# An Audiovisual System of Monitoring of Participants in the Smart Meeting Room

Al.L. Ronzhin
SPIIRAS
39, 14th line, St. Petersburg, Russia
ronzhinal@iias.spb.su

**Abstract**

The problem of automatic event recording in smart meeting room is considered. The proposed monitoring system based on speaker localization and video tracking of meeting participants is implemented for recording of the current situation in the room and extends possibilities to automate audio-, video and presentation equipment control.

**Index Terms:** Audiovisual processing, ambient smart space, sound source localization, automatic camera pointing.

## I. Introduction

The main idea of smart space is recognition of the current situation, user's behavior analysis and satisfaction his/her requirements in natural and practically inconspicuous form. One of the examples of such smart space is a smart room, which equipped by software modules network, activation devices, multimedia toolkits and audiovisual sensors. Implementation of multimodal interface, which processes participant speech, motion, pose, gestures for detection their intentions and requirements allows us to create natural interaction with the smart room. The information about participant location in the room, their current activity and preferences, as well as role in the current event helps to automate work of integrated soft-hardware modules and provide well-timed control of multimedia and other devices [1]. Similar smart meeting rooms often operate in automanual mode, when hidden experts support performance of the smart space modules.

Choosing current active speaker and recording his/her activity during an event are the main tasks for meeting recording and supporting teleconference systems [2-3]. Panoramic and personal cameras could be employed for simultaneous recording of all participants. Such approach is suitable for small events, where all the participants are located at one table. Increase in participant number leads to space extension, which should to be processed, as well as cost of recording technical equipment.

Automatic analysis of audio and video data recorded during a meeting is not trivial task, since it is necessary to track a lot of participants, which randomly change position of their body, head and gaze. In order to detect participant activity several approaches based on using panoramic cameras, intelligent PTZ cameras, distributed camera systems were employed [4]. Besides video monitoring, motion sensors and microphone arrays could be implemented for detecting participant's location and selection of the current speaker. The sound source localization technique is effective for small lections or conference rooms. Personal microphones for all the participants or system of microphone arrays, which set on several walls of smart room, are employed for audio recording in medium rooms [1-2].

The developed smart room is intended for holding small and medium events with up to forty-two participants. Also there is the ability to support of distributed events with connection of remote participants. Two complexes of devices are used for tracking participants and recording speakers: (1) personal web-cameras serve for observation of participants, which are located at the conference table; (2) three microphone arrays with T-shape configuration and five video cameras of three types are used for audio localization and video capturing of other participants, which sit in rows of chairs in other part of the room. Status of multimedia devices and participant activity are analyzed for whole mapping current situation in the room [1].

## II. Functions and components of the monitoring system

The developed system of audiovisual monitoring of events consists of the four modules: (1) a multimodal control system of the smart room (MCSSR); (2) a multichannel system of personal web-cameras processing (MSPWCP); (3) a multichannel system of sound source localization (MSSSL); (4) a multifunction system for video monitoring (MSVM). Let us consider the description of each module, their features and joint work.

Figure 1 presents the MSPWCP module, which realizes multichannel audio video streams processing, coming from personal web-cameras, located on the conference table [4]. At first the search of face on the frames is made, if a face is detected then following tracking his position on the frame and web-camera pointing on the center of the face are carried out. The audio signal recorded by microphone of this camera is used for speech boundary extraction in multichannel audio stream. Time interval stamps, when face was detected (the buffer $B_1$) and speech was presented in audio channel (the buffer $B_2$) are accumulated for each camera. The data from the both buffers are used for video record segmentation of participant's presentation at the conference table.

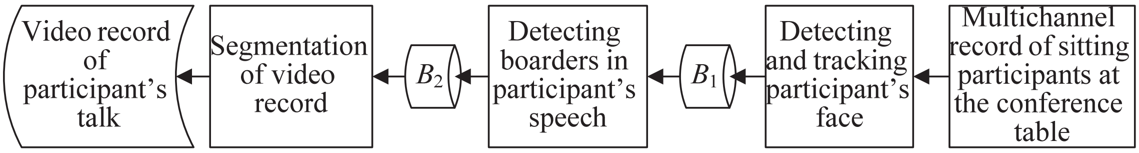
Video record of participant's talk
Segmentation of video record
$B_2$
Detecting boarders in participant's speech
$B_1$
Detecting and tracking participant's face
Multichannel record of sitting participants at the conference table

Fig. 1. Scheme of the MSPWCP module work

The MSSSL module captures audio streams from three microphone arrays and calculation coordinates of sound source in the room as well as estimation of speech message boundaries. If a signal was detected then the event $E_1$ is generated. The process of speech boundaries detection is begun by this event. The audio signal with recorded speech message is transmitted to the MCSSR module for following processing including recognition of speech commands for control by equipment in the smart room. The average coordinates of sound source obtained after processing streams from the three microphone arrays are sent to the MCSSR module and written in the buffer $B_3$. Figure 2 shows the MSSSL and MCSSR modules and them interaction.

