]. The model-driven principles presented in this article for requirements validation can also be adapted to the rest of the methodology. Attending to our practical experiences, we can highlight that this paradigm might provide suitable results to companies using some abstract concepts like metamodels or transformations. In fact, we consider that this paradigm, apart from being used by the research community in the last years, is now starting to offer results and may become a very useful mechanism for building software, as well as for its maintenance or management.

Research work presented in this paper allows future development in different lines of work. We would like to add more tools to the NDT-Suite to develop Web systems. Currently, we are working on a new tool, named NDT-Counter, that it is oriented to estimate the cost of a Web system at the beginning of the life cycle. It is based on the model-driven paradigm and applies the Use Case Point technique to measure the development time of a system.

The environment introduced in this paper, which is used in different companies, is certificated under ISO 9001:2008 [54], UNE EN 166002 [05] and ISO 14001[06]. We are increasing the possibility to include processes to support some other standards like CMMi level 2 and ITIL v3 (Information Technology Infrastructure Library) [07]. NDTQ-Framework offers a set of processes to deal with these standards. Thus, if a company uses NDT and NDT-Suite while they want to pursue the certification under these standards, they could take NDTQ-Framework processes as reference. The tool does not only offer these processes, but also defines metrics, outputs and techniques useful for this goal. The implementation of metrics help to overcome limitations to their application in SME [58].

Acknowledgements

This research has been supported by the project Tempros project (TIN2010-20057-C03-02) of the Ministerio de Ciencia e Innovación, Spain and NDTQ-Framework project of the Junta de Andalucia, Spain (TIC-5789).

References

[1] Molina F., Toval A., Integrating Usability Requirements that Can Be Evaluated in Design Time into Model Driven Engineering of Web Information Systems. Advances in Engineering Software. Vol. 40, Issue 12, December 2009, pp. 1306-1317

[2] Escalona, M. J., Koch, N. Requirements Engineering for Web Applications: A Survey. Journal of Web Engineering, Vol. II, Nº2, pp. 193-212, 2004

[3] Aguilar, J. A., Garrigós, I., Mazón, J. N., Trujillo, J. Web Egineering Approaches for Requirements Analysis- A systematic Literature Review. Proceedings of WebIST 2010, pp. 187-190, 2010

[4] Pressman, R. S. Software Engineering. A practitioner's approach. Mc Graw Hill, 2004

[5] Sommerville, I. Software Engineering. Addisson Wesley, 9th Edition. 2010

[6] Escalona, M. J., Aragón, G. NDT: A Model-Driven Approach for Web requirements, IEEE Transactions on Software Engineering. Vol. 34, No. 3, pp. 370-390, 2008

[7] Bernárdez, B., Durán, A., Genero, M. Empirical Evaluation and Review of a Metrics-based Approach for Use Case Verification. Journal of Research and Practice in Information Technology. Vol. 36, No. 4, pp. 247-258, 2004

[8] Leite, J. C. S. P., Eliciting Requirements Using a Natural Language-based Approach: The Case of the Meeting Scheduler Problem. Monografias em Ciência da ComputaÇao. No. 13, 1993

[9] Leite, J. C. S. P., Requirements Validation through Viewpoint Resolution. IEEE Transaction on Software Engineering. Vol. 17, No. 12, pp. 1253-1269, 1991

[10] Silva, J. R., dos Santos, E. A., Applying Petri Nets to Requirements Validation. ABCM Symposium. Series in Mechatronics. Vol. 1, pp. 508-517, 2004

[11] Zhu, H., Jin, Lingzi, Diaper, D., Bai, G. Software Requirements Validation via Task Analysis. The Journal of System and Software. No. 61, pp. 145-169, 2002

[12] Escalona, M. J., Gutierrez, J. J., Mejías, M., Aragon, G., Ramos, I., Torres, J., Domínguez-Mayo, F. J. An Overview on Test Generation from Functional Requirements. Journal of Systems and Software. No. 84, pp. 1379-1393, 2011

[13] Gemino, A. Empirical Comparison of Animation and Narration in Requirements Validation. Requirements Engineering. No. 9, pp. 153-168, 2004

[14] Uchitel, S., Chatley, R., Kramer, J., Magee J. Fluent-based Animation: Exploiting the Relation between Goals and Scenarios for Requirements Validation. Requirements Engineering. 2004

[15] Katasonov, A., Shakkien, M. Requirements Quality Control. A Unifying Framework. Vol. 11, No. 1, pp. 42-57, 2006

[16] Sulehri, L. H. Comparative Selection of Requirements Validation Techniques Based on Industrial Survey. Department of Interaction and System Design. School of Engineering. Blekinge Institute of Technology. Master Thesis. December 2009, Ronneby, Sweden

[17] Brambilla, M., Butti, S., Fraternali, P. WebRatio BPM: A Tool for Designing and Deploying Business Processes on the Web. International Conference on Web Engineering. pp. 415-429 Web Austria 2010

[18] Robles, E., Garrigós, I., Manzón, J. N., Trujillo, J., Rossi, G. An i*-based Approach for Modeling and Testing Web Requirements. Journal on Web Engineering. Vol. 9, Nº4, pp. 302-326, 2010

[19] Dargham, J., Semaan, R. A Navigational Web Requirements Validation through Animation. Third International Conference on Internet and Web Applications and Services. pp. 211-216, Greece, 2008

[20] Garrigós, I., Mazón, J. N., Trujillo, J. A Requirements Analysis Approach for Using i* in Web Engineering. International Conference on Web Engineering. LNCS 5648, Spain, 2009

[21] Yu, E. Towards Modeling and Reasoning Support for Early-Phase Requirements Engineering. 3rd International Symposium on Requirements Engineering. pp. 226-235, 1997

[22] Fraternali, P., Tisi, M. Multi-Level Tests for Model Driven Web Applications. International Conference on Web Engineering. pp. 158-172, Austria, 2010

[23] Business Process Management Initiative. Available in www.bpmn.org. Accessed January 2012

[24] Robles, E., Grigera, J., Rossi, G. Bridging Test and Model-Driven Approaches in Web Engineering. International Conference on Web Engineering. pp. 136-150, Spain, 2009

[25] IBM Rational Doors. Available in www-01.ibm.com/software/awdtools/doors. Accessed in September 2012

[26] HP Requirements Management. Available in www8.hp.com/us/en/ software/software-solution.html?compURI=tcm:245-937050. Accessed in September 2012

[27] Blueprint Requirements Center. Available in www.blueprintsys.com/ products, Accessed in September 2012

[28] IRQ-A. Version 3.0. Available in www.visuresolutions.com. Accessed in September 2012

[29] Polarion Requirements. Available in www.polarion.com. Accessed in September 2012

[30] Enterprise Architect 9.0. Available in www.sparxsystems.com.au. Accessed in September 2012

[31] NDT-Suite. Available in www.iwt2.org. Accessed in September 2012

[32] Object Constraint Language. Available in www.omg.org/spec/OCL/2.2/. Release 2.2. 2010. Accessed in September 2012

[33] Robles, E., Escalona, M. J., Rossi, G. A Requirements Metamodel for Rich Internet Application. ICSoft 2010 Selected paper. Communications in Computer and Information Science. Springer Verlag. To be published. 2012

[34] Query/View/Transformation. Available in www.omg.org/spec/QVT/1.1/. Release 1.1. 2011. Accessed in September 2012

[35] EMF Technologies. Available in www.eclipse.org/modeling/emft/. Accessed in September 2012

