ROBOSOFA-Low Cost Multimodal I/O Fusion for Smart Furniture

Rytis Maskeliunas and Vidas Raudonis
Information Faculty, Kaunas University of Technology, Lithuania

Abstract: *The smart furniture provide various services alone or by communication with other devices. The cost and time for building such devices is a barrier to the deployment of various novel applications. Since the smart devices are typically equipped with embedded or traditional computers, various sensors and multimodal I/O devices, it can provide various services in public spaces as well as in private spaces. The goal of our research was to introduce affordable techniques on how to build and model the acceptable multimodal HCIs and interactions with the environment all integrated in the mobile smart furniture device. In this paper we present a detailed analysis of the associated, promising state of the art techniques and the initial set of techniques we have used to realize the multimodal HCI interaction, followed by the descriptions of implemented control algorithms and some first-hand experimental evaluation analysis.*

Keywords: *Human computer interaction, information interfaces and representation, user interfaces, image processing and computer vision, natural language processing, algorithms and applications.*

Received August 16, 2011; accepted December 30, 2012; published online August 5, 2012

1. Introduction

Intelligent furniture is the future development trend of furniture industry [23, 77]. The so called Smart Furniture extemporaneously converts the legacy non-smart space into a Smart Hot-spot which consists of computational services [24]. Since the Smart Furniture is equipped with networked computers, sensors and various I/O devices, it can provide various services by alone or by coordinating with other devices. Many researchers introduced special rooms equipped with sensors, devices, networks, and computers to demonstrate ubiquitous computing environment, such as a personalized message board system, a zero-stop check-out system, etc., [67]. Novel realizations sometimes even have a built-in occupancy monitoring sensor system for intelligent furniture. For example in [31] authors present the related user and service scenarios in the context of living environment for older persons, describing details of the technical solution of the intelligent sofa and the test results of its laboratory use. However, the cost and time for building such a room is a barrier to the deployment of various ubiquitous applications. Since the smart furniture is equipped with a processing center (embedded or traditional computers, various sensors and multimodal I/O devices), it can provide various services in public space as well as in private space.

In this paper various techniques used in the smart furniture of the ROBOSOFA research study to realize the multimodal HCI interaction are described and the description of implemented control algorithms are offered.

The paper is organized as follows: First-we offer the educational review of the 3 main interaction modalities (due to space constraints and not having an implementation in our system choosing not to review other more niche I/O modalities), then we present our own implementations and the initial experimental evaluation.

2. State of the Art in Speech, Gaze and Touch I/Os

2.1. Controls Based on Gaze Tracking

The input devices, which allow to interact with computers or other physical devices without actually touching them opens a new communication potential for people with physical impairments. During the last decade the interest in called gaze tracking systems have been rapidly growing. The amount of the intensive investigation shows that the assistive technologies have huge potentiality of practical usage in the everyday life and more [16, 38].

There are a number of methods for extracting the point of the gaze or the motion of the eye relative to the head from the captured eye or face image. Most popular methods for gaze tracking techniques utilize either bright-pupil or dark-pupil and Purkinje images [35, 37, 38, 49, 71]. The bright-pupil technique illuminates the eye with a source that is on or very near the axis of the camera. The dark-pupil techniques illuminate the eye with an off axis light source. There are four visible Purkinje images: first is reflection of the outer surface of the cornea, second is reflection of the inner surface of the cornea. Third and fourth Purkinje images are the reflections from outer and inner surfaces of the lens, accordingly [12, 68].

Latest contact gaze tracking systems are mostly based on two video cameras and additional IR light source. First camera is dedicated to capture the eye images and second camera is designated to observe the surrounding. The user's eye is usually captured in two ways, camera is directed directly to the eye using special boom arm or the eye image is captured from the semi-transparent mirror or special prism. The additional light source are often design from 1 or more (up to 9) infrared light emitting diodes which are directed to the users eye. The IR diodes are mounted so that the dark or bright pupil is captured in the eye image and it do not dazzling the user with visible light. The cameras are usually mounted on the helmets [69], safety (ordinary) glasses or caps [25, 27, 28, 75].

Gaze detection systems which are based-on special hardware and sensors are also explored [5, 46]. For instance, low computational complexity tracking algorithm can be used by applying symmetry detector for rough pupil localization and triangulation for detection the accurate pupil position [35]. Experiments have shown that the algorithm was able to detect the pupil in 96% of cases and it is robust enough against drastic changes of ambient luminosity. However, proposed system may by uncomfortable, because of its weight and size, since it consists of four video cameras and two servo drives, which are mounted on the glasses. Another eye tracking algorithm is combining feature-based and model-based approaches [4, 16, 17]. Device consists of scene and eye cameras mounted on the glasses. Eye camera works in the near infrared (IR) light. Researches refer to this algorithm as "starburst" because how the pupil features are obtained [16, 55]. The ambient illumination occludes the Purkinje images.

The pupil detection in IR images can be established using an algorithm of symmetric mass center [38]. The pupil is detected in the digitized eye image that is acquired referring to the user defined threshold. Such predefined threshold cannot adaptively change depending on the ambient IR luminosity. Slightly bigger search window than the actual pupil size cannot capture whole dynamical motion of the eye. The accuracy detection of the gaze vectors strongly depends on the position measurements quality of Corneal Reflection (CR) and pupil centers. The task of CR detection often become non-trivial, when system works in the uncontrolled ambient illumination or eye rotates to extreme positions. Gaze tracking technique based on the structured IR illumination overcomes mentioned difficulties [37]. The light source is designed from nine IR emitting diodes which are placed in certain distance from each other and directed to the user's eye. Although, such technique is capable of measuring a wide range of eye movement, system blocks the large portion of field of view. It is uncertain how such mentioned systems react to the different head poses or in changes of calibration conditions.

Head mounted devices are good option when accurate gaze detection is needed while allowing relatively free head movements. Most head mounted trackers uses second video camera, which captures the scene view and allows tracking objects of interest [12, 25, 69]. The angular rotation of the eye ball relative to the head orientation and the position of the head relative to the observable scene have to be measured in order to compute the accurate gaze vector. Yu and Eizenman [72] have developed the gaze-point tracking methodology where the head location and orientation is determined according the location of the objects in the scene. Presented point of gaze tracking system can be used to assess visual patterns with 0,9 angular accuracy.

