User Preference Model for Conscious Services in Smart Environments

Andrey Ronzhin, Jesus Savage, and Sergey Glazkov

***Abstract*—Awareness of user preferences and analysis of the current situation makes capable to provide user with invasive services in various applications of smart environments. In smart meeting rooms context-aware systems analyze user behavior based on multimodal sensor data and provide proactive services for meeting support, including active control PTZ (pan, tilt and zoom) cameras, microphone arrays, context dependent automatic archiving and web-transmission of meeting data at the interaction. History of interaction sessions between a user and a service is used for knowledge accumulation in order to forecast user behavior during the next visit. The user preference model based on audiovisual data recorded during interaction and statistics of his/her speech activity, requests, movement trajectories and other parameters was implemented for the developed mobile information robot and smart meeting room.**

***Index Terms*—User preferences, context awareness, action recognition, mobile robot, smart meeting room.**

I. Introduction

THE notions of user model and context are fundamental for artificial intelligence and human-machine interaction in particular. Creation of user model or profile involves gathering user information during his/her interaction with a system. The primary aim of the system personalization is to improve user experience and get relevant service in the current situation [1]. The context change could be caused both a user and environments, in which interaction takes place.

The difference in abilities, interests, roles, location of a user as well as history of previous interaction sessions are main factors considered by context-aware systems concerned with acquisition, understanding of context and action based on the recognized context [2]. The problems of context representation, sensor uncertainty and unreliability are considered in numerous works. However, there is no any accepted opinion on types and number of context spaces and their attributes, as well as there is a lack of universal approaches to the problem of context prediction, especially for acting on predicted context [3].

Manuscript received June 29, 2011. Manuscript accepted for publication August 25, 2011.

This work is supported by Saint-Petersburg State University (project # 31.37.103.2011), the Russian Federal Targeted Program (contracts #P876 and #14.740.11.0357) of the Ministry of Science and Education of Russia.

Andrey Ronzhin is with St. Petersburg State University, 11, Universitetskaya nab., St. Petersburg, 199034, Russia. Jesus Savage is with the Universidad Nacional Autonoma de Mexico, Mexico City, Mexico (e-mail: savage@servidor.unam.mx). Sergey Glazkov is with the Russian Academy of Sciences St. Petersburg Institute for Informatics and Automation RAS, St. Petersburg, 39, 14 Line, 199178, Russia (e-mail: glazkov@iias.spb.su).

Location and time have been the commonly used components of the context. Computing context, user context and physical context were selected by Schilit et al. [4]. User's location, environment, identity and time were analyzed at the context definition by Ryan et al. [5]. Three different categories of contexts were proposed in [6]: (1) real-time (location, orientation, temperature, noise level, phone profile, battery level, proximity, etc.); (2) historical (for instance, previous location, and previous humidity and device settings); (3) reasoned (movement, destination, weather, time, user activity, content format, relationship, etc.).

In [7], the context information used for service personalization and designing of multimedia applications for heterogeneous mobile devices were divided into the five categories: spatio-temporal (place, time), environment, personal, task, social. A personalization service based on user profile retrieves user context and context history information from context management services. It helps the user to get relevant content and services in the current situation.

The human beings, the physical and informational environments were considered by Dai et al. in the framework of two types of contexts [8]: interaction context representing interactive situations among people and environment context describing meeting room settings. They use propositions that the interaction context of a meeting has a hierarchical structure and expresses the context as a tree. User's standing-sitting states, changing user's location, face orientation, head gestures, hand actions, speaker turns and other events are analyzed for the context prediction. A Finite State Machine framework was introduced in order to classify these meaningful participants' actions. However, before the classification an event should be detected, so particular issues of signal capturing and feature extraction are appeared.

The rest of the paper is organized as follows. Section 2 describes the appropriate audio and video processing techniques used for evaluation of user behavior as well as context acquisition and analysis in smart environments including smart meeting rooms and social robots. The issue of evaluation of user behavior and his preferences during interaction with intelligent services equipped by different types of user interface is considered in Section 3. The results of cognitive evaluation of three types of user interfaces for the developed information mobile robot are discussed in Section 4. The architecture of the meeting web-transmission system, which performs selection and transmission of the most actual multimedia content captured from video cameras, whiteboard, presentation slides, based on context analysis

during the meeting in the smart room, is presented in Section 5. Conclusions and plans for future work are outlined in Section 6.