[36] Valderas, P., Pelechano, V., Pastor, O. A Transformational Approach to Produce Web Application Prototypes from a Web Requirements Model. Vol. 3, Nº1, International Journal of Web Engineering and Technology. pp. 1476-1289, 2006

[37] Chavarriaga E., Macías J. A., A Model-driven Approach to Building Modern Semantic Web-based User Interfaces. Advances in Engineering Software. Vol. 40, N. 12, pp. 1329-1334, 2009

[38] Garcia-Garcia, J., Cutilla, C. R., Escalona, M. J., Alba, M., Torres, J. NDT-Driver, a Java Tool to Support QVT Transformations for NDT. The Twentieth International Conference on Information Systems Development (ISD) To be published. 2012

[39] Naresh, A. Testing From Use Cases Using Path Analysis Technique. International Conference on Software Testing Analysis & Review. 2002

[40] Ostrand, TJ, Balcer, MJ. Category-Partition Method. Communications of the ACM. 676-686. 1988

[41] Gutiérrez, J. J., Escalona, M. J., Mejías, M., Torres, J., Torres-Zenteno, A. H. A Case Study for Generating Test Cases from Use Cases. Proceedings of RCIS 2008, Morocco, pp. 223-228, 2008

[42] Capability Maturity Model Integration (CMMi). Available in www.sei.cmu.edu/cmmi/. Accessed in September 2012

[43] Kasser, J. A Prototype Tool for Improving the Wording of Requirements. 12th Annual International Symposium of the INCOSE, pp. 1-12, USA, 2002

[44] Mátrai, R., Kosztyán, Z. T. A New Method for the Caracterization of the Perspicuity of User Interfaces. Acta Polytechnica Hungarica. Journal of Applied Sciences, Vol. 9, N. 1, pp. 139-156, 2012

[44] Consejería de Cultura. www.juntadeandalucia.es/ccul. Accessed in September 2012

[45] Escalona, M. J., Aragón, G., Molina, A., Martinez-Force, E. A MDWE Methodological Environment for Culture Heritage. Technologies for Tourism Destination Management and Marketing: Tools and Trends. IGI Global. To be published. 2011

[46] Escalona, M. J., Gutiérrez, J. J., Rodríguez-Catalán, L., Guevara, A. Model-Driven in reverse. The Practical Experience of the AQUA Project. Euro American Conference on Telematics and Information Systems. pp. 90-95, Czech Republic, 2009

[47] Emasesa. www.aguasdesevilla.com. Accessed in September 2012

[48] RUP. Rational Unified Process. Available in www-01.ibm.com/software /awdtools/rup/. Accessed in September 2012

[49] Alcer. Federación Nacional de Asociaciones para la lucha contra las enfermedades renales. www.alcer.org. Accessed in September 2012

[50] Escalona, M. J., Aragón, G. NDT-Tool. A Model-Driven Tool to Deal with Web Requirements. ACM/IEE International Conference on Model Driven Engineering Languages and Systems. USA, 2007

[51] Escalona, M. J., Parra, C. L., Martín, F. M., Nieto, J., Llergó, A., Pérez P. A Practical Example for Model-Driven Web Engineering. Information System Development. Challenges in Practice, Theory and Education Springer Science + Business Media LCC, Vol. 1, pp. 157-168, 2008

[52] Escalona, M. J., Aragón, G., Molina, A., Martinez-Force, E. A Model-Driven Tool Framework for the Improvement in the Use of NDT. 8th International Conference on Software Quality Management. The British Computer Society. pp. 147-157, UK, 2010

[53] Fernandez-Sanz, L. and Misra, S., Practical Application of UML Activity Diagrams for the Generation of Test Cases, Proceedings of the Romanian Academy, Series A, Vol. 13, N. 3/2012, pp. 251-260

[54] ISO9001-2008. www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_ detail.htm?csnumber=46486. Accessed in September 2012

[55] UNE 166002. www.aenor.org. Accessed in September 2012

[56] 14001-2004. www.iso.org/iso/catalogue_detail?csnumber=31807. Accessed in September 2012

[57] Information Technology Infrastructure Library. ITIL Open Guide. Available in www.itil-officialsite.com. Accessed in September 2012

[58] Pusatli, O. T. and Misra, S., Software Measurement Activities in Small and Medium Enterprises: an Empirical Assessment, Acta Polytechnica Hungarica, Journal of Applied Sciences, Vol. 9, N. 1, pp. 139-156, 2012

Tribological Investigation of K Type Worm Gear Drives

Balázs Magyar[1], Bernd Sauer[1], Péter Horák[2]

[1] TU Kaiserslautern, Institute of Machine Element, Gears and Transmissions
Gottlieb-Daimler Str. Geb. 42, D-67661 Kaiserslautern, Germany
magyar@mv.uni-kl.de; sauer@mv.uni-kl.de

[2] Budapest University of Technology and Economics, Department of Machine- and Product Design, Műegyetem rkp. 3, H-1111 Budapest, Hungary
horak.peter@gt3.bme.hu

Abstract: This paper presents a calculation method to determine the locally changing tooth friction coefficient along each contact line based on the TEHD lubrication theory for worm gear drives operating in mixed lubrication conditions. This also involves a detailed presentation of a procedure to specify the proportions of boundary lubrication and hydrodynamic lubrication as well as temperature states in contact. By comparing the calculated tooth friction coefficients with experimental test results, it can be stated that the calculated results properly approximate the measurement results.

Keywords: worm gear drive; TEHD lubrication theory; mixed lubrication; numerical and experimental investigation; tooth friction coefficient

1 Introduction

There are different theories for the tribological investigation of gear teeth contact. Niemann [1] calculated the lubricant behaviour of worm gears according to the hydrodynamic lubrication theory. Wilkesmann [2] developed an algorithm to determine the speed and contact properties and losses of various worm gear drives of standard profiles. Predki [3] applied the elastohydrodynamic lubrication theory to evaluate the thickness of the lubricating film considering the deflection of the shaft. Bouché [4] considered the surface roughness and calculated the friction power loss of the worm gear teeth by assuming the mixed lubrication condition between the tooth surfaces. The aims of our investigation are to calculate the locally changing tooth friction coefficient of worm gear drives and to compare the calculated results with experimental tests. A knowledge of the tooth friction coefficient is necessary for further investigation into the dynamic behaviour of worm gear drives.

2 Geometry and Kinematics of the Investigated Drive

This study examines a worm gear set of a=100 mm axial distance, ZK profile, and i=40 ratio, with the worm shaft located at the bottom.

Knowledge of contact lines is required for determining the tribological properties of worm gear sets. Contact line points associated with the traverse positions of the worm shaft can be determined by a numerical solution of the equation of meshing. According to the equation of meshing, the normal vector (**n**) of the common surface is perpendicular to the relative velocity vector (**v**) of the bodies at the contact point [6]:

$$f\left(\theta,\zeta,\varphi_1\right)=\mathbf{n}\cdot\mathbf{v}=0 \tag{1}$$

Figure 1 shows instantaneous contact lines in case of the teeth parameters specified above, in 3 views, with different φ_1 angular position of the worm, in a stationary system of coordinates fixed to the casing [x_1, y_1, z_1].