All eye tracking systems must deal with frustrating problem of "Midas touch", i.e., the fixated gaze can have two meanings [47]. The combination of the gaze and head tracking techniques discriminates more efficiently the voluntary selection of the command from gaze fixation [13, 76]. The head gesture is employed to generate the decision.

It is obvious, that a good gaze tracking system must satisfy strong requirements: the system must work stable and steady in different lightning conditions; the user should be able to calibrate and recalibrate the system easily and independently; the system should be portable, flexible and miniature as possible. Most of reviewed systems are applicable to certain cases, but still struggles with three main tasks: accurate pupil detection, compensation of natural head motions and "Midas touch" problem. In this paper the novel gaze tracking system is proposed that consist from three modules: pupil detection is based-on adaptive grayness thresholding, natural head motions are evaluated using non-linear technique and "Midas touch" task is solved using accumulative luminosity.

2.2. Multilingual Speech Recognition Based Control Interactions

The spoken language techniques are developed for more than 50 years [1]. The effectiveness of the spoken language is astounding [10] and undeniably the speech is the most natural way of interaction. Voice user interface adds up quite well to the common graphical user interfaces [41] and allows overcoming various control problems [19]. It is already proven that the Multilingual Speech Recognition (MSR) models ease the creation of the Automated Speech Recognition (ASR) systems for the other languages an impact factor very important for the less popular tongues, typically with limited linguistic development resources.

The base terms of MSR are described in [3]: the poliphones-phonetical recognition units overlapping a few languages and monophones-phonemes without a clear analog in the other language. In [70] the multilingual English and Swedish phonetic and syntactical systems were developed allowing information querying in both languages. IBM research [11], further conducted an experiment of transferring the English phonemes to French language. 25 poliphones common to both languages were selected and specific English (24) and French (9) monophones were determined. The bilingual recognition analysis proved that the system was more accurate for English recognition (~7% WER) than French (~14% WER) for the artificial (nor real life) speech data. Microsoft has also conducted similar experiments [15, 39] trying to adapt the US English recognizer to Mandarin and Canton recognition. MIT analyzed the possibilities of transferring the Voyager system (US English) to other languages [22]. The spoken language dialogs were evaluated and some factors were determined for English, Japanese and Italian languages. Siemens utilized the OGI Multi Language Telephone Speech Corpus and formed unilingual and multilingual phoneme model [34]. Multilingual models have not improved the recognition accuracy, but overall reduced the number of necessary phonetical units and got somewhere closer to unilingual phone recognition (around 40 %). This type of research was continued [33] by evaluating several EU languages (SpeechDat corpora). The goal was to create a multilingual models and use them for the recognition model of the German language. The results showed that the multilingual model was quite accurate (85,5 % vs 89 % for a single language). For the adoption to German language only 100 German phrases were used and the overall 84.3 % recognition accuracy was achieved.

One of the first multilingual recognition experiments of a large vocabulary [36] was conducted trying to transfer the current English US and French large vocabulary recognition schemes to English UK and German languages. The overall recognition accuracy was 85%. Several institutions [7] researched the possibilities of transferring foreign phonetical models to large vocabulary Czech language recognition. Without using the Czech corpus (only using linguistic knowledge) quite high error rate was achieved (~80%), but with the added Czech data the WER was reduced to ~30%.

Important aspects were determined in the GlobalPhone project. The multilingual (15 European-Asian languages) corpus was developed [57]. The initial analysis of six languages allowed achieving the ~40% WER [58, 59, 60]. Some articulatory features were formed trying to achieve lesser determenencies with variable channel and noise [65, 66] decreasing the WER by 12.3%. Next the aspects of user identity, accent and language setting were evaluated [20, 26, 61]. Overall the English language was recognized most accurately (21-26% WER for single utterances, 43-48% WER for sentences). The phonetical model of all 15 languages allowed achieving the 25% WER for Portuguese [21] compared to Portuguese only model (~10%). Similar principles were adapted to Afrikaner language [48]. The authors utilized a lot of English data and a small number of Afrikaner data. The WER was from 48% to 14% growing with the percentage of Afrikaner data. In [6] the German, English and Spanish model allowed achieving the 85% of recognition accuracy for Slovene language. Later a multilingual model was developed [74] for the Slovene language recognition.

During more recent year the use of closed-source (commercial) English and Spanish recognizers for the recognition of Lithuanian language was analyzed. Some of the transcription (writing the words in recognizer specific phonemes) principles were formed in [29, 42, 43, 44, 45, 50, 51, 52, 53]. Practical evaluations and experimental research with various vocabularies allowed achieving about 95% recognition accuracy for limited size vocabularies. Spanish language proved to be most effective for the task of the recognition of Lithuanian voice commands.

This overview could be concluded with the following notes. First, the use of multilingual recognition seems promising for integration in international research projects. Second in all the experiments reviewed the multilingual recognition models had worse recognition accuracy than single language alternatives. Third-some of the linguistic resources (mostly acoustical data) are still necessary for adding the new language. Fourth-the use of multilingual models allows cutting the development cost. Fifth-there is no clear and universal solution yet.

2.3. Touch Based Solutions

Touch is another wide spread form of interaction not only in HCIs but also in human-human interactions. By general opinion touch input is a must in modern smart devices and in principle is compatible with most devices capable of detecting the movement and the direction of a pointer thus providing the direction vector used for control, though the addition of haptics is still undergoing [18]. Multi-touch interaction is a more free form of interaction suitable even for the disabled [73]. For the applications in virtual exhibition domain, the [9] paper proposes a multi-touch recognition method capable of 3D control actions. In another example a novel interaction technique was developed using a cubic device with five multitouch surfaces [14]. The design of a multitouch gesture sensing environment should allow the user to execute both independent and coordinated gestures. In [8] a test multitouch device built around FTIR technology is illustrated, where a vision system, driven by a visual dataflow programming environment, interprets the user gestures and classifies them into a set of predefined patterns, corresponding to language commands. In [2] an interesting and novel method to enable multi-touch interactions on an arbitrary flat surface using a pair of cameras mounted above the surface is presented allowing robustly identify finger tips and detect touch with a precision of a few millimeters above the surface.

