Audio-Visual Feature Extraction for Semi-Automatic Annotation of Meetings

Marián Képesi,
Michael Neffe
and Tuan Van Pham
SPSC Laboratory
Graz University of Technology
Inffeldgasse 16, 8010, Graz, Austria
Email: kepesi@tugraz.at,
michael.neffe@tugraz.at,
v.t.pham@tugraz.at

Michael Grabner
and Helmut Grabner
Inst. for Computer Graphics and Vision
Graz University of Technology
Inffeldgasse 16, 8010, Graz, Austria
Email: mgrabner@icg.tu-graz.ac.at,
hgrabner@icg.tu-graz.ac.at

Andreas Juffinger
Inst. for Theoretical Computer Science
Graz University of Technology
Inffeldgasse 16b, 8010, Graz, Austria
Email: andreas.juffinger@tugraz.at

***Abstract*— In this paper we present the building blocks of our semi-automatic annotation tool which supports multi-modal and multi-level annotation of a meeting database. The main focus is on the proper design and functionality of the modules for recognizing meeting actions. The key features, identity and position of the speakers, are provided by different audio and video modules. Three audio algorithms (Voice Activity Detection, Speaker Identification and Speaker Position Estimation) and three video modules (Detection, Tracking and Identification) form the low-level feature extraction components. Low-level features are automatically merged and the recognized actions are proposed to the user by visualizing them. The annotation labels are related but not limited to events during meetings. The user can finally confirm or if necessary, modify the suggestion, and then store the actions into a database.**

I. Introduction

The computational analysis of actions and interactions in meetings is a relative young discipline based on the results of social psychology and social network analysis. These disciplines study the influence of individuals on other individuals and groups. For the meeting scenario the influence of certain participants on decision making as well as the amount of attendance and contribution in certain topics is of primary interest to automatically identify experts and to understand the information flow within meetings. The knowledge about these things is highly important to begin optimizing the social network connectivity or to offer services like expert finding and meeting moderator selection. Basic elements, on which these high level behaviors relay on are active attendance of individuals such as speaking or not speaking, body postures while speaking or listening, communication patterns and spatial information for clique identification.

It is possible to derive different group behaviors from this information such as monologue, presentation and discussion [1] as well as attentional cues, which can be further inferred from gazing behavior and body postures [2]. In the Mistral Project [3], where we aim to identify actions, it is therefore necessary to extract low-level features such as the position of active and passive persons and their head position, voice activity, gender information and direction of arrival of speech signal. High-level features such as multi-modal person index, multi-modal localization and attendance are then derived from these unimodal low-level features.

To extract and learn actions, it is important to build a database for learning and evaluation of different methods. To produce a huge amount of annotated data, it is necessary to have a tool which supports fast annotation of multi-modal data on different levels - a semi-automatic, multi-modal and multilevel annotation system. The annotated data can then be used to discover higher level concepts and the evaluation of methods for group action learning and social network analysis. Such an annotation system tackles all our needs and we strongly believe that it would be very useful for the multi-modal information retrieval community.

A number of different systems for digital video and audio annotation have been proposed and implemented. In the first step these tools have been evaluated. For example the VideoAnnEx Tool [4] supports a wide range of functionality like static scene description, key object description, event description in one level only. Actions in meetings are of different granularity so it is necessary to annotate actions on multiple levels. Therefore it is essential for an annotation tool to support the creation of overlapping and non-overlapping segments. The ELAN Tool [5] supports multilevel segmentation and annotation as well as transcription of videos, but does not support spatial and spatio-temporal annotation of videos. Other video annotation tools like [6]–[9] provide similar functionality sets but none of them is intended for semi-automatic video annotation.

This paper aims to introduce the modules of a semi-automatic tool to annotate actions in the meeting database of the Mistral Project. Our main goal is to design a tool for semi-automatic transcription of an action-based meeting database based on features, learned by visual and acoustic processing of meeting recordings. The low-level features are then provided automatically to the user as suggestions.

The paper is structured as follows: first in section II the architecture of the tool and its structure is described followed by section III describing the Audio extraction modules and section IV describing the Video extraction modules. A short description of an acted meeting dataset and the visualization of the results from the core modules (audio and video) is given in section V.

II. Annotation Tool Architecture

The primary setup of the meeting data which is going to be annotated with the tool is designed for, but not limited to, the setup in Fig. 1. A linear microphone array provides an acoustic view of 180°, while the video stream contains information about the speakers sitting in front of the camera within a camera dependent angle.