II. Context Acquisition and Analysis in Smart Environments

In a smart meeting environment, to provide conscious services context-aware systems should analyze user behavior based on multimodal sensor data and provide proactive services for meeting support, including active control PTZ (pan, tilt and zoom) cameras, microphone arrays, context dependent automatic archiving and web-transmission of meeting data at the interaction. Automatic analysis of audio and video data recorded during a meeting is not a trivial task, since it is necessary to track a lot of participants, who randomly change positions of their bodies, heads and gazes. Audio-visual tracking has been thoroughly investigated in the framework of CHIL and AMI/AMIDA projects [9, 10].

Use of panoramic and personal cameras is suitable for recording a small-sized meeting, where all the participants are located at one table. In a medium size meeting room (~50 people), a larger space should be processed that affects on the cost of recording technical equipment too [11]. Distributed systems of microphone arrays, intelligent cameras and other sensors were employed for detecting participant's location and selection of a current speaker in the medium meeting room.

Let us consider several recent works devoted to analysis of meeting participant behavior. Zhang et al. proposed a speaker detector for the Microsoft RoundTable distributed meeting device [12]. It has a six-element circular microphone array at the base, and five video cameras at the top. The proposed algorithm fuses audio and visual information at feature level by boosting to select features from a combined pool of both audio and visual features simultaneously. Audio related features are extracted from the output of the maximum likelihood based sound source localization (SSL) algorithm instead of the original audio signal. They achieved a speaker detection rate of 93%, a person detection rate of 96%, and multimodal speaker detection of 98%.

A ceiling 4-camera tracking system, a 360° camera, a single microphone for speaker identification, and a circular 16-microphone array were used in the University of Southern California smart room [13]. A mixture particle filter was used for tracking an unknown number of acoustic sources. The angular estimates of source locations were obtained using a variant of time difference of arrival (TDOA) method for each microphone pair. Speaker detection rate was around 90% during four sessions with approximate length of 15 minutes.

Raykar et al. [14] compared the performance of GCC-PHAT, GCC-ML, Brandstein's pitch-based, and the method based on characteristics of the excitation source during the production of speech using an 8 element microphone array in an office room of dimension 5.67x4.53x2.68 m with an average reverberation time of about 0.2 s and noise level of about 40–50 dB. Signal from each channel is sampled at 8 kHz frequency. Some cases were considered during the experiments: the source was placed at a distance of 2.0 m from the center of microphone pair which are 1 m apart; the speaker moved in such a way that he was always facing the microphones. The error is generally lower for frames where signal energy is high, and also a lower error is obtained when larger frame sizes are used. Using a frame size of 500 ms with frame shift of 50 ms the localization error for the proposed method was lower 30 cm.

Multiband joint position-pitch algorithm for 24 channel circular microphone array was proposed to track a single speaker and multiple concurrent speakers in the meeting room measuring 6.02x5.32x3 m [15]. The array was placed in the center of the room; the loudspeakers were positioned at a constant distance of approximately 2 m from the array. Experiments using real-world recordings in a typically reverberant meeting room showed a frame-wise localization estimation score of about 95% for tracking a single speaker.

The approaches based on the signal (also interaural) level difference between different microphones, and TDOA were tested in a train compartment for aggressive behavior detection [16]. The experiments are concentrated in an area having a length of about 7.5 m with eight predefined candidate locations and four microphones. The mean square error of location estimation for sources near microphones was lower 50 cm, but the performance significantly decreased at the detection of far-field sources.

In the DICIT project the harmonic linear array of 13 microphones was used for detection of up to four persons in a room of dimension 3.4x5.0m, which control an interactive television. 4 person positions were investigated at 2.1 m distance from the microphone array [17]. Adaptive Eigenvalue Decomposition was implemented as an alternative to GCC-PHAT in TDOA estimation. A localization error was labeled either as gross, when it is larger than 0.5m, or as fine otherwise. Localization rate (LR) was defined as the percentage of fine errors over all the localization estimates. Localization accuracy is measured in terms of Root Mean Square Error (RMSE) of all the localization errors (fine and gross). In the 30dB SNR case the localization rate was about 97% and RMSE was lower 25 cm.

Summing up the review, it should be noted that distance between center of microphone array and sound source was lower 3 m in all the considered papers, where the SSL methods were evaluated. Positions of speakers were predefined in most of the applications and position number was up to six. The aim of our study is to select current active speaker in the medium meeting room with the number of sitting participants up to 42. Besides of smart environment, the social mobile robots, which are capable to natural interaction with a user, are actively investigated now.