Flat surfaces, such as tables provide a large and natural interface for supporting direct manipulation of visual content for human-to-human interactions. Such surfaces also support collaboration, coordination, and parallel problem solving [63] also presenting considerable challenges, including the need for input methods that transcend traditional mouse- and keyboard-based designs. In [54] authors present observations of user experience on interactive tables in different real world contexts. The effects of the touching manners and the motion directions of human finger in recognizing fine surface texture were investigated in [30]. The authors developed a measurement system to measure the human tactile sensation capability. It was found that the distinctive sensitivities of human tactile sensation in active-touch and passive-touch manners are different in discriminating between fine step-heights and that the directions of finger motion have little effect on the human tactile recognition of fine step-heights.

At the end of this section some examples of a novel touch based interaction systems should be mentioned. AirTouch [56] uses computer vision techniques to remove any need to physically touch the display. The user interacts with a virtual plane that rests in between the user and the display, not limited by requirements that users do not leave the frame or do not perform gestures accidentally, as well as by cost or specialized equipment. Another interesting scientific area is the development tangible user interfaces in augmented reality applications. In [62] authors introduce a novel touch-based interaction technique allowing a direct access and manipulation of virtual content on a registered tracking target. The Haptic Voice Recognizer [64] was developed as a multi-modal interface that combines speech and touch sensory inputs to perform voice recognition allowing to reduce the search space for speech recognition, thereby making the decoding process more efficient and suitable for portable devices with limited compute and memory resources. Sample hardware and algorithms for a real-time social touch gesture recognition system are presented in [32]. Early experiments involved a sensate bear test-rig with full body touch sensing, sensor visualization and gesture recognition capabilities. Algorithms were based on real humans interacting with a plush bear. The system demonstrates the infrastructure to detect three types of touching: social touch, local touch, and sensor-level touch.

To conclude we can say that touch based implementations are important to consider and ever evolving. More and more novel implementations are reaching us every year. Touch is becoming more and more familiar in daily computational devices. Simple touch controls populate nearly every laptop computer in the planet, touch and multitouch screens exist in nearly every smartphone or a more advanced modern entertainment gadget in the planet. Thus in our opinion it is important to include this modality in a multimodal implementation of HCI (if a usage scenario allows it) as most of the users can just “jump in” and use it.

3. The Fusion and Our Implementation of the Three Main Modalities

The performance and efficiency of the associative assistive technology is increased using combination of many signals or control algorithms, i.e., voice recognition, gaze tracking and touch interpretation. The method that combines all these signals must evaluate the data fusion challenge, because many of these signals are activated doing the same mental or physical action task. Therefore, the simple, effective and robust method is presented that makes proper control decision despite the fact that some signals can be more or less erroneous that others in different cases. The proposed integration of different assistive control algorithms is based on up-to-date information, which brings different levels of control to certain user needs. The proposed multimodal interface must incorporate information about the age, skills, native language status, cognitive styles, sensory and motion impairments or other temporary illnesses of the user. For example, a visually impaired user or one with a repetitive stress injury may prefer speech input or touch input. In contrast, a user with a hearing impairment or accented speech may prefer touch or gaze input. The certain modality is adapted based on general information about the user which is acquired by filling the special form.

In the beginning the priorities of control modalities are identified (preset) either by the user or automatically by the environmental (lighting conditions, noise levels, etc.,) scanning features (currently being developed and not implemented) as shown in Figure 1. In this case, the control signal from a higher-priority modality is used first. The control signal of modality with second priority is used when control command is not recognized from the first modality. In this way the proposed multimodal interface is fully depended on the user needs. It allows to create adaptive, flexible and user friendly way of control of mobility device.

In the next part of the paper our proprietary implementation of multimodal associative control interface is presented. All the descriptions of the techniques used are accompanied by the results of initial experimental evaluations.

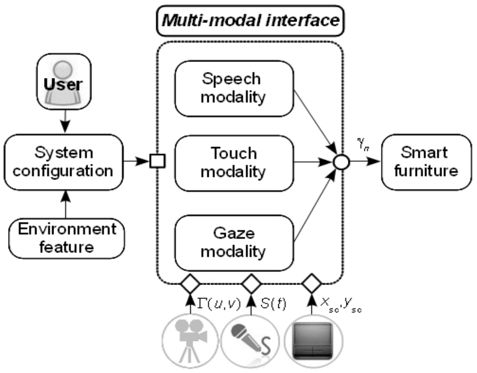

Figure 1. Block diagram of multi-modal fusion.

3.1. The Implementation of Gaze Tracking

We have chosen the monocular video camera which is mounted on the glasses and directed to the user's eye is used for the eye tracking in this work as shown in Figure 2.

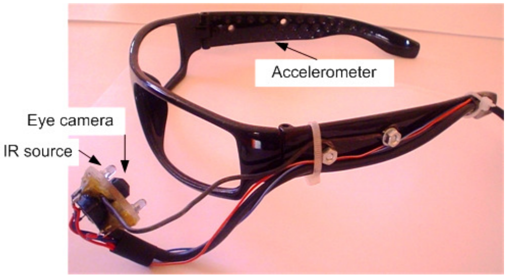

Figure 2. The proprietary hardware of our monocular eye tracking system.

The user's eye is illuminated with three external IR light diodes that does not cause eye discomfort and gives mostly stable ambient IR luminosity. The IR image of the human eye is captured with dark pupil, because IR source is placed of the optical axis. Head orientation in three directions is measured using accelerometer, which is attached to the side of the glasses. The signals from the accelerometer are used for the displacement compensation that accurse because of natural head motions.

In this work we propose the eye tracking algorithm as shown in Figure 3 that is based on three fundamental processes: the evaluation of accumulative luminosity function, adaptive thresholding of grayscale images and the measurements of the head orientation. The luminosity function is used to separate the images of the closed and opened eye. The adaptive thresholding of grayscale images enables the precise detection of the eye pupil in the different layers of gray color, regardless of how the lightening is changing. The diameter measurements of the dark pupil are compared with the limits of possible minimal and maximal pupil size. The pupil center coordinates (x_{pupil}, y_{pupil}) are detected, when the diameter fits the limits.

