Multimodal Analysis for Universal Smart Room Applications

Lutz Goldmann, Amjad Samour, Thomas Sikora

Technical University of Berlin, Communication Systems Group
Einsteinufer 17, 10587 Berlin, Germany

Abstract

Abstract This paper presents a multimodal system incorporating smart room technologies (SRT) for conference room applications. Although, the audio-visual analysis requires only rather basic equipment, the system works reliably and supports various applications such as recognizing persons using different modalities, localizating visible speakers, controlling the camera view and summarizing the AV data.

1 Introduction

Recent developments in communication technologies have lead to exciting applications that might change the way people communicate and interact. One application that recently gained significant attention are multimodal, unobtrusive smart room technologies (SRT). The goal is to monitor persons based on audio-visual information in order to infer their location, identity, and behaviour in an environment. Due to the multiple modalities used for the analysis, research in diverse topics including object detection and tracking, visual person recognition, speech detection, and speaker recognition is involved.

In this paper, a multi-modal approach providing smart room technologies for talks(presentations, discussions) is proposed. In contrast to other systems [1,10] only rather basic equipment is neccessary, allowing for an easy and portable setup. While Busso et al. [1] utilize multiple calibrated and synchronized cameras and microphone arrays, this system already operates with a single camera and a single microphone. Nevertheless, for some applications two cameras, a static camera and a pan-tilt-zoom (PTZ) camera, are required. Although the use of this basic equipment provides less information than complex equipment the proposed system works reliably. It allows to detect and recognize visible persons as well as detect speech and recognize the speakers. Furthermore the speakers can be located in the video and a PTZ camera can be controlled if required.

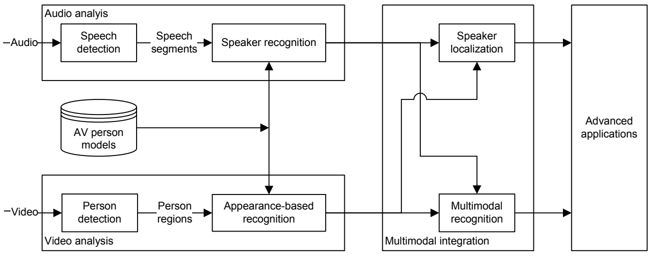


Fig. 1 System overview.

2 System overview

Figure 1 gives an overview of the system which mainly consists of three parts: audio analysis, video analysis, and multimodal integration. The audio analysis itself is composed of speaker detection and speaker recognition while the video analysis consists of person detection and appearance-based recognition. Within the multimodal integration part, the audio and video information is combined in different modules based on the application, namely multimodal recognition and speaker localization. Furthermore different advanced applications, such as camera control, summarization, and indexing & retrieval are supported.

2.1 Audio analysis

Speech detection: The goal of the speech detection stage is to separate human speech from other audio data, such as silence, music, and environmental sounds. Therefore, different features such as short time energy (STE), high zero crossing rate (HZCR) and band periodicity (BP) are extracted. The speech is distinguished from the others using a rule-based classifier that exhibits well-known characteristics of these different types of audio data in a heuristic way [5].

Speaker recognition: In the speaker recognition stage the speech segments are classified as belonging to specific speakers. This is based on a supervised learning approach, where pretrained models for each speaker are compared to the segments and the speaker with the highest probability is assigned to this segment. Mel-frequency cepstrum coefficients (MFCC) and their derivatives [8] are used for describing the cepstral properties of each speaker's voice. A Gaussian mixture model (GMM) [2] is trained for each speaker and the recognition is based on the maximum a posteriori (MAP) criterion [8].

2.2 Video analysis

Person detection: Since a smart room environment is a naturally quite static, it is possible to use a static camera for capturing the video. Thus, background differencing techniques can be used to segment foreground objects such as persons from the background.