Robot Neel, developed by an Indian group HitechRoboticSystemz Ltd, is an autonomous reference robot, which provides information services to visitors in shopping mall [18]. The robot navigation system is based on laser

sensors and route planning for a given map. The robot is equipped with a touch screen with graphical menus, menu items can be synthesized by Microsoft Windows TTS. The system of interaction with a user applies speech synthesis and a graphical menu. A user selects goods or services on the touch screen, the robot pronounces his/her choice and the response to the user's query. Neel robot is connected to the information network of the shopping center and notified of all changes, availability of goods and services. Also, when interacting with people the robot creates a database of visitors and their preferences based on analysis of user queries. Currently, the robot is able to independently navigate a given route and to identify obstacles. The user interface is based on JavaFX, which allows quick change of graphical part of the interface.

System with a multimodal user interface, including at least speech recognition and synthesis, in addition to the graphical menu, will benefit for a lot more groups. For example, visually impaired people can interact with the system in a natural way using speech. An example of such systems is a robot FriDA, which was developed by Korean company DASA TECH Co. Ltd. FriDA. This robot is equipped with a touchscreen monitor, speakers, and a microphone array. The monitor has standard graphic menus, as well as speakers and microphones to ensure system of synthesis and speech recognition. The robot is designed to provide reference information at the airport in a verbal dialogue mode and can display and pronounce data required by user.

Systems with three-dimensional avatar of the human head are able to communicate with hearing disabled people. Lip movements of avatars are synchronized with the speech signal, which makes possibility for lip reading. For example, a robot secretary HALA, developed at the University of Carnegie Mellon, is equipped with a touch screen, which displays animated avatars, speaker, microphone and an infrared sensor to determine the presence of a user [19]. HALA can lead voice dialogue with a user in Arabic and English, the avatar is used for verbal expressions (movements of the lips are applied in the process of speech synthesis) and nonverbal means (shaking his head, facial movements).

Recently there was a tendency to create humanoid robots with the approximate shape of the hull, with varying degrees, to the human body shape. Such robots are able to interact with a person, not only through speech but also with gestures. Typically, these robots are not equipped with monitors, therefore they have not any graphical interface. For example, the robot Robotinho, developed at the University of Bonn in Germany, has a humanoid form, and can interact with humans through speech, gestures and facial expressions [20]. The robot uses mixed system of dialogue, and is able to determine position of a user and his face, as well as to recognize and synthesize speech. Robotinho can express its emotional state and communicate with many people simultaneously. Since the robot has a humanoid body shape, it can nonverbally communicate with users through gestures during the dialogue, as well as attract users' attention to itself or to the objects of the environment by gestures or gaze direction. The robot detects a user with two laser range finders, and then he finds a human face with two video cameras. When interacting with users it creates a database of users containing user's face images and his/her preferences, based on the query history. In future the robot will be able to identify user.

Thus, the appointment of the robot and the possibilities of potential users are necessary to consider at the development of multimodal interfaces for a social robot. Ways of interaction must be easy-to-use and do not require special training of users. So, speech and multimodal interfaces with speech processing, are being actively researched and applied in robotic systems. Despite the fact that user interaction with social robots in most cases takes place in an environment with high noises, speech interfaces, and multimodal, including speech and gesture processing, are being actively studied and applied research in robotic systems [21, 22, 23].

III. Influence of User Interface on User Behavior

Fundamental principles of the field of human-computer interaction lays the basis for the design of dialogue models, also the capabilities of modern hardware and software that implement the input, output and processing of information channels available to the user are taken into account. With the development of socially oriented services, it became clear that the interfaces for interaction of the system with a user should be simpler, more intuitive and do not require additional knowledge and training.

The standard interface is a graphical user menu, which includes information inputting by a user in manual mode (keyboard, mouse, touchscreen monitor). The most widespread of such interface has received in a self-service machine, such as payment terminals or ATM services. This kind of interaction is not always convenient for a user, and often even impossible, for example, people with disabilities are not able to interact in this way (blind, armless, etc.). To increase the opportunities of graphical user interface voice prompts to the menu should be used in self-service machines and robots are used.

The standard graphical user interface remained the most common before the appearance of complex interactive systems for mass services. Much greater attention is now given to the development of queuing systems with multimodal user interfaces based on analysis of speech, gestures, and graphical user interface, three-dimensional model of a human head with a strong articulation of speech, facial expressions and other natural means of communication for interpersonal communication.