The control of the external device strongly depends on the calibration conditions and how they are maintained during the usage process. During calibration the relationship between the center coordinate of the pupil in the image and computer mouse position (x_{sc}, y_{sc}) is established. The cursor control problem in the paper is solved using certain virtual calibration grid which stabilizes the cursor position on one point, if the pupil remains in one predefined region. The grid is acquired during the system calibration. For this process the user must observe four corner points of the computer screen and the calibration field is obtained. The grid, which is built from N even sectors, is separately applied to the computer screen and the area of possible pupil trajectory. Each area on the screen has a unique number that is equal to the number of the area on the eye image. The algorithm assigns the center coordinates of the area to the cursor position having the same number as the area in which the pupil center appears on the eye image.

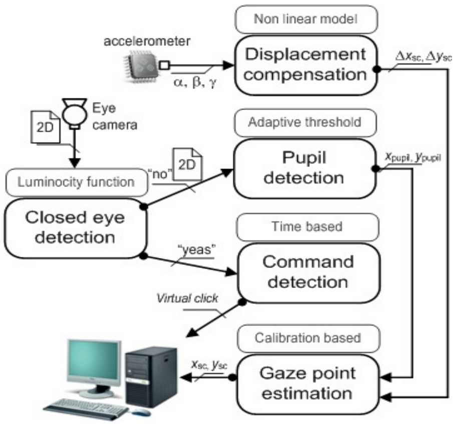

Figure 3. Block scheme of the proposed eye tracking algorithm.

The signals (α, β, γ) captured by orientation measuring device (accelerometer), used to design the technical facility for the eye tracking system. The device is attached to the eyeglasses together with the video camera and can measure all head poses. The device consists of one axis accelerometer and three-axis magnetic field B sensor that measures the direction of the magnetic field in gauss G. The head movement compensation is realized using artificial neural networks, which finds the relationship between the measured direction of the magnetic field and the corners of the calibration field. The relation between the measured angle values and the coordinates is non-linear, so a properly trained neural network is able overcome the non-linearity and give desirable compensation accuracy. The ANN-based compensation algorithm can be considered as a "black box" model and it is essential advantage of the proposed algorithm. It finds the relation based on the training data not using certain predefined coefficients or parameters.

The specially made text writing application was used for testing the proposed gaze tracking system. Ten participants were asked to write a word "HELLO EYETRACKER". The zooming functionality is used in this case, which expands the letters of the button when it is pushed. The mouse clicking operation is executed when user's eye is closed. The application allows selecting one letter at the time using only two mouse clicks. The selected letter is displayed in the text region. The classification rate and the time, which was needed for the establishing the task was recorded during the experiment. Created writing application consists from two regions, i.e., regions of buttons, selected text as shown in Figure 4.

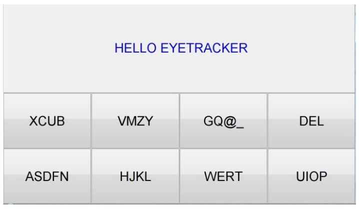

Figure 4. EyeWriter experimental application.

The initial experimental performance results are depicted in Figure 5. The average time that is needed to write the word "EYETRACKER" is equal to 75.2 seconds. About 20 commands per second are generated using proposed gaze tracking algorithm, because algorithm is able to discriminate the processes of the selection and gazing. The tendency was noticed, that the users learn to use the eye tracking system more efficient so the time which is needed for accomplishing the task has been shortened. The text writing application was not designed for the fast text writing, because main task was to show the functionality of proposed eye tracking system.

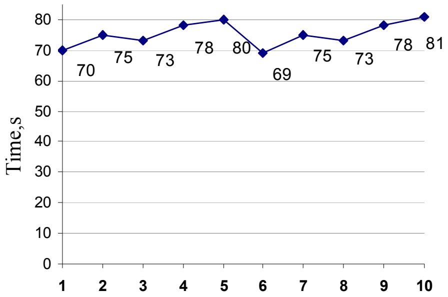

Figure 5. Experimental results of eyewriter application.

3.2. The Implementation of Semantical Keyword Selection Based Speech Recognition

For the speech recognition part we have utilized a proprietary MFCC based HMM, speaker independent bilingual recognizer, running on a netbook computer along the rest of our software, capable of a 95% recognition accuracy in speaker independent mode. A headset was used for collecting the speech input $S(t)$. In case of a paralyzed user who is unable of uttering a clear speech, a proprietary, speaker dependent DTW based ASR recognizer can be used and trained to the specific uttering of a paralyzed person. The algorithm of a dialog model capable of recognizing spoken commands (γ_n) is presented in Figure 6.

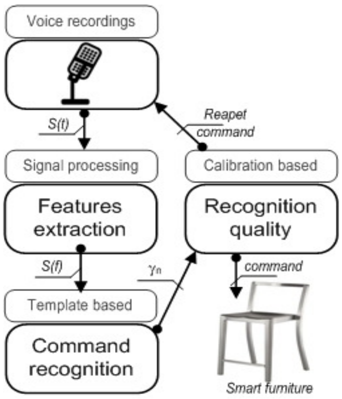

Figure 6. Block scheme of the proposed speech recognition algorithm.

At the start of the dialog a user is prompted (either visually with an added speech, or only by speech, or GUI-depending on the type of application) to enter a command (either by simple voice commands or by the traditional means).

For a more natural flow of "voice interaction" the algorithm features a keyword spotting mode. This way a system is preprogrammed to use a specific set of complex grammar rules, allowing keyword (the important words with a specific semantic value) selection in the user's utterance. This way a user can speak naturally (for example: "please GO to my KITCHEN") and a system only catches the important words (in this case "GO" and "KITCHEN"), assigns the appropriate semantic values and passes for further processing and finally jumps to a next stage in dialog. A correction sub-algorithm is also possible in this case, and if available, a user is offered a list of selection (by voice, or by GUI). Finally the confidence value of the recognized phrase is measured and if it is high enough the semantic value is used in further processing. In case of an unclear recognition (system sees a few choices as similar) an n-best strategy sub-algorithm might be powered and a user might be offered not to repeat the phrase, but to choose between the ones offered to him (the most similar results - i.e., "Did you say: Bath or Path?"). After a successful gathering of the input, the semantic value is processed and the application proceeds to the next stage of a dialog. The biggest advantage of this approach over the isolated words-is the added naturalness, while still maintaining (hopefully) the high enough recognition accuracy.