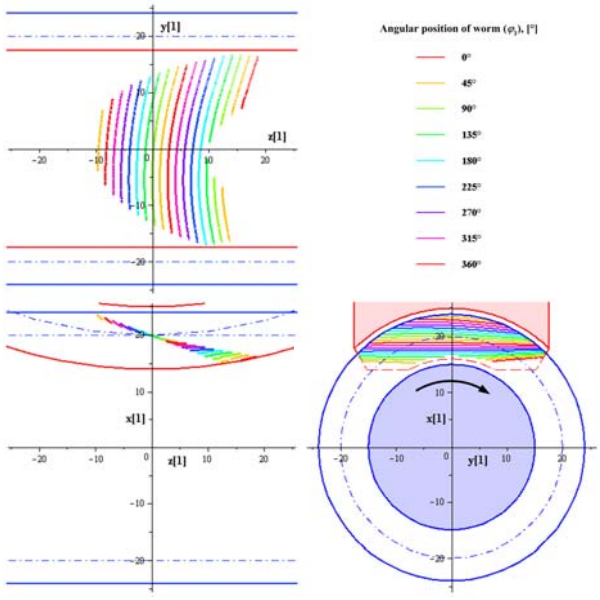


Figure 1
Contact lines calculated in a coordinate system fixed to the housing

Further tribological calculations require knowledge of speed and curvature properties at the contact line points. Speed conditions are determined in the manner described by Predki [3] and Bouché [4]. A kinematic procedure was developed by Litvin [6] to determine the contact point curvatures for spatial gears.

For the worm gear set examined, Figure 2 shows the course of $v_{sliding}$ sliding velocity and the course of v_{Σ} effective sum velocity along the contact lines at an input rpm of n_1=1500 1/min in the transverse section of the worm gear set.

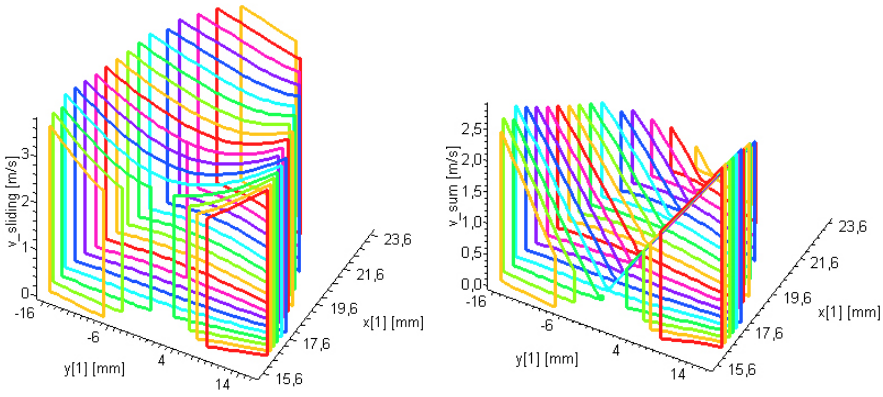


Figure 2

Changes in sliding velocity and in effective sum velocity along contact lines (transverse section)

As shown in Figure 2, sliding velocity does not change considerably along contact lines. On the contrary, there are significant changes in effective sum velocity. At entry (where the velocity of the worm shaft and the wheel are in the opposite direction) it assumes a large value. At the middle of the tooth, it is reduced to zero and it reaches its maximum at exit (here the velocity of the worm shaft and the wheel are in the same direction). Figure 3 shows the changes in ρ_E reduced radii of curvature.

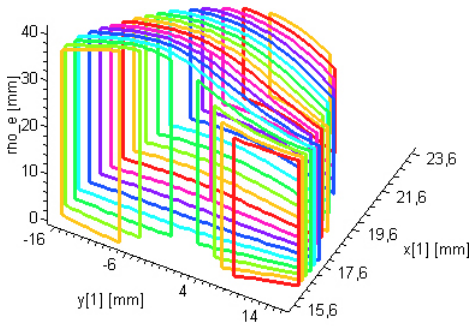


Figure 3

Changes in reduced radii of curvature along contact lines (transverse section)

3 Tribological Simulation

In the dimensions examined, the worm gear flanks operate in a mixed lubrication state [7]. This means that the asperities of contacting teeth are only partly separated by the film created. A precondition for determining the friction coefficient is to obtain the proportions of load distribution between the asperities and the film and the friction coefficient associated with boundary and hydrodynamic lubrication, respectively.

3.1 Model of the TEHD Lubrication of Worm Gear Teeth

In order to calculate these tribological properties, tooth flanks must be discretized. In the course of meshing, the worm shaft and the wheel are in contact with each other along the lines described above (see Figure 1). Between two contact points of a single contact line, tooth flanks can be substituted by rolls whose radius coincides with the reduced radius of curvature of tooth flanks, and the rolls perform rotational motion of the same or opposite direction at the velocity valid for the given contact point. This conceptual model is shown in Figure 4. This approximation, coming from Niemann [1], was successfully applied by Predki [3], as well as by Bouché [4].

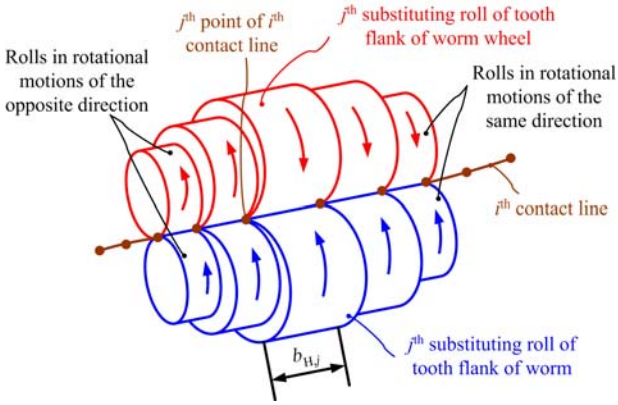


Figure 4
Substituting rolls for the tribological modelling of meshing

Based on the output torque, the height of the rolls, and the reduced radii of curvature, the load of each of the rolls and their deformations can be determined on the basis of Hertz's theory [4]. In the knowledge of these data, the EHD lubrication theory can be applied to substituting rolls for calculating the film thickness between tooth flanks. In our investigations, equal Stribeck's contact pressure is assumed to exist along the length of a roll. The pressure mound above

the flattening was approached according to Hertz, since the relatively low contact load-bearing capacity of the worm wheel made of bronze cannot produce pressures to give rise to a Petrushevich's peak and to narrow down the lubrication gap in the exit area. Accordingly, a lubrication gap of constant height was assumed for the entire contact area.

3.2 Determining Film Thickness between Substituting Rolls

Venner published a combined equation to determine minimum film thickness ($h_{\min}$) [8]. This equation yields accurate results in all four EHD ranges defined by Johnson [5]. The equation can be stated in the following form:

$$H_{\min} = \left[\left\{ \left(0,99 \cdot M^{-1/8} \cdot L^{3/4} \cdot t \right)^r + \left(2,05 \cdot M^{-1/5} \right)^r \right\}^{s/r} + \left(2,45 \cdot M^{-1} \right)^s \right]^{1/s}, \tag{2}$$

$$t = 1 - \exp\left\{ -3,5 \cdot M^{1/8} \cdot L^{-1/4} \right\},$$

$$r = \exp\left\{ 1 - 3/\left(L + 4 \right) \right\},$$

$$s = 3 - \exp\left\{ -1/\left(2 \cdot M \right) \right\}$$

where H, L and M are dimensionless EHD parameters. This equation yields the minimum film thickness under isothermic conditions. Murch and Wilson introduced a thermal correction factor (ϕ_ϑ) [9] for determining film thickness even in non-isothermic conditions. This correction factor depends on oil viscosity (η_0), the temperature dependence of oil viscosity (β), the thermal conductivity of oil (λ_ϑ), and effective sum velocity (v_Σ).

$$\phi_\vartheta = \frac{3,94}{3,94 + \left(\dfrac{\eta_0 \cdot \beta \cdot v_\Sigma^2}{\lambda_\vartheta} \right)^{0,62}} \tag{3}$$

In order to determine the central non-isothermic lubrication gap size (h_0), the $h_{\min}$ film thickness yielded by Venner's equation must be multiplied by 1.2 and the Murch and Wilson's correction factor [10]:

$$h_0 = 1,2 \cdot \phi_\vartheta \cdot h_{\min} \tag{4}$$

Figure 5 shows changes in the central non-isothermic film thickness thus determined for the worm gear drive examined, in the case of n_1=1500 1/min input rpm, T_2=570 Nm output torque, and FVA 4 mineral oil lubrication, at ϑ_S=60°C oil bath temperature. As can be seen, film thickness is affected by changes in effective sum velocity. Maximum film thickness is assumed at the end of contact lines; its value is reduced to zero at the middle.