Besides of interface type, various factors influence on user behavior, for example, the general context and peculiar features of the task; experience of human-computer interaction. The point is that the user usually keeps in his mind all the experience of the same kind, so time after time he/she tends to use one and the same algorithm of interaction,

ignoring new modalities and options of a system. The main purpose of the present investigation is to assess user behavior and his preferences during interaction with intelligent services. Let us consider several types of interface, which are used in our experiments during testing an inquiry system.

Visual interface gives complete information; an inquiry is outputted to the screen, variants of answers to choose by pushing the menu items on the touchscreen. In this instance minimum of speech actions is expected from the user, especially speech communication with the robot. The potential client sees the interface assuming tactile-visual interaction, and is not ready to think of possibility of speech modality, even if this function is available. However predisposition to a choice of a touch modality instead of the speech one depends on the visual components of the interface.

In the case of a visual-speech interface, questions are synthesized by voice without any text duplication on the screen, and variants of answers are outputted to the display. Presence of output speech modality stimulated the user to give speech responses.

In speech interface (even combined with a visual component of the dialogue-system), both questions and variants of answers are voiced, a speech modality can be preferred, even if a touch modality is available. The choice of the speech interface can be made as the most natural.

In the case of both speech and visual interface with a full duplication of speech by the text on the touchscreen, it is expect that user behavior will similar to the variant with a completely visual interface owing to more informativity of the visual modality.

All the described interface cases are suitable for those tasks of dialogue interaction when there is no obvious requirement for a combination of interfaces (for example, speech and touch ones). Depending on type of a problem and type of information used during the interaction, as well as user experience, the necessary modality combination will be chosen by the user.

IV. Use Case: User Interface for Information Mobile Robot

The developed mobile robot consists of a mobile information platform and information desk. Multimodal user interface, developed earlier for the stationary information kiosk, was used in the design of the mobile version [24]. First of all the combination of audio source localization, voice activity detection and face tracking technologies was realized in the developed multimodal infokiosk equipped by the standard means for information input/output (touch-screen and loudspeaker) and the devices for contactless HCI (microphone array and web cameras). This test-bed model is able to determine the client's mouth coordinates and to detect boundaries of speech signal appeared in the kiosk speech dialogue area. The model was used for cognitive evaluation of three types of user interfaces: a) a speech interface; b) a speech-and-text interface; c) text interface.

Experiments were performed by questioning users with help of different types of the interface. There were questions of two kinds: with some variants of answer and without them. Testing of the three variants of the interface was carried out by means of questions of the first category only. For a reception of a spontaneous answer from the user and assessing his/her behavior in the limits of a spontaneous interaction the second kind of the questions was used, by means of the text-speech interface. To define influence of experience on the subsequent interactions for different groups of users' sets of questions were alternated. All the informants were students. Each student had 20 questions to answer; the first 10 questions had variants of answer, and the last 10 implied spontaneous and long answers. The test bench asked the students in three modes: 1) question in a synthesized voice; 2) question in a synthesized voice, duplicated upon the screen with a text; 3) text only.

TABLE I
User behavior during the experiments

Symbol	Number of phenomena	Number of students
Question to the associates	24	15
Attempt to control the dialogue	6	4
Silence	37	15
Voiced pause	41	15
Thoughts aloud	82	22
Self-correction	17	8
Multiple pressing the button	3	3
Repeated answer	5	3

The students were distributed into three groups, 10 students in the group, and each group was questioned in one mode. The progress bar and announcement about speech recording were outputted to the display. The informants were not instructed about behavior, all the decisions were to be made in the course of the test. The informants were tested one-by-one and did not see previous sessions. During the experiments a constant record of answers and monitoring of button-pressing was made. Table 1 presents the types and number of phenomena (i.e. reaction of informants) registered during the test. As it is well shown in the table, a half of the students asked their associates for help — perhaps, they did not trust the computer completely or just could not find ways to ask the computer itself. It was very typical of situations of hesitations about modality choice ("Should I press *the button* here?"), type of required answer ("Should I *just name the number*, eh?") or when the informant just did not know what to do ("*What must I tell* if I know no answer?").

The majority of the students kept silence if they knew no answer. Sometimes the pause was vocalized by sustaining some sound (a vowel or a sonant), cough, laugh and so on. But if the informant knew the answer, it was given in no time. Sometimes the students expressed their thoughts aloud.