The recognition accuracy was verified using a 35 speaker corpus (no speech impairment), containing a set of 10 control utterances (100 pronunciations for each speaker, 3500 total) used for a direct control of the device. The results are offered in Figure 7.

The overall 94,55% recognition accuracy was achieved for this corpus. However it is very important to note that this was not a real life conditions, such as normal living conditions or an outdoor environment, and this experiment will be repeated once again in near future, this time with real live persons interacting with our device itself. At this stage the DTW based recognizer was not evaluated.

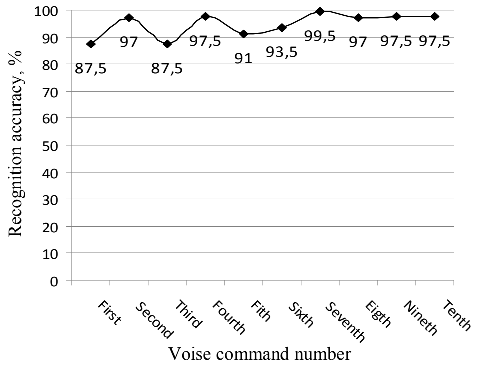

Figure 7. Recognition accuracy of our speech recognition system.

3.3. The Implementation of Touch Input Recognition

For the simplicity of use and the familiarity with most users we have chosen to implement an eyes free touch interface. On the hardware part for the touch recognition we use a large USB touch surface (touchpad) without buttons (only "left click" is registered). A simple control algorithm was developed using a .NET framework controls to allow registering gestures and swipes, and by determining the semantic values further processing them and using for a device control. An algorithm is illustrated in Figure 8.

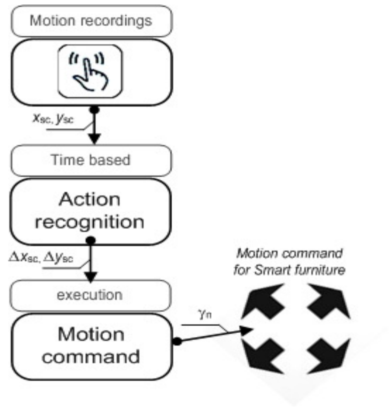

Figure 8. Block scheme of the proposed touch recognition algorithm.

At this stage the touch serves mostly for a direct control of the mobility feature of our smart furniture project. Basically a user swipes a command (x_{sc}, y_{sc}) (for example, left-to go left, forward-to go forward, pinch to slow down, etc.,). Then a timer is fired as a safety feature. A person must lift his finger (or tongue) after each swipe. If he does not lift a finger after a pre-set time, the command is rejected and he is asked to repeat it. If the action is determined (Δx_{sc}, Δy_{sc}) the database of actions is queried and if there is such an action the semantic value (γ_n) of it is passed further down application for further processing.

The specially made application was used for testing the proposed algorithm of the motion recognition. Ten participants were asked to repeat randomly selected motion command for certain time. The recognition accuracy was calculated during the experiment. Created application randomly generates one of the four commands (e.g., UP, DOWN, LEFT, WRITE) time at the time. The recognition is correct when the generated motion command is equal to the motion command generated by the participant.

The experiment was executed in order to show the functionality of proposed system. The experimental results are shown in Figure 9. The average recognition accuracy (97.7%) is relatively high. It can be explained by the simplicity of the proposed control interface (very intuitive) and the ability of human's adaptation.

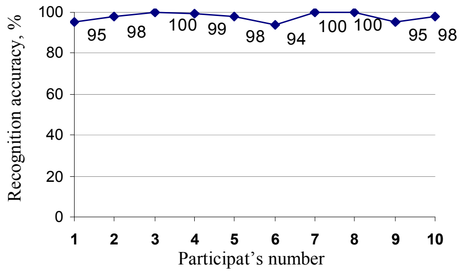

Figure 9. Experimental results of motion recognition testing.

4. Discussion and Conclusions

The initial analysis of the works by the other authors shown that gaze tracking system must satisfy strong requirements: the system must work in different lightning conditions; the user should be able to calibrate and recalibrate the system easily; the system should be portable, flexible and miniature as possible and suitable for performing typical control tasks. Most of reviewed systems are applicable to certain cases, but still struggles with the accurate pupil detection, the compensation of natural head motions and "Midas touch" problems. Our approach overcomes these limitations by implementing a pupil detection based on adaptive grayness thresholding, evaluating natural head motions using non-linear techniques and solving "Midas touch" using accumulative luminosity. About 20 commands per second are generated using our system as algorithm is able to discriminate the processes of the selection and gazing. The tendency was noticed, that the users learn to use the eye tracking system more efficiently so the time which is needed for accomplishing the task can be shortened during the process. The experiments have shown that proposed gaze tracking system can be applied for control purposes and evaluated further.

The current status of multilingual speech recognition research can be evaluated like so. First, the use of multilingual recognition seems promising for integration in international research projects. Second in all the experiments reviewed-the multilingual recognition models had worse recognition accuracy than single language alternatives. Third-some of the linguistic resources (mostly acoustical data) are still necessary for adding the new language. Fourth-the use of multilingual models allows cutting the development cost. Fifth-there is no clear and universal solution yet. We have achieved a 94,55 % recognition accuracy for a small set of control commands. However it is very important to note that this was laboratory test, and this experiment will be repeated once again in near future, this time with real live persons interacting with our device itself.

It is clear that the touch based implementations are important to consider and ever evolving. More and more novel solutions are reaching us every year. Touch is becoming ever familiar in daily computational devices. Simple touch controls populate nearly every laptop computer in the planet, touch and multitouch screens exist in nearly every more modern smart gadget in the planet. Thus in our opinion it is important to include this modality in a multimodal implementation of HCIs (if a usage scenario allows it) as most of the users can just “jump in” and use it. The experiment was executed in order to show the functionality of proposed system for control task. The experimental results of our system were concluded with relatively high recognition accuracy (97.7%). It can be explained by the simplicity of the proposed control interface (it was very intuitive) and low learning curve.

Acknowledgements

Parts of this work were done under the research project “LITVOCROBOT”, No.: 20101216-90 funded by EU SF project “Postdoctoral Fellowship Implementation in Lithuania” and also under the individual scientific research project “ROBOSOFA”, No. MIP-037 funded by the Lithuanian Science Council.