The MCSSR module is intended for control of smart room devices, messages between other modules as well as expert support of the smart room performance. Location of participants, photos of the registered participants, location of the current sound source (talking participant), status of the devices, and current state of the event, recognized speech commands and other data useful for expert are displayed on the dialogue window of the module MCSSR. A buffer $B_4$ is refreshed when status of devices was changed by speech command, touch screen, remote web-interface or some other way. This buffer stores information about current status of used devices in the room. The speech commands, which were selected during analysis of speech activity in the room, are saved as audio files and its file name includes information about recording start time and location of sound source.

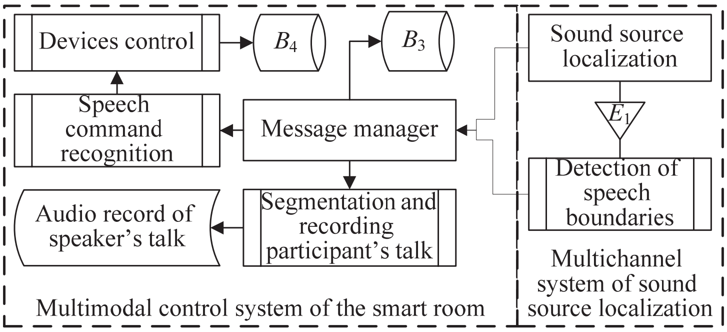
Devices control
$B_4$
$B_3$
Sound source localization
Speech command recognition
Message manager
$E_1$
Detection of speech boundaries
Audio record of speaker's talk
Segmentation and recording participant's talk
Multichannel system of sound source localization
Multimodal control system of the smart room

Fig. 2. Scheme of joint work of MSSSL and MCSSR modules

The MSVM module realizes video recording of whole event including recording of main speaker, active participants in auditorium during discussion and serves for automatic participant registration by set of the Internet-cameras AXIS [5]. Figure 3 presents three operations, which are sequentially performed in the main process of this module: (1) detection and tracking participants inside the room by processing of frames coming from panoramic ceiling camera mounted in the middle of the smart room; (2) detection of occupied sits by analysis of zone of chairs; (3) choose of video monitoring work mode based on analysis of data, which include information about participants location, status of group of lights (the event $E_2$) and audio activity in the room (the event $E_3$), these data are stored in the MCSSR module's buffers $B_4$ and $B_3$ accordingly.

One of the four sub threads (participant's registration, recording of active speaker in the zone of chairs, view on auditorium or main speaker) or their combination are started in accordingly with selected video monitoring work mode. The event $E_4$ launches the process of PTZ camera pointing on active speaker in the zone of chairs and following recording of his speech. The events $E_5$ and $E_6$ start sub threads of recording a main speaker and view on the auditorium accordingly.

The registration of participants in the zone of chairs is started after the event $E_7$. The buffer $B_5$ stores results of analysis of the zone of chairs, which contains numbers of chairs with sitting participants. If the occupied chairs are detected the sub thread of sitting participants face detection is launched. The intersection method of histogram comparison [6] is used for detecting of occupied chairs and face detection method based on AdaBoosted classifier [7] is used for checking presents of sitting participants. This sub thread carries out additional checking of sitting participant presence. When whole buffer $B_5$ is processed data about detected nonregistered participants are saved in the buffer $B_6$. The sub thread for participant's registration is launched in case of the event $E_7$

and presence of nonregistered participants. This sub thread controls of PTZ camera pointing on participant's faces and photographing them. There are two types of registration: passive and active, depending on the current video monitoring work mode. In first case only participant's photo is saved during the registration, in other case participant reports his\her main personal data in dialog mode. Thus, during active registration besides participant's photo, audio records with participant's main personal data are saved.