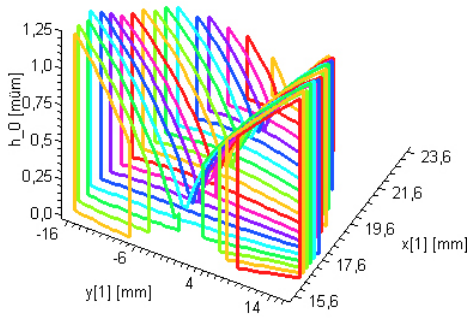


Figure 5
Changes in non-isothermic central film thickness along contact lines (transverse section)

3.3 Determining the Proportions of Friction Mechanisms

As a part of the present investigations, a calculation method was developed for determining the proportions of boundary lubrication and hydrodynamic lubrication. This requires knowledge of the roughness and material properties of contacting surfaces and of dimensionless film thickness.

The first model to describe rough surfaces – still frequently applied today – was developed by Greenwood and Williamson (the GW model) [11]. According to this model, asperities can be modelled by spherical caps of identical radius but stochastic height, and material properties are isotropic. Contact between two rough surfaces can be perceived as contact between a derived rough surface modelled by spherical caps and a smooth surface. The GW model also assumes that spherical caps do not affect adjacent spherical caps and that deformation is brought about only within single spherical caps (see Figure 6).

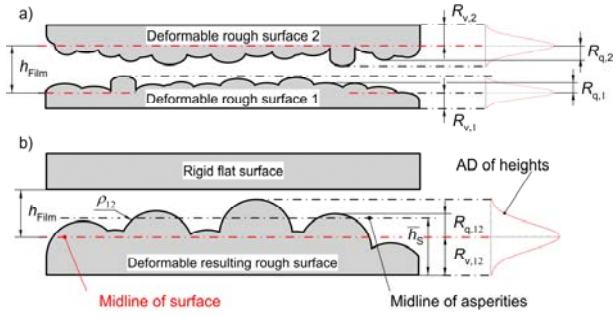


Figure 6
Generating derived surfaces from rough surfaces based on the GW model a) Two contacting rough surfaces, b) Contact model of a rigid flat surface and a deformable rough surface

The ρ_{12} radius of derived spherical caps can be specified by reduced radius generation according to the Hertzian theory:

$$\frac{1}{\rho_{12}} = \frac{1}{\rho_1} + \frac{1}{\rho_2} \tag{6}$$

The maximum valley depth ($R_{v,12}$) and rms-roughness ($R_{q,12}$) of the derivative surface can be calculated by the following correlations:

$$R_{v,12} = R_{v,1} + R_{v,2} \tag{7}$$

$$R_{q,12} = \sqrt{R_{q,1}^2 + R_{q,2}^2} \tag{8}$$

These values can be used for calculating the average height (h_S) and the standard deviation of heights (σ_S) of derivative spherical caps as follows:

$$\bar{h}_S = R_{v,12} + 0,82 \cdot R_{q,12} \tag{9}$$

$$\sigma_S = 0,71 \cdot R_{q,12} \tag{10}$$

Assuming that the height of asperities follows a Gaussian distribution, the probability density function of the derivative surface can be stated in the following form by using (9) and (10) [12]:

$$\varphi(h_S) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{h_S - \bar{h}_S}{\sigma_S} \right)^2 \right) \tag{11}$$

According to the GW model, contacts between each deformable spherical cap and the rigid flat surface can be examined separately. A spherical cap will first suffer elastic deformation then plastic deformation (see Figure 7).

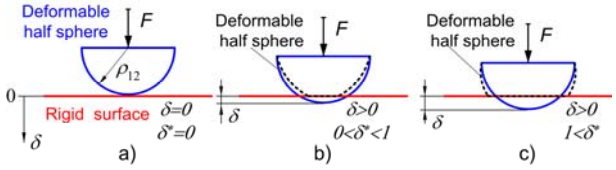


Figure 7

Contact model of a deformable half sphere and a rigid flat surface

a) contact, b) elastic deformation, c) plastic deformation

The boundary of elastic and plastic deformation is a distinctive point. The force to generate elastic deformation is called critical force (F_{Krit}), and the associated deformation is called critical deformation (δ_{krit}). Force and deformation figures are normalized by these values.

In the knowledge of spherical cap height, film thickness determines the rate of deformation δ of the spherical cap (see Figures 6 and 7).

The force F between the surfaces can be determined from this deformation δ. The present calculations used these complex equations to determine the force generated by deformation.

In order to determine the proportions of friction mechanisms, the generation of reduced surfaces is followed by film thickness modifications between a maximum value and zero, and the force arising between each contacting spherical cap and the rigid flat surface is determined for each film thickness, as well as its relation to the force that would arise in the case of nothing but boundary lubrication. This way a curve can be produced to represent the rate of boundary lubrication (ψ) in function of dimensionless film thickness (λ). Figure 8 shows curves determined by this procedure for steel-steel contact pairs of 3 different roughness levels.

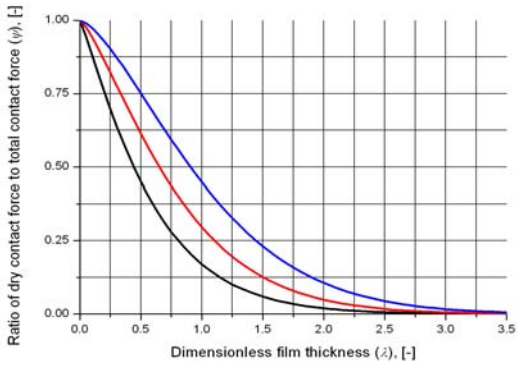


Figure 8

Separation curves of frictions mechanisms determined for steel-steel contact pairs of 3 different roughness levels

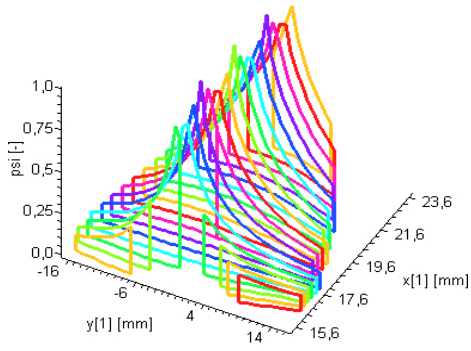


Figure 9

Changes in the proportion of boundary lubrication along contact lines (transverse section)

Figure 9 shows the proportions of boundary lubrication at the worm gear drive examined along contact lines, in case of n_1=1500 1/min input rpm, T_2=570 Nm output torque and FVA 4 mineral oil lubrication, at ϑ_S=60°C oil bath temperature.

At the middle of the tooth, where the film is reduced nearly to zero, is predominantly characterized by boundary lubrication.