References

[1] Abushariah M., Ainon R., Zainuddin R., Elshafei M., and Khalifa O., “Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus,” *The International Arab Journal of Information Technology*, vol. 9, no. 1, pp. 84-93, 2012.

[2] Agarwal A., Izadi S., Chandraker M., and Blake A., “High Precision Multi-Touch Sensing on Surfaces using Overhead Cameras,” *in Proceedings of the 2nd Annual IEEE International Workshop on Horizontal Interactive Human-Computer Systems*, Newport, pp. 197-200, 2007.

[3] Anderson O., Dalsgaard P., and Barry W., “On the use of Data-Driven Clustering Technique for Identification of Poly- and Mono-Phonemes for Four European Languages,” *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, SA, vol. 1, pp. 121-124, 1994.

[4] Araujo E., “Fuzzy Eye Model,” *in Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, Montreal, pp. 3379-3383, 2007.

[5] Barea R., Boquete L., Mazo M., and Lopez E., “System for Assisted Mobility using Eye Movements Based on Electrooculography,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 4, pp. 202-218, 2002.

[6] Bub U. and Koehler J., “In-Service Adaptation of Multilingual Hidden-Markov-Models,” *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, vol. 2, pp. 1451-1454, 1997.

[7] Byrne W., Beyerlein P., Huerta J.M., Khudanpur S., Marthi B, Morgan J., Peterek N., Picone J., Vergyri D., and Wang T., “Towards Language Independent Acoustic Modeling,” *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, vol. 2, pp. II1029-II1032, 2000.

[8] Celentano A. and Minuto A., “Gestures, Shapes and Multitouch Interaction,” *in Proceedings of the 19th International Workshop on Database and Expert Systems Application*, pp. 137-141, 2008.

[9] Wang C., Dai J., and Wang L., “3D Multi-Touch Recognition Based Virtual Interaction,” *in Proceedings of the 3rd International Congress on Image and Signal Processing*, vol. 3, pp. 1478-1481, 2010.

[10] Chapanis A., “Interactive Communication: A Few Research Answers for a Technological Explosion,” *Nouvelles Tendances de la Communication Homme-Machine (New Trends in Human-Machine Communication)*, pp. 33-67, 1979.

[11] Cohen P., Dharanipragada S., Gros J., Monkowski M., Neti C., Roukos S., and Ward T., “Towards a Universal Speech Recognizer for Multiple Languages,” *in Proceedings of Automatic Speech Recognition and Understanding*, Santa-Barbara, pp. 591-598, 1997.

[12] Coutinho L. and Morimoto H., “Free Head Motion Eye Gaze Tracking using a Single Camera and Multiple Light Sources,” *Brazilian*

Symposium on Computer Graphics and Image Processing, Manaus, pp. 171-178, 2006.

[13] Crisafulli G., Iannizzotto G., and Rosa-La F., "Two Competitive Solutions to the Problem of Remote Eye-Tracking," *in Proceedings of IEEE International Conference on Human System Interaction*, Catania, pp. 356-362, 2009.

[14] Luz T., "3D Interaction for Puzzle Solving with the Cubtile," *in Proceedings of the IEEE Symposium on 3D Multitouch Device 3D User Interfaces*, USA, pp. 129-130, 2011.

[15] Deng L., "Integrated-Multilingual Speech Recognition using Universal Phonological Features in a Functional Speech Production Model," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1007-1010, 1997.

[16] Dongheng L., Babcock J., and Parkhurst D., "OpenEyes: a Low-Cost Head-Mounted Eye-Tracking Solution," *in Proceedings of the Symposium on Eye Tracking Research and Applications*, USA, pp. 95-100, 2006.

[17] Dongheng L., Winfield D., and Parkhurst D., "Starburst: A Hybrid Algorithm for Video-Based Eye Tracking Combining Feature-Based and Model-Based Approaches," *in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 79-99, 2005.

[18] El-Saddik A., "The Potential of Haptics Technologies," *IEEE Instrumentation & Measurement Magazine*, vol. 10, no. 1, pp. 10-17, 2007.

[19] Fezari M. and Bousbia-Salah M., "Implementation of a Hybrid Voice Control System for a Colony of Robots," *The International Arab Journal of Information Technology*, vol. 6, no. 1, pp. 67-71, 2009.

[20] Fugen C., Stuker S., Soltau H., Metze F., and Schultz T., "Efficient Handling of Multilingual Language Models," *in Proceedings of Automatic Speech Recognition and Understanding*, pp. 441-446, 2003.

[21] Garcia E., Mengusoglu E., and Janke E., Multilingual Acoustic Models for Speech Recognition in Low-Resource Devices," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. IV-981-IV-984, 2007.

[22] Glass J., Flammia G., Goodine D., Phillips M., Polifroni J., Sakai S., Seneff S., and Zue V., "Multilingual Spoken-Language Understanding in the MIT Voyager System," *Speech Communication*, vol. 17, no. 1-2, pp. 1-18, 1995.

[23] Wuliji D., "Creative Design of Intelligent Children Furniture," *in Proceedings of IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design*, Wenzhou, pp. 1345-1348, 2009.

[24] Ito M., Iwaya A., Saito M., Nakanishi K., Matsumiya K., Nakazawa J., Nishio N., Takashio K., and Tokuda H., "Smart Furniture: Improvising Ubiquitous Hot-Spot Environment," *International Conference on Distributed Computing Systems*, pp. 248-253, 2003.

[25] Jian-nan C., Peng-yi Z., Si-yi Z., Chuang Z., and Ying H., "Key Techniques of Eye Gaze Tracking Based on Pupil Corneal Reflection," *WRI Global Congress on Intelligent Systems*, vol. 2, pp. 133-138, 2009.

[26] Jin Q., Schultz T., and Waibel A., "Speaker Identification using Multilingual Phone Strings," *in Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, USA, vol. 1, pp. I-145-I-148, 2002.

[27] Gao J., Shuqiang Z., and Wei L., "Application of Hough Transformation in Eye Tracking and Targeting," *in Proceedings of International Conference on Electronic Measurement and Instruments*, Beijing, pp. 751-754, 2009.

[28] Han T., Kriegman D., and Ahuja N., "Appearance-Based Eye Gaze Estimation," *in Proceedings of IEEE Workshop on Applications of Computer Vision*, pp. 191-195, 2002.