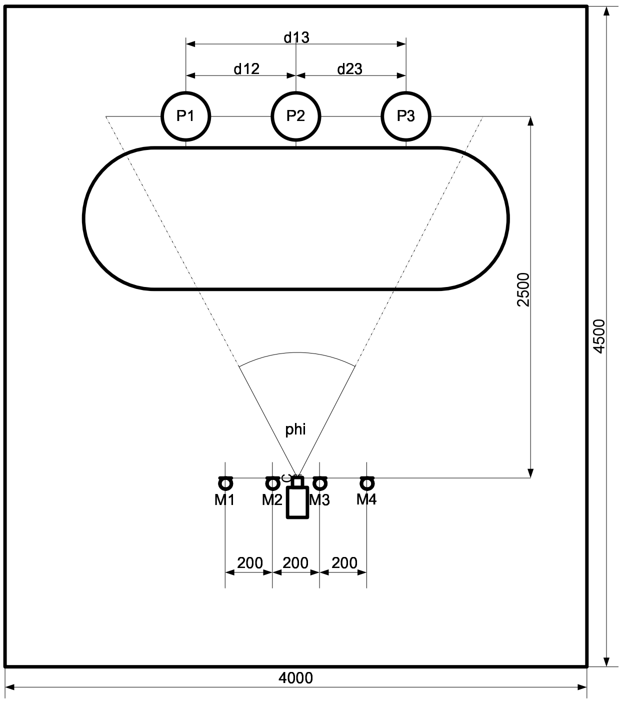

Fig. 1. Meeting Scenario Setup: P1, P2, and P3: initial location of participants, M1-M4: microphone array, distances in millimeters.

For merging the unimodal information derived from video and audio it is necessary to align them in time and space. For the time alignment we will provide a synchronization tool where the annotator can align the different streams by selecting a certain event in the video stream and the appropriate noise. The spatial alignment between pixels in a video frame and the azimuth derived from the DoA method during voice activity can be calculated as follows:

$$l_x = \frac{res_x}{2} + \frac{tan(\alpha_{DoA})}{tan(\varphi/2)} * res_x$$

where l_x corresponds to the xth column in the image, α_{DoA} is the angle derived from the DoA audio module, and res_x states the horizontal camera resolution. Due to this closed analytical form it is possible to calculate pixel information from the azimuth and vice versa, so that the user of the system will be able to select the preferred information.

Figure 2 shows the basic modules of the annotation tool, whereby the user interface is the central point of the system. Within this module it will be possible to load different video and audio streams as well as the action lexicons (sets of predefined actions of interest) and participant (needed to load facial and speech models). The results of main unimodal modules, described in detail in the next sections, will be visualized to support fast annotation and to give recommendations to the user about certain actions of persons such as P2 is speaking ”at 5°” or ”at 340x270”, respectively. These recommendations can then be accepted or corrected by the user. The resulting annotated data can be used by the underlying modules as labeled data to improve their performance as well as basis for the high-level features to learn actions such as agreement, disagreement, dialog and monologe.

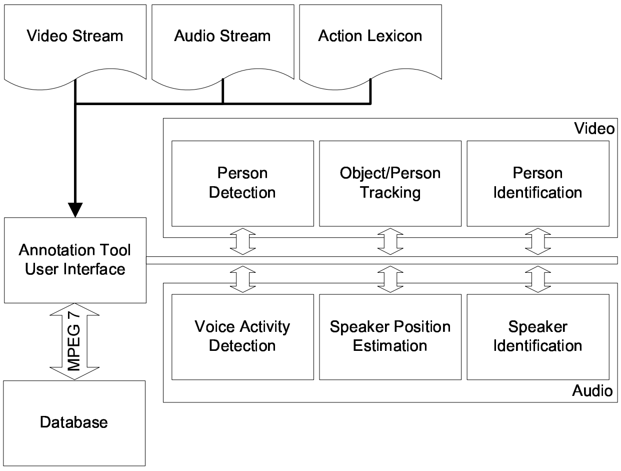

Fig. 2. Annotation Tool Architecture: Schematic description of interactions between the different modules, the user interface and the database

III. Audio Modules

The audio module is composed of three extraction units: Voice Activity Detection (VAD), Speaker Identification (SI) and Direction of Arrival (DoA) estimation unit.

A. Voice Activity Detection

Many voice activity detection techniques have been proposed in the last decades with the common methodology being to draw some features from the input speech frame and then to take a hard decision against predetermined thresholds. In this work, an energy-based VAD unit is applied with blockwise update of background noise-level and the 15ms/200ms conversation rule [10]. The method from Aurora group [11] is used by employing the quantile technique to estimate the noise energy for threshold calculation. The VAD module serves as a feeding block for both audio (SI and DoA Estimation modules) by preparing only cut speech segments from the marked piece of recording.