A few students acted fussily, they pushed buttons several times and repeated answers. It is to be noted especially, that the dialogue was "one-sided", i.e. the computer just received some information and confirmed it. The informant did not want anything from the machine, so he had no fear, that

interaction would not be very successful. Some students told after the test, that they were confused by a long time given for answering.

During the experiments an answer to question means, the answer of any type and by any means was received: by speech, by pressing buttons; for questions with variants repetition of a variant or naming only its number was allowed; or just a "dunno-answer". Unlike usual examinations or test, the students knowing the answer, recognized it without trying to think up something, or kept silence, expecting a following question. In questions with offered variants of answers uncertainty was expressed by words like, "appears", "it's something like" "maybe", etc. More tangled and dim answers were recorded. The majority of answers was not similar to short orders and commands, they were supplemented with other words, characterizing the degree of confidence to the machine, reflexions etc. Also for answers of users (especially at the speech interface) were peculiar if the question or variants of answers was badly remembered. Thus respondents did not look forward to hearing to these questions, and used them only as the discourses markers often used in dialogue between people.

V. Use Case: Smart Meeting Room

The developed smart room is intended for holding small and medium events with up to forty-two participants. Two groups of devices are used for tracking participants and recording speakers: (1) personal web-cameras serve for observation of participants located at the conference table; (2) four microphone arrays with different configurations and five video cameras of three types are used for audio source localization and video capturing of participants, who sit in rows of chairs in another part of the room.

In our research, three major types of conscious services are studied: (1) an active controlling PTZ camera to point on active speakers; (2) an automatic archiving of meeting data, including photos of participants' faces, video records of speakers, presentation slides and whiteboard sketches based on online context analysis; (3) selection and web-transmission of the most actual multimedia content during the meeting in the smart room. The meeting web-transmission system, which deals with the latter service and uses some results of other services, is considered here.

The developed meeting web-transmission system (MWTS) consists of five main software complexes and one control server. Figure 1 presents all six modules, which are marked by digits. The first complex is Multimedia Device Control System (MDCS), which joins modules that control all multimedia hard-software. This multimedia hard-software records behavior of participants and displays some presentation data. Second complex is Multichannel Personal Web Camera Processing System (PWPS), which captures and processes both audio and video streams from the personal web-cameras. The third complex stores the recorded audio and video data of the meeting in the smart room. The fourth complex is a database, which includes information about the meeting. Meeting Control Server (MCS) (№ 6 in Figure 1) receives and analyses data from all other modules and gives information about received data to displaying web-system (DWS) (№ 5 in Figure 1). DWS joins modules, which transmit multimedia content to remote participants. Content Management System (CMS) consists of third, fifth and sixth complexes.

The first complex MDCS is responsible for multimedia devises work. Sketch Board System (SBS) allows subjects to use the plasma panel with the touch screen for drawing and writing notes. Presentation Control is responsible for loading, displaying and switching presentation slides. Multichannel Sound Localization System (MSLS) gives information about audio activity in the smart room. Multichannel Video Processing System (MVPS) is responsible for processing and recording of video streams incoming from the cameras, which are focused on the auditorium, presenter and sitting participants in the zone of chairs.

MPVPS consists of client modules, such as PWC, which supports work of personal cameras located on the conference table, as well as PWPS, which processes data from the PWCS modules. Audio files in the *wav* format and video files in the *avi* format, which were received from the personal cameras and processed by the MCS (change of the format, resolution and file name) images from MVPS, PCS, SB and PWC are added to the file storage. The meeting database is realized by MySQL server and includes two tables: (1) basic information on all scheduled meetings and; (2) information about the current meeting, which includes some data for the meeting display system. DWS works as a web-page with several forms [25]. The data about form content are processed based on the AJAX technology. The transmitting of audio data to the client-computer is based on the RTMP stream server and the Abobe Flash technology. MCS receives and analyses data from all the modules, as well as chooses of audio and video content for DWS. This analysis is based on the logical-time model. Software modules of MWTS were installed on several personal computers joined in one local network, connection between them is based on transmitting messages in a string format by UDP packets.