The approach of Horprasert et al. [3] was adopted since it is quite robust and its complexity is rather low in comparison to other approaches [4], allowing for real-time performance. The background is modeled using a special model that separates luminance and chrominance information. Based on this model the luminance and chrominance distortion of an unknown pixel are calculated and used for classifying this pixel as either foreground, background, highlight or shadow.

The resulting binary foreground mask passes through a connected component labeling stage resulting in objects consisting of single connected blobs. Persons are detected by applying heuristic rules to these blobs based on size and shape criteria.

Appearance-based recognition: Visual person recognition can be based on both biometric (face, gait) and non-biometric (appearance) features. Since biometric features might not be available due to the resolution or content of the video, appearance based features are used in this system.

For describing the appearance of persons color and texture features are suitable [7]. In order to allow for real-time performance, the following decriptors were choosen: the average RGB color (AvgRGB) for describing the color and the intensity histogram based features (IH) [6] for describing the texture.

Following the idea of a supervised learning approach Gaussian mixture models (GMM) [2] are adopted. The results of different descriptors are fused using post mapping fusion (product rule) [9] and the person is recognized using maximum a posteriori (MAP) criterion.

2.3 Multimodal integration:

Audio and video can be integrated in different ways depending on whether the information is used sequential or parallel. The first case can be called *assistance* since one modality provides information to the other modality in some task. The latter one is the actual *fusion* where both modalities are combined in order to improve the performance in comparison to each modality alone.

Speaker localization: The goal of speaker localization is to determine the position of the actual speaking person within the environment. Using just a 2D environment the task is to choose one specific person (the person the speech belongs to) out of all visible persons in the video. The key for that lies in the audio-visual person model which allows to establish a correspondence between visible and speaking persons. Since audio is restricted to just one speaking person and the video may contain multiple persons, the audio assists the video in that matter. Given a person recognized by the audio, its ID is compared to the ID's of the persons recognized from the video. If the person can be found, it is considered visible and its position can be obtained from its visual description. If the person cannot be found, it is considered invisible and might belong to the audience. Both informations can be used for advanced applications.

Multimodal person recognition: Assuming that only one person is visible in the video and this person is the actual speaker, audio and video information can be fused in order to improve the reliability of the person recognition. Post mapping fusion [9] is applied in our system based on the outcome of the speaker recognition module and the visual person recognition module. Depending on the used classifiers, decision (hard decision) or opinion level (soft decision) concepts can be used. The GMMs used for audio- and video-based recognition provide probabilities for each person. These opinions are fused using unweighted product rule and a combined decision is made using (maximum a posteriori) MAP criteria.

2.4 Advanced applications:

Based on the information provided by audio and video analysis as well as the multimodal integration several applications can be derived

Camera control: If two cameras (a static and a PTZ camera) are available the PTZ camera can be controlled by the static camera using the results of the speaker localization module. If the speaker is known and visible it might zoom in to establish a close-up view of his face. On the other hand if the speaker is known and not visible the PTZ camera might pan and tilt to find the speaker in the audience.

File	Audio	Video	Multimodal
S1	97.9	100.0	99.5
S2	98.0	100.0	99.9
S3	100.0	88.8	96.5
All	98.4	96.8	99.0

Table 1 Recognition rates (in %) of speaker recognition, appearance-based recognition and multimodal recognition for different single person videos.

Summarization: Especially for meetings or panel discussions where multiple persons discuss a topic a summary of the activities is useful. Interesting facts are the number of speakers, the length of the speech segments, their visual appearance (face), and position within the environment (podium, audience). All this information can be provided as a short audio-visual summary.

Indexing & retrieval: When analyzing AV data from talks (presentations, discussions) one might be interested to find the audio, visual or audio-visual presence of a certain person. Another issue is the cooccurence of multiple visible persons at the same time.