B. Speaker Identification

In the last decades it had been shown that for Speaker Identification tasks the universal background model (UBM) approach is to favor among others. To overcome the problems of mismatch between UBM and real speech recorded by the microphone array the approach of [12] has been applied. In this case an UBM model is used that is built directly from the data of a speaker but with a reduced number of Gaussian components compared to that of the built speaker model. To account for the acoustic mismatch caused by the changing influence of the environment on the recorded speech two normalization approaches have been tested. The first is the well known cepstral mean subtraction and variance normalization, where the mean removes the global shift affecting the cepstral coefficients and variance normalization compensates the main effect of the acoustic distortion. Nonlinear effects can't be treated by this method. The second method called histogram equalization [13] is often used in digital image processing. Recently it has been also adapted for speech and speaker recognition. This method transforms the histogram of each feature vector independently onto a reference histogram which is to be assumed Gaussian in the experiment. Non of these two methods outperforms the other for this setup. As features 20 mel-warped frequency cepstral coefficients (MFCCs), augmented by its corresponding delta and delta delta MFCCs are used.

C. Direction of Arrival Estimation

The speakers position is estimated by using a linear microphone array with four omni-directional electret microphones spaced at 20cm from each other. At first, cross correlation matrix is calculated for 3 pairs of microphones, using speech segments of 300msec with an overlap of 10msec, one matrix per each VAD segment. The dominance of main correlation peaks is evaluated by normalizing the peak value by the square of its width. The dominance value of each peak is then thresholded, and only surviving peak positions are processed further. Candidate correlation peaks corresponding to (-90° to +90°) in azimuth are considered, and a histogram of candidates with 3° resolution is created. Three main candidate positions are kept for each microphone pairs, with a certain tolerance in azimuth, while all the other candidates are dropped, unless the speakers seem to move continuously during the VAD region. Finally candidates are cross-checked at each time-point by searching the same candidate in each of the three channels. Winner DoA candidates at a certain time instance are the ones with the highest probability being present in all of the three channels. The DoA is finally stored into two hierarchical levels: 1) High-level localization: one DoA value per VAD extracted from the main histogram peak position and decoded as azimuth. 2) Detailed DoA description: each segment of a VAD region with a resolution of 10ms is stored with corresponding DoA values. This yields 20-200 values per VAD.

D. Evaluation

The performance of the VAD and azimuth estimation (DoA) algorithms has been presented in previous work [14] where the VAD was evaluated by using Precision and Recall, the two measures often used in the domain of Information Retrieval (IR) [15], while the DoA algorithm was evaluated by its correctness in percent. The performance of the VAD method on noise-free acoustic data showed a recall of 91% and a precision of 79 %, while the DoA correctness was 86 %. We need to note, that since the DoA relies on the VAD estimation, its performance strongly depends on the VAD performance. DoA precision dramatically drops by VAD insertions.

The performance of the Speaker identification module has been tested with pretrained data using 30 seconds of speech for each speaker. Each UBM had been modeled by 3 and each speaker model by 38 Gaussian components. The Speaker detection task in this first step was performed on the VAD output. The segment length analyzed had a mean duration of 1.83 sec, the min/max segment length was 20msec/9.4sec. The recognition rate for the given setup was 77 %.

IV. Video Modules

The video module provides information about the actual position (xy coordinates) and also about the identity of the participants. For reliably extracting this information we are using 3 different components. First, a detector which has been trained off-line for detecting frontal views of faces, is applied to each frame. Second, in order to obtain trajectories and to obtain continuous information about the location even when we can not see the frontal view, a tracker is initialized with each detection which can then track the head independent from its pose. Finally, video applies recognition on the trajectories for providing person's identity information to the semi-automatic annotation tool. Note that all three components can process the video information in real-time.

A. Detection

Due to its popularity and power we use the classical idea of Viola and Jones [16] to off-line train a classifier for detecting objects. In this case a large variety of objects can be learned when a large number of labeled training samples is available. The main assumption is that a set of generic features can separate an object category from the background. This feature selection is done by boosting. Once the detector is trained, it is simply evaluated by an exhaustive search at many possible positions and scales on the image, which is not costly anymore because of the use of integral data-structures for feature extraction.

B. Tracking

The main idea is to formulate the tracking problem as a binary classification task as proposed by [17]. Robustness to changes in appearance of the target object (in our case of the tracked head) is achieved by continuously updating the classifier. Once the target object has been detected, it is assumed to be a positive image sample for the tracker. At the same time negative examples are extracted by taking regions of the same window size from the surrounding background. These images are used to make several learning iterations of on-line Adaboost (based on [18]) in order to find an initial ensemble of features. Note that these iterations are only necessary for the initialization of the classifier.