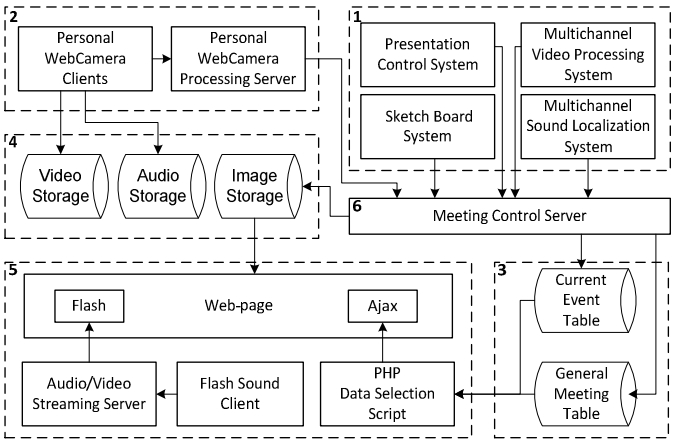

Fig. 1. Architecture of the meeting web-transmission system.

The work of the meeting web-transmission system and its components depends on the situation in the room. The component status and synchronization of audio and video content depend on the incoming events from the modules for audio localization, video monitoring, multimedia devices control. CMS manages by the multimedia content output, which is accessible for remote meeting participant. The events, which are generated by MCS and influenced on the meeting web-transmission system work, can be divided into four types by the following criteria: (1) by time; (2) by activity of the main speaker; (3) by activity of sitting participants; (4) by use of the presentation devices.

Experimental results were obtained with a natural scenario, where several people discussed a problem in the meeting room of 8.85x7.15x4.80m. One of the participants stayed in the presentation area and used the smart desk and the multimedia projector. Other participants were located at the conference table. The main speaker started his talk, when all the participants came together in the meeting room. Every participant could ask any questions after finish of the presentation. During the experiments the most of errors were made by the algorithm for detection of the active speaker, such errors occur when a participant at the conference table asks a question, but an image of other participant, which sits nearly, was displayed on the web-page. The accuracy of switching between the active participant and the presenter is higher. In total, about 97% of whole meeting time the graphical content were correctly selected at the analysis of the current situation in the meeting room.

VI. Conclusion

User profile and context modeling are the most important challenges of the ambient intelligence design. Development of the context-aware meeting processing systems gives appreciable benefits for automation of recording, archiving and translation of the meeting stream. The analysis of user behavior and multimedia equipment statuses is used for the context prediction and selection of audio and video sources, which transmit the most actual multimedia content for perception of the meeting and user provision with the relevant service. The developed meeting web-transmission system allows remote participants to perceive whole events in the meeting room via personal computers or smartphones. Further work will be focused on enhancement of abilities of remote participation during events in the intelligent meeting room and interaction with mobile information robot.

References

[1] T. Laakko, "Context-Aware Web Content Adaptation for Mobile User Agents," in *Studies in Computational Intelligence*, R. Nayak et al. (Eds.): SCI 130, Evolution of the Web in Artificial Intelligence Environments, 2008, pp. 69–99.

[2] C. Bolchini, C.A. Curino, E. Quintarelli, F.A. Schreiber, and L. Tanca, "A data-oriented survey of context models," *SIGMOD*, 36(4), 2007, pp. 19–26.

[3] A. Boytsov and A. Zaslavsky, "Extending context spaces theory by proactive adaptation," S. Balandin et al. (Eds.): *NEW2AN/ruSMART 2010*, LNCS 6294, Springer, 2010, pp. 1–12.

[4] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Proc. of the Workshop on Mobile Computing Systems and Applications*, Santa Cruz, CA, USA, 1994, pp. 85–90.

[5] D.R. Morse, N.S. Ryan, and J. Pascoe, "Enhanced reality fieldwork using hand-held computers in the field," *Life Sciences Educational Computing*, 9 (1), 1998, pp. 18–20.

[6] B. Moltchanov, C. Mannweiler, and J. Simoes, "Context-Awareness Enabling New Business Models in Smart Spaces," S. Balandin et al. (Eds.): *NEW2AN/ruSMART 2010*, LNCS 6294, Springer, 2010, pp. 13–25.

[7] K.H. Goh, J.Y. Tham, T. Zhang, and T. Laakko, "Context-Aware Scalable Multimedia Content Delivery Platform for Heterogeneous Mobile Devices," in *Proc. of MMEDIA 2011*, Budapest, Hungary, 2011, pp. 1–6.

[8] P. Dai, L. Tao and G. Xu, "Audio-Visual Fused Online Context Analysis Toward Smart Meeting Room," J. Indulska et al. (Eds.): *UIC 2007*, LNCS 4611, Springer, 2007, pp. 868–877.