3 Experiments

Since no suitable data was available a new database was built. It consists of audio-visual data captured using a standard MiniDV camcorder. For training the audio-visual person models, videos containing single persons giving a presetnation are used. For testing, videos with multiple persons in presentations or discussions are considered. The database contains about 30 minutes of material and 3 different persons.

Figure 2 shows a typical example of the system applied to one of the captured discussion scenarios. Figure 2(a) shows the results of the speaker recognition. Figure 2(b) shows the result of the appearance-based recognition, which recognizes all persons correctly. The result of the speaker localization based on the previous steps is shown in figure 2(c). The actual speaker is marked in “white” and the non-speaking persons in “black”. Finally, a simulated view of the PTZ camera obtained by translation and zoom is shown in figure 2(d).

Table 1 shows the results for the person recognition within the single person videos. It shows the recognition rates (RR) for the audio-based (speaker recognition) and video-based recognition (appearance-based recognition) as well as the results obtained using multimodal fusion. In general the results are quite encouraging with overall recognition rates of 98.4% and 96.8% for audio and video respectively. Furthermore, its interesting to see the complementary results for the audio and video part. Thus an improved overall recognition rate with 99.0% is achieved using the multimodal approach.

File	Speaking				Visible		
	Amj.	Lut.	Mus.	Sil.	Amj.	Lut.	Mus.
S1	66.7	1.7	0	31.7	93.5	0	0
S2	0	61.2	1.3	38.6	0	92.4	0
S3	0	0	49.5	50.5	9.9	0	78.3
M1	33.4	26.3	27.3	12.9	82.1	78.8	26.0
M2	34.1	39.3	15.5	11.1	70.4	59.0	60.8
All	29.6	28.3	21.1	21.1	10.7	12.3	7.7

Table 2 Visibility and speaker proportions (in %) of each persons in different videos and all videos together.

One of the advanced applications based on the multimodal analysis is illustrated in table 2. It gives a summary of the videos by providing the proportions, when a certain person is visible or speaking. This may be used to measure the level of participation for each person, to identify the roles of different persons within a discussion, or to discover the type (presentation, meeting) of the talk. Some examples:

- Video “S1”: “Amjad” is the dominating speaker and also the only visible person over the whole video, which leads to the conclusion that this video contains a presentation of “Amjad”.
- Video “M2”: All persons are visible for a long time during the video, which is typical for discussions. The speaking proportions show that “Amjad” and “Lutz” are quite active, while “Mustafa” is more passive.