[29] Kasparaitis P., "Lithuanian Speech Recognition Using the English Recognizer," *Informatica*, vol. 19, no. 4, pp. 505-516, 2008.

[30] Kawamura T., Otobe Y., and Tani K., "Effect of Touching Manner and Motion Direction of Human Finger on Human Tactile Recognition," *in Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication*, Toyama, pp. 998-1003, 2009.

[31] Kivikunnas S., Strommer E., Korkalainen M., Heikkila T., and Haverinen M., "Sensing Sofa and Its Ubiquitous Use," *in Proceedings of International Conference on Information and Communication Technology Convergence*, Jeju, pp. 559-562, 2010.

[32] Knight H., Toscano R., Stiehl W., Chang A., Yi W., and Breazeal C., "Real-Time Social Touch Gesture Recognition for Sensate Robots," *in Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, MO, pp. 3715-3720, 2009.

[33] Kohler J., "Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, vol. 1, pp. 417-420, 1998.

[34] Kohler J., "Multi-Lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds," *in Proceedings of 4th International*

Conference on Spoken Language, Proceedings, Philadelphia, vol. 4, pp. 2195-2198, 1996.

[35] Kumar N., Kohlbecker S., and Schneider E., "A Novel Approach to Video-Based Pupil Tracking," *IEEE International Conference on Systems, Man and Cybernetics*, San Antonio, pp. 1255-1262, 2009.

[36] Lamel L. and Gauvain J., "Issues in Large Vocabulary, Multilingual Speech Recognition," *in Proceedings of Eurospeech*, pp. 185-188, 1995.

[37] Li F., Kolakowski S., and Pelz J., "Using Structured Illumination to Enhance Video-Based Eye Tracking," *IEEE International Conference on Image Processing*, vol. 1, pp. 373-376, 2007.

[38] Lijima A., Haida M., Ishikawa N., Minamitani H., and Shinohara Y., "Head Mounted Goggle System with Liquid Crystal Display for Evaluation of Eye Tracking Functions on Neurological Disease Patients," *in Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4, pp. 3225-3228, 2003.

[39] Lin H., Li D., Droppo J., Dong Y., and Acero A., "Learning Methods in Multilingual Speech Recognition," *NIPS Workshop,* available at: http://ssli.ee.washington.edu/people/hlin/papers/nips2008WSL1_03.pdf, last visited 2008.

[40] Long X., Tonguz O., and Kiderman A., "A High Speed Eye Tracking System with Robust Pupil Center Estimation Algorithm," *International Conference of the Engineering in Medicine and Biology Society*, pp. 3331-3334, 2007.

[41] Maskeliunas R., Ratkevicius K., Rudzionis A., and Rudzionis V., "SALT-Markup Language for Speech-Enabled Web Pages," *Information Technology and Control*, vol. 34, no. 2, pp. 145-152, 2009.

[42] Maskeliūnas R., Rudžionis A., Ratkevičius K., and Rudžionis V., "Investigation of Foreign Languages Models for Lithuanian Speech Recognition," *Electronics and Electrical Engineering*, vol. 3, no. 91, pp. 15-21, 2009.

[43] Maskeliunas R., Rudzionis A., Ratkevicius K. and Rudzionis V., "User Identification Based on Lithuanian Digits Recognition," *in Proceedings of the 15th International Conference on Information and Software Technologies*, pp. 256-262, 2009.

[44] Maskeliunas R., Rudzionis A., and Rudzionis V., "Analysis of the Possibilities to Adapt the Foreign Language Speech Recognition Engines for the Lithuanian Spoken Commands Recognition," *in Proceedings of Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, Berlin, pp. 409-422, 2009.

[45] Maskeliunas R., Rudzionis A., and Rudzionis V., "Advances on the use of the Foreign Language Recognizer," *in Proceedings of Development of Multimodal Interfaces: Active Listening and Synchrony*, Berlin, vol. 5967, pp. 217-224, 2010.

[46] Miyashita H., Hayashi M., and Okada K., "Implementation of EOG-Based Gaze Estimation in HMD with Head-Tracker," *in Proceedings of the 18th International Conference on Artificial Reality and Telexistence*, pp. 20-27, 2008.

[47] Moriyama T., Jing X. and Cohn J.F., "Meticulously detailed eye model and its application to analysis of facial image," *in Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, vol. 1, pp. 629-634, 2004.

[48] Nieuwoudt C. and Botha E., "Cross-Language use of Acoustic Information for Automatic Speech Recognition," *Speech Communication*, vol. 38, no. 1-2, pp. 101-113, 2002.

[49] Park M., Lee H., and Choi J., "Wearable Augmented Reality System using Gaze Interation," *in Proceedings of IEEE International Symposiumon Mized and Augmented Reality*, Cambridge, pp. 175-176, 2008.

[50] Rudzionis A., Puniene J., Rudzionis V., Punys G., and Maskeliunas R., "The Problems of Cross-Modal Analysis in Verbal and Non-verbal Communication in the Framework of the EU COST2102 Program," *in Proceedings of the 14th Conference on Information and Software Technologies*, pp. 78-83, 2008.

[51] Rudzionis A., Maskeliunas R., Ratkevicius K., and Rudzionis V., "Investigation of Voice Servers Application for Lithuanian Language," *Electronics and Electrical Engineering*, vol. 6 no. 78, pp. 43-46, 2007.

[52] Rudzionis A., Ratkevicius K., and Maskeliunas R., "Adaptation of English Speech Recognition Engines for Lithuanian Speech Recognition," *in Proceedings of the 3rd Baltic Conference on Human Language Technologies*, pp. 265-271, 2007.

[53] Rudzionis V., Maskeliunas R., and Rudzionis A., "On the Adaptation of Foreign Language Speech Recognition Engines for Lithuanian Speech Recognition," *Business Information System, Springer LNBIP37*, pp. 113-118, 2009.

[54] Ryall K., Morris R., Everitt K., Forlines C., and Chia Shen, "Experiences with and Observations of Direct-Touch Tabletops," *in Proceedings of the 1st IEEE International Workshop on Horizontal Interactive Human-Computer Systems*, USA, pp. 89-96, 2006.

[55] Ryan J., Woodard L., Duchowski T., and Birchfield T., "Adapting Starburst for Elliptical Iris Segmentation," *in Proceedings of IEEE International Conference on Biometrics: Theory,*

Applications and Systems, Arlington, pp. 1-7, 2008.