3.4 Calculations of Temperature

It is essential to know the temperature course of the contact zone to be able to determine hydrodynamic lubrication.

It can be observed that the heat flow generated in the course of friction (q_{friction}) exits to adjacent walls by conduction (q_1, q_2), and heats up the oil by convection (q_k):

$$\dot{q}_{\text{friction}} = \dot{q}_1 + \dot{q}_2 + \dot{q}_k \tag{12}$$

The temperature course of both the oil film and the contacting surfaces is to be determined on this basis.

3.4.1 Surface Temperature Calculations

The point of departure for determining surface temperatures is Fourier's Law for heat conduction, to be stated generally in the following form:

$$c \cdot \rho \cdot \frac{\partial \vartheta}{\partial t} = -\text{div}(-\lambda \cdot \text{grad}\,\vartheta) \tag{13}$$

Assuming that heat conduction is effected only in the direction of motion (x direction) and into the specimens (y direction), and that material properties are constant, and by taking the motion velocity of the specimens (v_1, v_2) into consideration, equation (12) will be simplified to the following form (index 1 pertains to the worm shaft, and index 2 to the worm wheel):

$$\frac{c_1 \cdot \rho_1 \cdot v_1}{\lambda_1} \cdot \frac{\partial \vartheta_1}{\partial x} - \left(\frac{\partial^2 \vartheta_1}{\partial x^2} + \frac{\partial^2 \vartheta_1}{\partial y^2} \right) = 0$$

$$\frac{c_2 \cdot \rho_2 \cdot v_2}{\lambda_2} \cdot \frac{\partial \vartheta_2}{\partial x} - \left(\frac{\partial^2 \vartheta_2}{\partial x^2} + \frac{\partial^2 \vartheta_2}{\partial y^2} \right) = 0 \tag{14}$$

Gnilke developed a procedure to solve these equations, determining the course of temperature in the form of Fourier Integrals [13]. This procedure was further developed by Plote to determine the temperature distribution of bodies moving into both identical and opposite directions [14].

Based on [14], surface temperature rises can be stated in a dimensionless form as follows:

$$\theta_1 = \frac{\lambda_1 \cdot \vartheta_1}{\mu \cdot p_H \cdot v_S \cdot a_H}, \quad \theta_2 = \frac{\lambda_2 \cdot \vartheta_2}{\mu \cdot p_H \cdot v_S \cdot a_H} \tag{15}$$

On the basis thereof, the dimensionless form of Fourier's equation for heat conduction can be stated as follows:

$$Pe_1 \cdot \frac{\partial \theta_1}{\partial \overline{x}} - \left(\frac{\partial^2 \theta_1}{\partial \overline{x}^2} + \frac{\partial^2 \theta_1}{\partial \overline{y}^2} \right) = 0$$

$$Pe_2 \cdot \frac{\partial \theta_2}{\partial \overline{x}} - \left(\frac{\partial^2 \theta_2}{\partial \overline{x}^2} + \frac{\partial^2 \theta_2}{\partial \overline{y}^2} \right) = 0 \tag{16}$$

Where Pe_1 and Pe_2 are Peclét numbers to be specified as follows:

$$Pe_1 = \frac{c_1 \cdot \rho_1 \cdot v_1 \cdot a_H}{\lambda_1}, \quad Pe_2 = \frac{c_2 \cdot \rho_2 \cdot v_2 \cdot a_H}{\lambda_2} \tag{17}$$

Surface temperature rises can be determined in the following manner from differential equation (17), by disregarding the inferences included in [14], and using Fourier Integrals:

$$\theta_1(\overline{x}) = \frac{1}{\pi} \int_0^\infty \left[c_{11} \cdot \cos(\overline{x} \cdot \varphi_z) - c_{12} \cdot \sin(\overline{x} \cdot \varphi_z) \right] \cdot d\varphi_z$$

$$\theta_2(\overline{x}) = \frac{1}{\pi} \int_0^\infty \left[c_{21} \cdot \cos(\overline{x} \cdot \varphi_z) - c_{22} \cdot \sin(\overline{x} \cdot \varphi_z) \right] \cdot d\varphi_z \tag{18}$$

The solution of (18) requires that the dimensionless flux of heat also be expressed by Fourier Integrals; this can be stated as follows:

$$\dot{Q}(\overline{x}, z) = \frac{\dot{q}(\overline{x})}{\mu \cdot v_S \cdot p_H} = \frac{1}{\pi} \int_0^\infty \left[a_Q(\varphi_z) \cdot \cos(\overline{x} \cdot \varphi_z) + b_Q(\varphi_z) \cdot \sin(\overline{x} \cdot \varphi_z) \right] \cdot d\varphi_z,$$

$$a_Q(\varphi_z) = \int_{\overline{x}_1}^{\overline{x}_2} \left[\dot{Q}(\overline{x}, z) \cdot \cos(\overline{x} \cdot \varphi_z) \right] \cdot d\overline{x},$$

$$b_Q(\varphi_z) = \int_{\overline{x}_1}^{\overline{x}_2} \left[\dot{Q}(\overline{x}, z) \cdot \sin(\overline{x} \cdot \varphi_z) \right] \cdot d\overline{x} \tag{19}$$

Two more boundary conditions are required for determining c_{ij} integration constants in equation 18. In determining these constants, Plote departed from the fact that the entire heat generated flows into the contacting surfaces (20) and that heat fluxes are distributed to the same measure between the surfaces (21):

$$-\frac{\partial \theta_1}{\partial \overline{y}_1}-\frac{\partial \theta_2}{\partial \overline{y}_2}=\dot{Q}\left(\overline{x},z\right) \tag{20}$$

$$\frac{\partial \theta_1}{\partial \overline{y}_1}=\frac{\partial \theta_2}{\partial \overline{y}_2} \tag{21}$$

Now the temperature rise of the contacting specimens can already be determined. By adding the melt temperature thereto, the absolute temperature of the surfaces can be calculated.

3.4.2 Determining Oil Temperature

As a result of internal friction in the oil film and compression, heat is generated, which flows within the oil through heat transport towards the walls and in the direction of motion. This energy balance is described by the correlation termed as energy equation in the EHD theory:

$$\underbrace{-\lambda_{c1}\cdot\frac{\partial \vartheta_1}{\partial y_1}-\lambda_{c2}\cdot\frac{\partial \vartheta_2}{\partial y_2}}_{\text{conduction of heat}}\underbrace{-\rho_F\cdot c_{p,F}\cdot h_0\cdot v_\Sigma\cdot\frac{\partial \vartheta}{\partial x}}_{\text{convection of fluid}}+\underbrace{\eta_{eff}\cdot\frac{v_s^2}{h_0}}_{\text{shear of fluid}}+\underbrace{\rho_F\cdot\frac{\partial v}{\partial \vartheta}\cdot\vartheta\cdot h_0\cdot v_\Sigma\cdot\frac{dp}{dx}}_{\text{compression of fluid}}=0 \tag{22}$$

In order to solve equation (22), it is required to know the temperature profile of the oil along the height of the film (y direction). Accurate specification thereof requires lengthy calculations. Eller departed from the fact that the oil temperature profile in the lubrication gap can be described by a parabola [15] (see Figure 10). This requires knowledge of the surface temperatures (ϑ_1, ϑ_2) and the median integral temperature (ϑ_L). This latter is unknown, and therefore equation (22) needs to be solved in an iterative manner. The temperature profile of the oil is described by equation (23).