[9] *Computers in the human interaction loop*. Ed. A. Waibel and R. Stiefelhagen, Berlin: Springer, 2009.

[10] G. Garau and H. Bourlard, "Using Audio and Visual Cues for Speaker Diarisation Initialisation," in *Proc. of ICASSP'2010*, 2010, pp. 4942–4945.

[11] Y. Rui, A. Gupta, J. Grudin, and L. He, "Automating lecture capture and broadcast: Technology and videography," *Multimedia Systems*, 10, 2004, pp. 3–15.

[12] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-Based Multimodal Speaker Detection for Distributed Meeting Videos," *IEEE Transactions on Multimedia*, Vol.10, No.8, 2008, pp.1541–1552.

[13] V. Rozgic, C. Busso, P.G. Georgiou, and S.S. Narayanan, "Multimodal meeting monitoring: Improvements on speaker tracking and segmentation through a modified mixture particle filter," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2007, pp. 60–65.

[14] V.C. Raykar, B. Yegnanarayana, S.R. Prasanna, and R. Duraiswami, "Speaker Localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, Volume 13, Issue 5, Part 2, 2005, pp. 751–761.

[15] T. Habib and H. Romsdorfer, "Concurrent Speaker Localization Using Multi-Band Position-Pitch (M-PoPi) Algorithm with Spectro-Temporal Pre-Processing," in *Proc. of Interspeech 2010*, Makuhari, Japan, 2010, pp. 2774–2777.

[16] J. Voordouw, C. Yang, L. Rothkrantz, and M. Capg, "A Comparison of the ILD and TDOA Sound Source Localization Algorithms in a Train Environment," in *Proc. of EuroMedia 2007*, Delft, 2007.

[17] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Proc. of Hands-Free Speech Communication and Microphone Arrays (HSCMA'2008)*, Trento, Italy, 2008.

[18] C. Datta, A. Kapuria, and R. Vijay, "A pilot study to understand requirements of a shopping mall robot," in *Proc. of HRI'2011*, 2011, pp. 127–128.

[19] M. Makatchev, I. Fanaswala, A. Abdulsalam, B. Browning, W. Ghazzawi, M. Sakr, and R. Simmons, "Dialogue Patterns of an Arabic Robot Receptionist," in *Proc. of HRI'2010*, 2010, pp. 167–168.

[20] M. Nieuwenhuisen, J. Stuckler, and S. Behnke, "Intuitive Multimodal Interaction for Service Robots," in *Proc. of HRI'2010*, 2010, pp. 177–178.

[21] A.C. Tenorio-Gonzalez, E.F. Morales, and L. Villaseñor-Pineda, "Teaching a robot to perform tasks with voice commands," in Grigori Sidorov, Arturo Hernandez Aguirre, Carlos Alberto Reyes Garcia (Eds.): *Proc. of the 9th Mexican international conference on Advances in artificial intelligence: Part I (MICAI'10)*, Springer-Verlag, 2010, pp. 105–116.

[22] G. Carrera J. Savage, and W. Mayol-Cuevas, "Robust feature descriptors for efficient vision-based tracking," in Luis Rueda, Domingo Mery, and Josef Kittler (Eds.): *Proc. of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications (CIARP'07)*, Springer-Verlag, 2007, pp. 251–260.

[23] A.C. Ramirez-Hernandez, J.A. Rivera-Bautista, A. Marin-Hernandez, and V.A. Garcia-Vega, "Detection and Interpretation of Human Walking Gestures for Human-Robot Interaction," in *Proc. of the 2009 Eighth Mexican International Conference on Artificial Intelligence (MICAI '09)*, IEEE Computer Society, Washington, DC, USA, 2009, pp. 41–46.

[24] V. Budkov, M. Prischepa, and A. Ronzhin, "Dialog Model Development of a Mobile Information and Reference Robot," *Pattern Recognition and Image Analysis*, Pleiades Publishing, Vol. 21, No. 3, 2011, pp. 442–445.

[25] V.Yu. Budkov, Al.L. Ronzhin, S.V. Glazkov, and An.L. Ronzhin, "Event-Driven Content Management System for Smart Meeting Room," S. Balandin et al. (Eds.): *NEW2AN/ruSMART 2011*, LNCS 6869, Springer-Verlag, 2011, pp. 550–560.