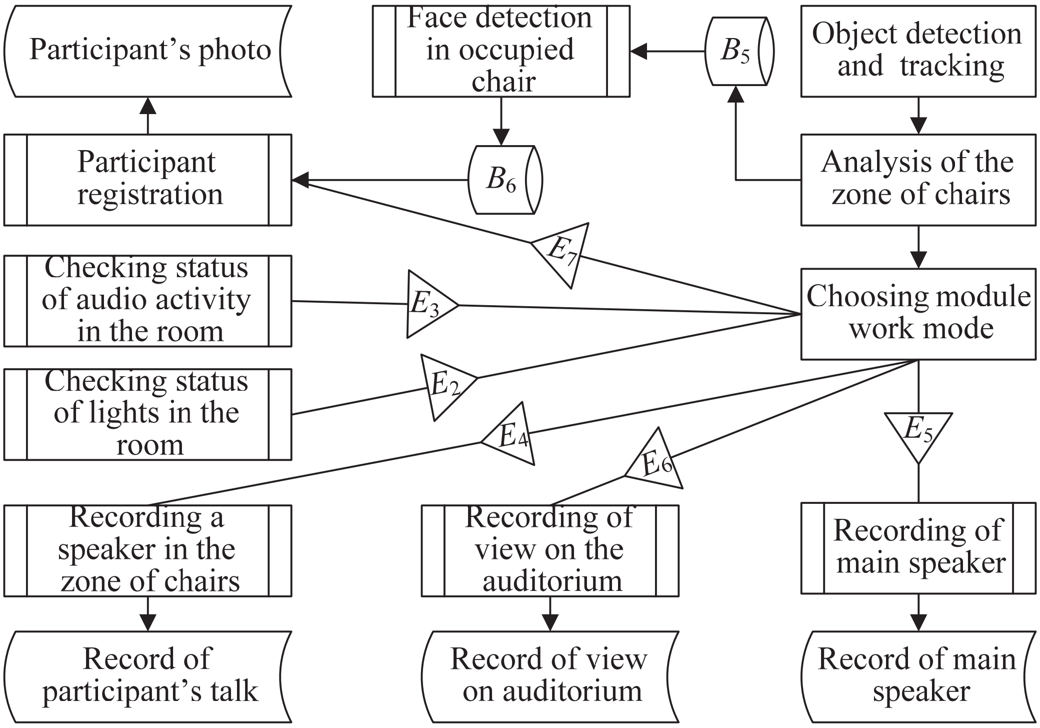
Participant's photo
Face detection in occupied chair
$B_5$
Object detection and tracking
Participant registration
$B_6$
Analysis of the zone of chairs
$E_7$
Checking status of audio activity in the room
$E_3$
Choosing module work mode
Checking status of lights in the room
$E_2$
$E_4$
$E_6$
$E_5$
Recording a speaker in the zone of chairs
Recording of view on the auditorium
Recording of main speaker
Record of participant's talk
Record of view on auditorium
Record of main speaker

Fig. 3. Scheme of the MSVM module work

The current version of the MSVM module works in the five video monitoring modes: observation, tracking participants, participant's registration, presentation and discussion. In first mode frames from the ceiling camera are only processed in the main thread. When a participant is detected in the room, the module comes to the tracking mode with recording of view on auditorium. If nonregistered participant is detected then registration mode is started. In the presentation mode tracking and recording participant located in the presentation zone is performed. In case of audio activity detection in the zone of chairs (the event $E_3$) the MSVM module changes work mode from the presentation to the discussion, in which sub thread of recording the active speaker in the zone of chairs is lunched. In this work mode, streams from the five video cameras installed in the smart room are processed in parallel by the five sub threads of the MSVM module, which were previously described. The developed algorithm of choosing the video monitoring mode and methods of video streams processing are described in [5].

During meeting processing the audiovisual monitoring system creates the six databases, which store audiovisual information about event and meeting's participants, who sit at the conference table and in the zone of chairs. The recorded databases are used for event annotation and generation of multimedia content of web-system for teleconference support [8].

## III. Testing of the system in the smart room

The developed audiovisual monitoring system is a part of the smart room technological platform, which includes complex of soft- hardware modules, and multimedia and information streams between them are performed on several multiprocessor computers. The two techniques, which are oriented on functional test and estimation of system performance quality, were proposed for testing the developed technological platform.