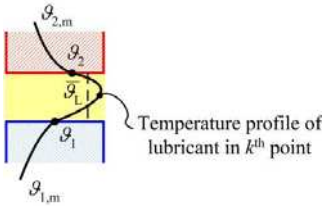


Figure 10
Oil temperature profile in the lubrication gap

$$\vartheta_L\left(y\right)=\left(-6\cdot\overline{\vartheta}_L+3\cdot\vartheta_1+3\cdot\vartheta_2\right)\cdot\left(\frac{y}{h_0}\right)^2+\left(6\cdot\overline{\vartheta}_L-4\cdot\vartheta_1-2\cdot\vartheta_2\right)\cdot\frac{y}{h_0}+\vartheta_1 \tag{23}$$

The term caused by compression and included in equation (23) is quite frequently disregarded in the literature due to its insignificance. Oil convection was also

demonstrated to be insignificant compared to the heat fluxes flowing into the walls. In our experience, consideration of the convective member in the case of low film thickness rates (such as the ones at the middle of contact lines) leads to numerical instability. So it is neglected for this reason and with a view to literature data. Accordingly, energy equation (22) is stated in the following form:

$$\underbrace{-\lambda_{c1}\cdot\frac{\partial\vartheta_1}{\partial y_1}-\lambda_{c2}\cdot\frac{\partial\vartheta_2}{\partial y_2}}_{\text{conduction of heat}}+\underbrace{\eta_{E,\text{eff}}\cdot\frac{v_s^2}{h_0}}_{\text{shear of fluid}}=0 \tag{24}$$

The first two members of equation (24) can be determined on the basis of the oil profile and wall temperatures registered. Oil viscosity depends on both pressure and temperature. The effective viscosity included in equation (24) is the integral mean value of the viscosity values along the height of the lubrication gap. This is why knowledge of the median integral temperature (ϑ_L) is also required for specifying it. As the former is unknown, equation (24) can only be solved iteratively.

The pressure and temperature dependence of viscosity was taken into consideration using the Rodermund equation [16]; this equation assumes the following form:

$$\ln\left(\frac{\eta(\vartheta,p)}{A}\right)=\frac{B}{C+\vartheta}\cdot\left(\frac{p-p_0}{F}+1\right)^{\left(D+E\cdot\frac{B}{C+\vartheta}\right)} \tag{25}$$

The effective viscosity determined by integrating equation (25) presumes the lubrication oil's Newtonian behaviour. The oil in the lubrication gap is exposed to considerable shear, and therefore it demonstrates non-Newtonian behaviour; this can be taken into consideration by Eyring's material law [14]:

$$\eta_{E,\text{eff}}=\eta_{N,e}\cdot\frac{\frac{\tau}{\tau_{Ey}}}{\sinh\left(\frac{\tau}{\tau_{Ey}}\right)} \tag{26}$$

where τ is oil film shearing, and τ_{Ey} is the Eyring shear stress. This is the stress where the so far Newtonian behaviour of the fluid changes to non-Newtonian.

The equations describing surface temperature rise and median fluid temperature are connected equations to be solved together.

Two examples for the temperatures determined by this method are to be presented, one for the rotational motion of the same direction, the other for the rotational motion of the opposite direction of the rolls.

Figure 11 shows wall and median temperatures as well as the course of effective viscosity in case of rotational motion of the same direction of the rolls within the contact area.

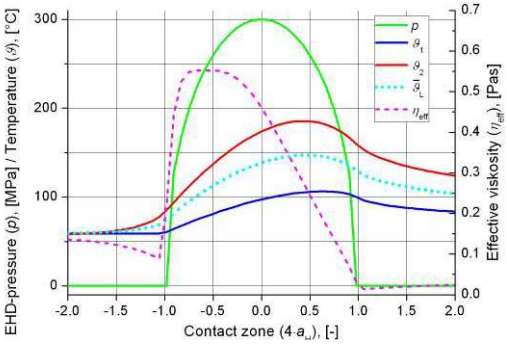


Figure 11

Temperature and viscosity changes in the contact zone in case of rotational motion of the same direction

Figure 12 shows wall and median temperatures as well as the course of effective viscosity in the case of rotational motion of the opposite direction.

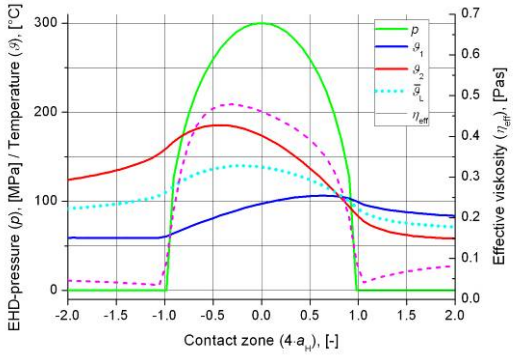


Figure 12

Temperature and viscosity changes in the contact zone in case of rotational motion of the opposite direction

Both the worm shaft and the wheel get more heated up in the case of rotational motion of the opposite direction. Accordingly, the median oil temperature will be higher and the effective oil viscosity will be lower.

3.5 Determining Hydrodynamic Lubrication

Hydrodynamic lubrication arises from film shearing. Shearing is two-directional in the case of worm gear drives, consisting of a component in the direction of motion (x direction) and a component in the direction of the tangent of the contact line (z direction). Taking this and the non-Newtonian behaviour of the oil, shearing stress at a given point of the Hertzian penetration can be specified as follows:

$$\tau_{\mathrm{EHD}} = \sqrt{\left(\tau_{\mathrm{Ey}} \cdot \operatorname{arsinh}\left(\frac{\eta_{\mathrm{N,e}} \cdot v_{\mathrm{rx}}}{\tau_{\mathrm{Ey}} \cdot h}\right)\right)^2 + \left(\tau_{\mathrm{Ey}} \cdot \operatorname{arsinh}\left(\frac{\eta_{\mathrm{N,e}} \cdot v_{\mathrm{rz}}}{\tau_{\mathrm{Ey}} \cdot h}\right)\right)^2} \tag{27}$$

By integrating the (27) equation along the penetration surface, the $F_{\mathrm{F,EHD}}$ frictional force arising from hydrodynamic lubrication is yielded:

$$F_{\mathrm{F,EHD}} = \int_A \tau_{\mathrm{EHD}} \cdot dA \tag{28}$$

Thereby the force arising from hydrodynamic lubrication is known for substituting rolls $F_{\mathrm{F,EHD}}$); the friction coefficient characterizing hydrodynamic lubrication (μ_{EHD}) can already be determined therefrom.

3.6 Determining the Friction Coefficient in Mixed Lubrication

In mixed lubrication, the friction coefficient consists of a component arising from boundary lubrication and one arising from hydrodynamic lubrication. These components are weighted by the function ψ depending on the dimensionless film thickness λ presented above. Hydrodynamic calculations are presented in the previous chapter. Measurement figures are specified by Bouché [4] for boundary lubrication between the tooth flanks of the worm gear drive. According to [4], it ranges between μ_{dry}=0.1÷0.14. For the present investigations, its value was specified at μ_{dry}=0.13.

As presented above, the coefficient pertaining to the state of mixed lubrication can be determined as follows:

$$\mu_{\mathrm{mixed}} = \psi \cdot \mu_{\mathrm{dry}} + (1-\psi) \cdot \mu_{\mathrm{EHD}} \tag{29}$$

Figure 13 shows changes – along contact lines – of the friction coefficient determined by the method described above in the case of n_1=1500 1/min input rpm, T_2=570 Nm output torque and FVA 4 mineral oil lubrication, at ϑ_{S}=60 °C oil bath temperature.