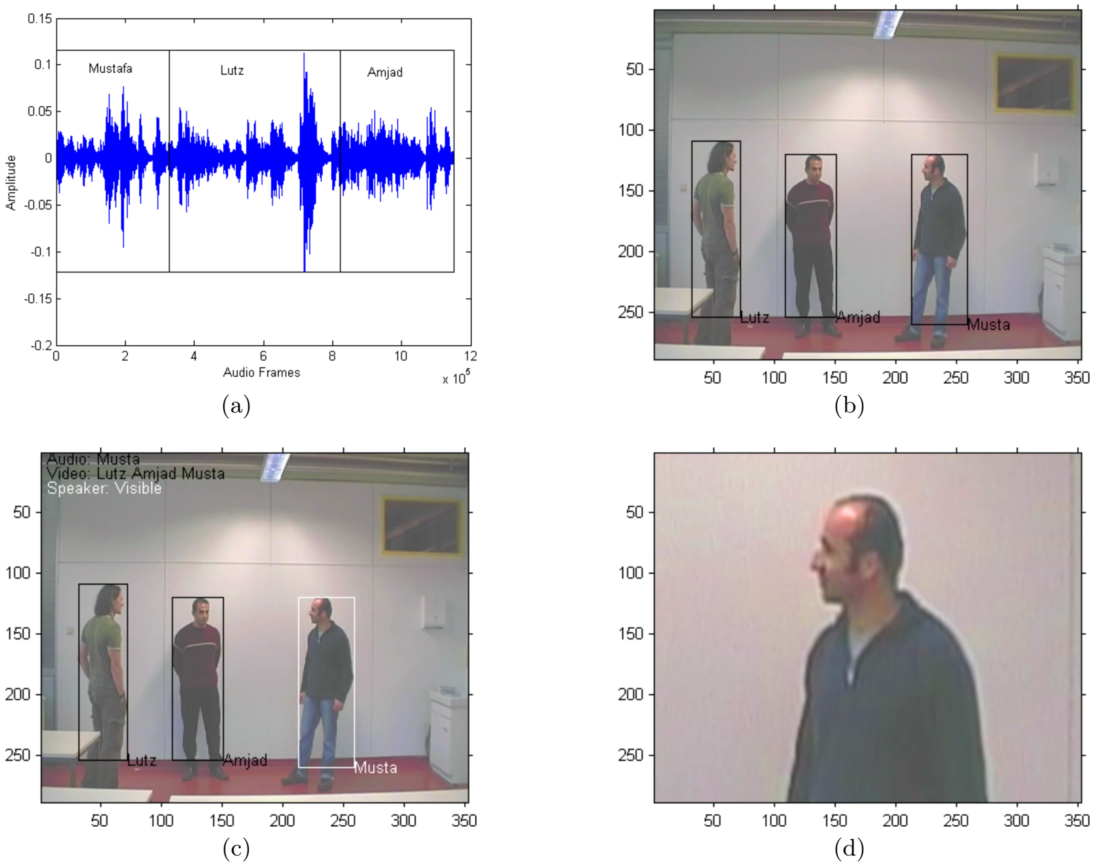
4 Conclusion

The proposed system reliably recognizes visible persons, detects speakers and localizes them within the 2D environment. Although it lacks some features of systems that use more complex equipment, such as 3D position and different viewing angles it allows to extract useful information for various applications.

An issue for future work is the use of a more comprehensive databases including more persons, different types of talks, different environments and a longer duration. At the moment the system does not use an object tracking module, which may improve the results by exploiting temporal constistencies. In order to distinguish a higher number of persons, a face recognition module will be considered.

At present the whole system is based on a supervised learning approach. Thus only previously trained persons can be recognized and located. In order to make the system more flexible, unsupervised and incremental learning approaches will be investigated.

Acknowledgement: The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content – K-Space.



(a) (b)

(c) (d)

Fig. 2 Examples illustrating the different parts of the overall system: (a) speaker recognition, (b) appearance-based recognition, (c) speaker localization, (d) PTZ camera control (simulated).

References

1. C. Busso, S. Hernanz, C.-W. Chu, S.-I. Kwon, S. Lee, G. Panayiotis, I. Cohen, and S. Narayanan. Smart room: Participant and speaker localization and identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.
2. R. O. Duda, H. P., and E. Stork. *Pattern Classification*. Wiley, 2001.
3. T. Horprasert, D. Harwood, and L. Davis. A statistical approach for real-time robust background substraction and shadow detection. Technical report, Computer Vision Lab., University of Maryland, USA, 1999.
4. M. Karaman, L. Goldmann, D. Yu, and T. Sikora. Comparison of static backgorund segmentation methods. In *Proceedings of Visual Communications and Image Processing (VCIP)*, 2005. Published.
5. L. Lu, H. Jiang, and H. J. Zhang. A robust audio classification and segmentation method. In *ACM Multimedia*, pages 203–211, 2001.
6. A. Materka and M. Strzelecki. Texture analysis methods – a review. COST b11 report, University of Lodz, 1998.
7. C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full body person recognition system. *Pattern Recognition*, 2003.
8. D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72–83, 1995.
9. C. Sanderson. *Automatic Person Verification Using Speech and Face Information*. PhD thesis, Griffith University, Queensland, Australia, 2002.
10. M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darell. A multi-modal approach for determining speaker location and focus. In *International Conference on Multimodal Interfaces*, 2003.