The technique of functional test is based on sequential query of all the software modules and if answering signal does not return in the appropriate time interval then a recovery procedure restarts the dropped module and devices connected to it. The technique of estimation of system performance includes the criteria presented in Table 1 and based on calculation statistic errors *FA* and *MR*. The estimations are calculated after the end of event with manual checking the recorded audiovisual data.

TABLE I
List of criteria estimation of system work quality

| Criteria | Formula |
|---|---|
| $A_p$ – quality of participants detection in the smart room | $A_p = \frac{N_p - N_{FA\_p} - N_{MR\_p}}{N_p}$, where $N_p$ is the maximum number of participants inside the smart room, $N_{FA\_p}$ is the number of falsely detected participants, $N_{MR\_p}$ defines the number of missed participants |
| $A_{o\_ch}$ – quality of occupied chair detection | $A_{o\_ch} = \frac{N_{ch} - N_{FA\_ch} - N_{MR\_s\_p}}{N_{ch}}$, where $N_{ch}$ is the number of chairs installed in the room, $N_{FA\_ch}$ is the number of falsely detected occupied chairs, $N_{MR\_s\_p}$ is the number of missed chairs with sitting participants |
| $A_{s\_p}$ – quality of detection of sitting participants in the zone of chairs | $A_{s\_p} = \frac{N_{ch} - N_{FA\_p\_f} - N_{MR\_p\_f}}{N_{ch}}$, where $N_{FA\_p\_f}$ is the number of false detected participant's faces in the zone of chairs, $N_{MR\_p\_f}$ is the number of missed participant's faces in the zone of chairs |
| $A_m$ – quality of the work mode | $A_m = \frac{N'_m}{N_m}$, where $N_m$ is the number of work mode changing, during an event, $N'_m$ is the number of correctly selected work modes |
| $A_{m\_s}$ – quality of camera pointing on speaker in the presentation zone | $A_{m\_s} = \frac{N'_{m\_s}}{N_{m\_s}}$, where $N_{m\_s}$ represents the number of frames, which were recorded, when speakers were in the presentation zone, $N'_{m\_s}$ is the number of frames with a speaker |

The estimation of video monitoring system work quality based on recordings of five events in the smarts room with total number of one hundred ten participants. After manual analysis of recordings average estimations were calculated:

$$\overline{A}_p = 88\%,\ \overline{A}_{o\_ch} = 89\%,\ \overline{A}_{s\_p} = 91\%,\ \overline{A}_m = 97\%,\ \overline{A}_{m\_s} = 90\%.$$

## IV. Conclusion

The audiovisual monitoring system of participants was developed for automation of devices control and recording events in the smart room. It consists of the four main modules, which realize multichannel audio and video signal processing for participants localization, detection of speakers and recording them. The proposed system allows us to automate control of audio and video hardware as well as other devices installed in the smart room by distant speech recognition of participant command. The verification of the system was accomplished on the functional level and also the estimations of detection quality of participants, work modes and camera pointing on speaker after some events was calculated.

## Acknowledgment

This work is supported by The Russian Program “Research and Research-Human Resources for Innovating Russia in 2009-2013” (contracts #П2360, #14.740.11.0357), RFBR (projects 10-08-00199-a) and Nokia.

## References

[1] R.M. Yusupov and A.L. Ronzhin, “From smart devices to smart space,” *Herald of the Russian Academy of Sciences, MAIK Nauka*, vol. 80, num. 1, pp. 63–68, 2010.

[2] B. Erol, Y. Li, “An overview of technologies for e-meeting and e-lecture,” *Proc. IEEE International Conference on Multimedia and Expo.*, pp. 6-12, 2005.

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

[4] R. Rienks, A. Nijholt, P. Barthelmess, “Pro-active meeting assistants: attention please!,” *AI & Society, Springer*, vol. 23(2), pp. 213-231, 2009.

[5] Al.L. Ronzhin, M.V. Prischepa, A.A. Karpov, “A Video Monitoring Model with a Distributed Camera System for the Smart Space,” *Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2010*, LNCS 6294, pp. 102-110, 2010.

[6] B. Schiele and J. L. Crowley, “Object recognition using multidimensional receptive field histograms,” European Conference on Computer Vision (vol. I, pp. 610–619), April 1996.

[7] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In Proc.of IEEE ICCV, pages II: 734–741, 2003.

[8] A.L. Ronzhin, V.Yu. Budkov, A. Karpov. “Multichannel System of Audio-Visual Support of Remote Mobile Participant at E-Meeting” *Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2010*, LNCS 6294, pp. 62–71, 2010.