As can be observed, in accordance with the partial results presented so far, the friction coefficient assumes its maximum value at the middle of the tooth dominated by boundary lubrication. There are considerable changes in the friction coefficient even along a single contact line.

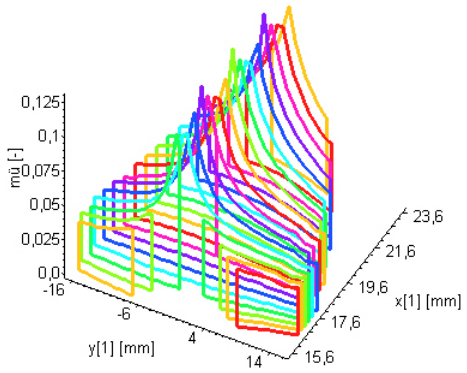


Figure 13
Friction coefficient changes along contact lines (transverse section)

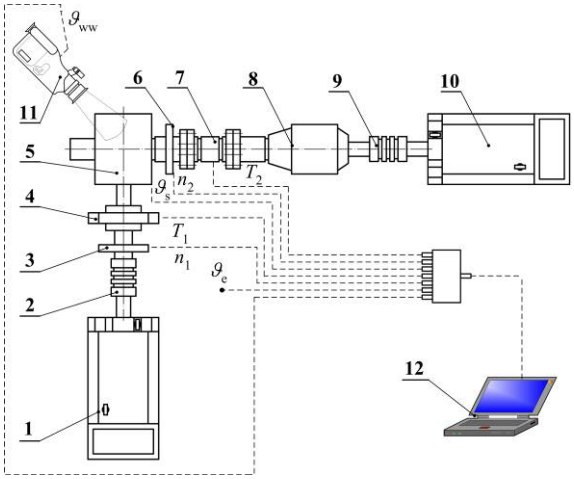
4 Experimental Investigations

The locally changing tooth friction coefficient of worm gear drives cannot be determined directly by gear drive measurements.

The median tooth friction coefficient can be determined indirectly from gear drive efficiency measurements. For this purpose, a test bench was set up at the gear technology laboratory of TU Kaiserslautern Institute of Machine Elements, Gears and Transmissions for testing the efficiency of worm gear drives [17]. The test bench is shown in Figure 14.

In the course of testing, the drive gear set was tested under various torque and rpm loads. Mineral oil with additives of ISO VG 150 viscosity classification was applied as a lubricant, as recommended by the manufacturer. The oil bath precisely covered the worm shaft, and its temperature was set at ϑ_S=60 °C.

Worm gear drive losses can be traced back to four basic reasons: losses by tooth friction, oil churning, bearings and shaft seals [18]. Each source of loss can be divided into load dependent (P_{LP}) and no load dependent (P_{L0}) components. The energy balance of the drive gear can be stated on the basis of the input (P_1) and output power (P_2) of the drive gear and sources of loss. The literature provides a number of solutions for defining loss components. Nass set up equations based on measurements to determine load dependent and no load dependent total loss components [19].



1 30 kW AC drive motor
2 Clutch
3 Speed sensor
4 Torque sensor
5 Worm gear drive
6 Speed sensor
7 Torque sensor
8 Planetary gear
9 Clutch
10 30 kW AC break motor
11 Thermo camera
12 Measurement computer

Figure 14
Test equipment structure

These equations were taken as a basis for evaluating our measurements, adjusted to our specific criteria. For the sake of verification, the values thus yielded were compared to the data resulting from the calculations of [18]. Based on the energy balance of the gear drive, tooth efficiency (η_z) can be specified as follows:

$$\eta_z = \frac{|P_2| + |P_{L02}| + |P_{LP2}|}{|P_1| - |P_{L01}| - |P_{LP1}|} \tag{30}$$

The tooth friction coefficient (μ_{zm}) can be calculated from tooth efficiency by taking tooth geometry into consideration as follows:

$$\mu_z = \tan\left(\arctan\left(\frac{\tan(\gamma_m)}{\eta_z}\right) - \gamma_m\right) \cdot \cos(\alpha_n) \tag{31}$$

Thus, efficiency measurements conducted at n_1=500, 750, 1000, 1500, 2000, 2500 1/min input rpm and T_2=270, 570, 1000 Nm braking torque yielded the average tooth friction coefficient associated with these load points. By averaging the locally changing tooth friction coefficient determined by simulation to a complete revolution of the worm shaft (summarizing the values of Figure 13 in a single coefficient of friction) an average tooth friction value can be obtained which can be compared to the values measured. The simulation took into account the material properties of FVA 4 mineral oil without additives, as the corresponding data of the mineral oil with additives of ISO VG 150 viscosity classification were not available. Here, it must be taken into consideration that the viscosity of FVA 4 oil is significantly higher than that of the oil used for testing.

Figure 15 shows the values of the average tooth friction coefficient determined for T_2=270 Nm load and 5 rpm figures by experimenting and simulation, respectively. The shape of the curves is similar, but it can also be observed that the values measured are overestimated by simulation. The average discrepancy is 0.01.

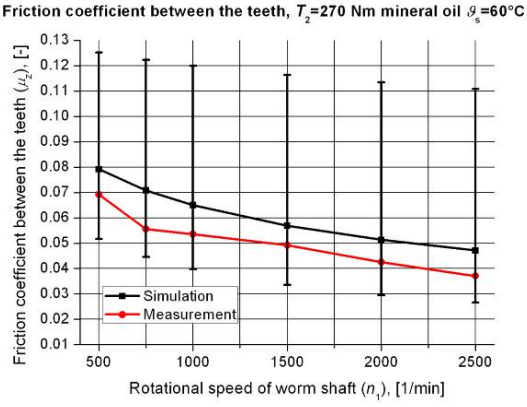


Figure 15
Course of the measured and simulated tooth friction coefficient at 270 Nm output torque

Figure 16 shows the values measured and calculated at T_2=570 Nm load. The shape of the curves is similar again, but it is also conspicuous that the discrepancy is getting larger at an increasing rpm.

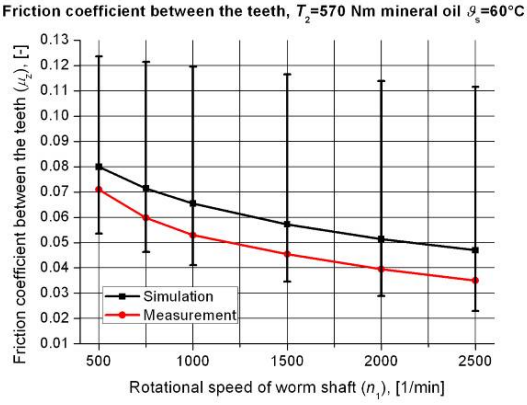


Figure 16
Course of the measured and simulated tooth friction coefficient at 570 Nm output torque

Figure 17 shows the course of the simulated friction coefficients and of the measurement results available at T_2=1000 Nm load. Here, the simulation approximates the measured values with the smallest error, although the initial discrepancy corresponds to that of the earlier two loads.