[56] Schlegel R., Chen C., Xiong C., Delmerico A., and Corso J., "AirTouch: Interacting with Computer Systems at a Distance," *in Proceedings of IEEE Winter Vision Meetings: Workshop on Applications of Computer Vision*, pp. 1-8, 2011.

[57] Schultz T., "GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University," *in Proceedings of ICSLP*, vol. 1, pp. 345-348, 2002.

[58] Schultz T. and Waibel A., "Experiments on Cross-language Acoustic Modeling," *in Proceedings of Eurospeech*, pp. 2721-2725, 2001.

[59] Schultz T. and Waibel A., "Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets," *in Proceedings of Eurospeech*, pp. 371-374, 1997.

[60] Schultz T. and Waibel A., "Multilingual and Crosslingual Speech Recognition," *in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 259-262, 1998.

[61] Schultz T., Jin Q., Laskowski K., Tribble A., and Waibel A., "Speaker, Accent, and Language Identification using Multilingual Phone Strings," *in Proceedings of the Human Language Technology Meeting*, pp. 125-131, 2002.

[62] Seichter H., Grasset R., Looser J., and Billinghurst M., "Multitouch interaction for Tangible User Interfaces," *in Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality*, pp. 213-214, 2009.

[63] Shen C., Ryall K., Forlines C., Esenther A., Vernier F., Everitt K., Wu M., Wigdor D., Morris M., Hancock M., and Tse E., "Informing the Design of Direct-Touch Tabletops," *IEEE Computer Graphics and Applications*, vol. 26 , no. 5, pp. 36-46, 2008.

[64] Sim K., "Haptic Voice Recognition: Augmenting Speech Modality with Touch Events for Efficient Speech Recognition," *in Proceedings of IEEE Spoken Language Technology Workshop*, Berkeley, pp. 73-78, 2010.

[65] Stuker S., Metze F., Schultz T., and Waibel A., "Integrating Multilingual Articulatory Features into Speech Recognition," *in Proceedings of Eurospeech*, pp. 1033-1036, 2003.

[66] Stuker S., Schultz T., Metze F., and Waibel A., "Multilingual Articulatory Features," *in Proceedings of ICASSP*, vol. 1, pp. I-144-I-147, 2003.

[67] Tokuda H., Takashio K., Nakazawa J., Matsumiya K., Ito M., and Saito M., "SF2: Smart Furniture for Creating Ubiquitous Applications," *in Proceedings of International Symposium on Applications and the Internet*, pp. 423-429, 2004.

[68] Villanueva A. and Cebeza R., "Evaluation of Corneal Refraction in a Model of a Gaze Tracking System," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 12, pp. 2812-2822, 2008.

[69] Vockeroth J., Dera T., Boening G., Bartl K., Bardins S., and Schneider E., "The Combination of a Mobile Gaze-Driven and a Head-Mounted Camera in a Hybrid Perspective Setup," *in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Montreal, pp. 2576-2581, 2007.

[70] Weng F., Bratt H., and Stolcke A., "A Study of Multilingual Speech Recognition," *in Proceedings of Eurospeech*, pp. 359-362, 1997.

[71] Xu Z., Xinyan Z., and Yingwei Y., "Employing United Delauney Triangulation in Contour Lines Generalization," *in Proceedings of International Conference on Geoinformatics*, Fairfax, pp. 1-6, 2009.

[72] Yu H. and Eizenman M., "A New Methodology for Determining Point-of-Gaze in Head-Mounted Eye Tracking Systems," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 10, pp. 1765-1773, 2004.

[73] Yu Y., Ying L., and Barner K., "Tactile Gesture Recognition for People with Disabilities," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 461-464, 2005.

[74] Zgank A., Kacic Z., Diehl F., Vicsi K., Szaszak G., Juhar J., and Lihan S., "The COST278 MASPER Initiative-Crosslingual Speech Recognition with Large Telephone Databases," *in Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 2107-2110, 2004.

[75] Zhang Y., Cheng B., Feng J., and Li W., "Real-Time Driver Eye Detection Method using Support Vector Machine With HU Invariant Moments," *in Proceedings of the 7th International Conference on Machine Learning and Cybernetics*, Kunming, pp. 2999-3004, 2008.

[76] Zhang J., Zhao M., Zhou L., and Shen L., "A Low Complexity Method for Real-time Gaze Tracking," *in Proceedings of the 10th IEEE Workshop on Multimedia Signal Processing*, Cairns, pp. 883-886, 2008.

[77] Zongdeng Z. and Wenjin L., "The Innovative Design Method of Intelligent Furniture Intelligent System Design and Engineering Application," *in Proceedings of International Conference on Intelligent System Design and Engineering Application*, vol. 2, pp. 673-677, Changsha, 2010.

Rytis Maskeliunas received his PhD degree in computer science, in 2009 from Kaunas University of Technology, Lithuania. He is a senior scientific researcher and a project manager in computer science field at Kaunas University of Technology, Information Technology Development and Automation and Control Systems Institutes, with an expertise in development and analysis of multimodal interfaces, automatic speech recognizers. He has won various awards/honours including the National Science Academy Award for Young Scholars of Lithuania (2010), the Postdoctoral Research Fellowship (2010), the Best Master (2004) and Master Work (2006). He has coordinated/participated in several research projects in computer science domain and was involved in the EU COST actions 278, 2102 and is an MC member (Lithuania) of the currently running COST IC1002. He is a member of an iEEE, author/co-author of over 30 refereed scientific articles and serves as a reviewer for a number of refereed journals. His research interest include modelling, development and analysis of multimodal interfaces, engineering of virtualization systems, programming web and telephony servers and applications.

Vidas Raudonis received his PhD degree in computer science filed, in 2010, from Kaunas University of Technology, Lithuania. He is a lecturer and scientific researcher in the Department of Control Technologies at the Faculty of Electrical and Control Engineering at Kaunas University of Technology. He is participating in various research projects with industrial companies like Siemens, Beijer Electronics. His research interest include application of computation intelligence in human-computer interaction, computer vision, assistive technology and robotics, video input technologies, ANN based systems, automation systems. He published over 20 refereed scientific papers in these fields. Together with colleagues from Lithuania.