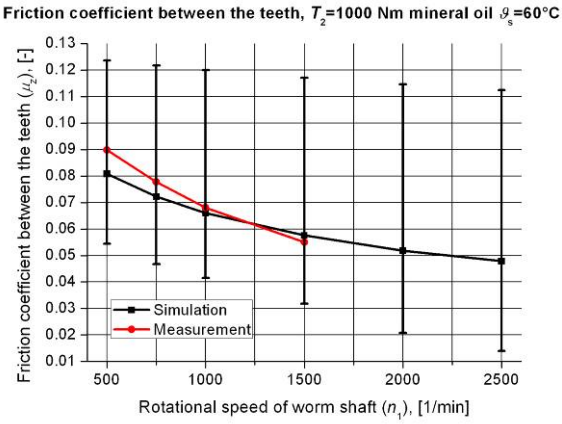


Figure 17
Course of the measured and simulated tooth friction coefficient at 1000 Nm output torque

The measured values were consecutively overestimated by the simulation. This can be attributed to the fact that an oil of higher viscosity was taken into account in the simulation than the oil used for testing. Results can also be affected by the fact that the curve to determine the rate of boundary lubrication models the distribution of asperity heights by Gaussian distribution. The roughness characteristics used for generating substituting surfaces come from 2D roughness measurements. 3D roughness measurements of tooth flanks would yield not only more accurate roughness indices, but would also provide additional information for the statistical description of the height distribution of asperities.

Conclusions

This paper presented a complex calculation procedure to determine the tooth friction coefficient locally changing along the length of the contact lines of worm gear drives. It was also explained how to generate dimensionless Stribeck curves from locally changing tooth friction coefficients determined by such complex calculation; these curves enable the simple still accurate determination of the changing tooth friction coefficient during multibody simulation calculations. The tooth friction coefficients determined by simulation were compared to measured values, and demonstrated satisfactory correspondence.

The authors expect a more accurate agreement by the further development of the calculation algorithm as follows:

- Consideration of the load condition of substituting rolls by load curves based on finite element calculations.
- Application of heat flux distribution changing point by point, describing real conditions more appropriately, in determining surface temperatures.

- Generation of curves to divide boundary and hydrodynamic lubrication using curves based on 3D rough measurements. Assumption of non-Gaussian distribution function if necessary.

Acknowledgements

The authors hereby express their thanks to Dipl.-Ing. Csaba Fábián for his assistance in preparing and conducting measurements. Grateful acknowledgements are also due to Dipl.-Ing. Viktor Aul for useful tribological consultations.

This work was supported by research project OTKA K62875.

References

[1] Niemann, G.: Schneckengetriebe mit flüssiger Reibung. Berlin: VDI-Verlag 1942

[2] Wilkesmann, H.: Berechnung von Schneckengetrieben mit unterschiedlichen Zahnprofilformen. Thesis. München: FZG 1974

[3] Predki W.: Hertzsche Drücke, Schmierspalthöhen und Wirkunggrade von Schneckengetrieben. Thesis. Bochum: LMG 1982

[4] Bouché B.: Reibungszahlen von Schneckengetriebeverzahnungen im Mischreibungsgebiet. Thesis. Bochum: LMG 1991

[5] Johnson K. L.: Regimes of Elastohydrodynamic Lubrication. Journal Mechanical Engineering Science 12 (1970) 1, pp. 9-16

[6] Litvin F. L., Fuentes A.: Gear Geometry and Applied Theory. Cambridge: University Press 2004

[7] Weisel, C.: Schneckengetriebe mit lokal begrenztem Tragbild. Thesis. München: FZG 2009

[8] Venner C. H.: Multilevel Solution of the EHL Line and Point Contact Problems. Thesis. Twente: Emschede 1991

[9] Murch L. E., Wilson W. R. D.: A Thermal Elastohydrodynamic Inlet Zone Analysis. Transaction of the ASME (1975) 4, pp. 212-216

[10] Wiśniewski M. Elastohydrodynamische Schmierung. Renningen-Malmsheim: Expert Verlag 2000

[11] Greenwood, J. A.; Williamson, J. B. P.: Contact of Nominally Flat Surfaces. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. Vol. 295, pp. 300-319, 1966

[12] Whitehouse, D. J.; Archard, J. F.: The Properties of Random Surfaces of Significance in their Contact. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. Vol. 316, pp. 97-121, 1970

[13] Gnilke, W.: Theorie der Mischreibung. Freiberger Forschungshefte: Theorie der Mischreibung und ihre Anwendung auf Gleitlager. Leipzig: VEB Deutscher Verlag für Grundstoffindustrie 1982

[14] Plote, H.: Zur Berechnung Thermo-Elasto-Hydrodynamischer Kontakte. Thesis. Clausthal: IRM 1997

[15] Eller, G.: Ein Beitrag zur Berechnung des stationären, nichtisotermen elastohydrodynamischen Schmierfilms. Thesis. Karlsruhe: IML 1987

[16] Rodermund H.: Extrapolierende Berechnung des Viskositätsverlaufes unter hohen Drücken. Tribologie und Schmierungstechnik 27 (1980) 1, pp. 3-5

[17] Magyar, B.; Horák, P.; Sauer, B.; Fábián, Cs.: Experimentelle Untersuchung der Zahnreibungszahl von Schneckengetrieben mit der Flankenform K. GÉP LXI (2010) 9-10, pp. 51-54

[18] E DIN 3996:2005-08 Tragfähigkeitsberechnung von Zylinder-Schneckengetrieben mit sich rechtwinklig kreuzenden Achsen. Beuth-Verlag: Berlin 2005

[19] Nass U.: Tragfähigkeitssteigerung von Schneckengetrieben durch Optimierung der Schneckenradbronze. Thesis. Bochum: LMGK 1995

2012 Reviewers

Ancza, Erzsébet
Antal, Margit
Badacsonyi, Ferenc
Bakó, András
Balas, Valentina
Baranyi, Péter
Bedő, Zsolt
Beinschróth, József
Benedek, András
Berki, Zsolt
Bilicz, Sándor
Bognár, Rita
Borgulya, Ágnes
Borgulya, Istvánné
Bóta, Gábor
Brtka, Vladimir
Czifra, Árpád
Csapó, Ádám
Csapó, Benő
Csiszár, Csaba
Csopaki, Gyula
Dobay, Péter
Dobrai, Katalin
Farkas, András
Farkas, Ferenc
Frey, Andreas
Gál, Zoltán
Galambos, Péter
Galántai, Aurél
Gázmár, Zoárd
Goda, Tibor
Gvozdenac, Dusan
Gyökér, Irén
Györe, Attila
Halász, Sándor
Horváth, Csaba
Horváth, László
Jagasics, Szilárd
Jánosi, László
Kabor, Jozef
Kádár, Péter
Kárász, Péter
Karlovic, Igor
Kelemen, Jozef
Király, Ágnes
Kis, Tibor
Kispál-Vitai, Zsuzsanna
Kiss, Tibor
Komlósi, Sándor
Komócsin, Mihály
Kovács, Szilveszter
Kovács, Zsolt
Kovács-Coskun, Tünde Anna
Kozlowski, Miklós
Kukolj, Dragan
Kundrák, János
László, István
Lóránd, Balázs
Lovassy, Rita
Nádai, László
Nagy, István
Nagy, Lóránt
Nagy, Tamás
Nemcsics, Ákos
Nemeskéri, Zsolt
Németh-Katona, Judit
Ősz, Rita
Pap, Endre
Pentelényi, Pál
Petkovic, Imre
Petrovic, Predrag
Philippe, Emilia
Poór, József
Precup, Radu-Emil
Preitl, István
Pusztai, József
Raisz, Dávid
Réger, Mihály
Roósz, András
Ruszinkó, Endre
Schuster, György
Szakács, Tamás
Szentgyörgyvölgyi, Rozália
Takács, András
Takács, Márta
Takarics, Béla
Tar, József
Terdik, György
Vajda, István
Vámossy, Zoltán
Várkonyi-Kóczy, Annamária
Zöldy, Máté