The tracking step is based on the classical approach of template tracking [19]. We evaluate the current classifier at the surrounding region of interest and obtain for each sub-patch a confidence value. Afterwards we analyze the obtained confidence map and shift the target window to the new maximum location. Next the classifier has to be updated in order to adjust it to possible changes in appearance of the novel view of the head and to become discriminative to a different background. The current target region is used for a positive update of the classifier while again surrounding regions represent the negative samples. This update policy has proved to allow stable tracking in natural scenes. As new frames arrive, the whole procedure is repeated and the classifier is therefore able to adapt to possible appearance changes and in addition becomes robust against background clutter. Note that the classifier adapts to the possible appearance changes of the head while at the same time tries to distinguish it from its surrounding background.

C. *Identification*

Identification of participants is based on the idea of the well known Scale-Invariant Feature Transform (SIFT) approach [20]. A grid of orientation histograms is applied at each frame to the tracking window for extracting a descriptor of the specified object which is then matched to a reference descriptor which has been computed off-line from sample images of the object. If the distance of the best match is lower than a certain threshold, the recognizer marks the tracked object as identified.

V. Meeting database

While there has been a huge amount of meeting data produced in different projects, like AMI (Augmented Multiparty Interaction) and ICSI (International Computer Science Institute) with the focus on automatic speech transcription, mostly by using close-enough mounted microphones. Our task is different: The main goal is a distant-acquisition-based meeting indexing without using any Automatic Speech Recognition (ASR) module for speech-to-text transcription. We are interested in learning and transcribing actions, which could be easily acquired by a distant camera and a distant microphone array. The recordings used for this work are part of the meeting database of the Mistral Project [3] acquired in different meeting rooms with different acoustic complexity (room size, reverberation level, noise type and level). To evaluate the core units we used recordings made in a small library - in order to use the room with the lowest echo level. The audio data was recorded by the *RME Fireface* external sound card using a 4-channel linear microphone array, sampled by 48kHz. In parallel, a 20fps video signal has been captured by a Webcam with a resolution of 640x480 pixels.

The scenario shows three persons (see Figure3 and 4) sitting at a table being 2.5m from the camera and reporting and discussing things, while looking to the direction of the camera. Both, the camera and microphone array were placed in the same position in the room at 1.2m from the floor. The noise level of the original recordings was low thanks to the absence of projectors and big computers, which are the main non-human source of noise in meeting recordings, this keeping the signal-to-noise ratio (SNR) of above 35 dB.

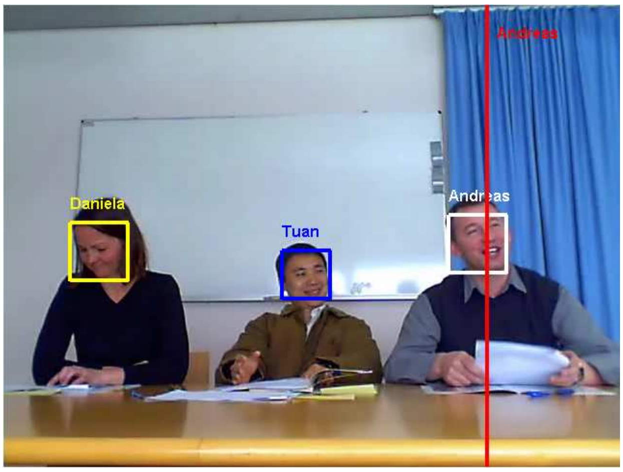

Fig. 3. Visualization of extracted audio (red vertical line) and video information (colored rectangles) at a single timestamp (frame 2175) as suggestion for the user of the annotation tool. Audio and video show the correct result.

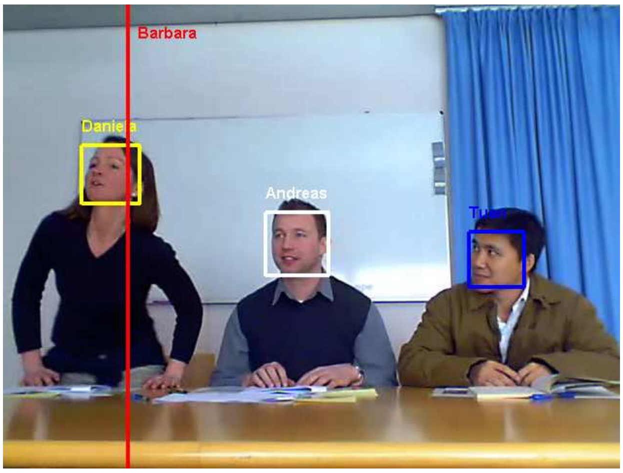

Fig. 4. Visualization of a conflict (frame 4783) between the audio and video models (human identification differs).

Figure 3 and 4 depict the visualization of extracted audio and video features while playing back the video. In the first case the extracted multi-media features correspond, i.e. video information suggest the same person as audio-based speaker identification, and they also show the same position and direction of the speaker (the audio-based vertical line is in the middle of the video-based rectangle). However, the second case shows a conflict: although the audio-based DoA is pointing to a real speaker position, the extracted speaker from audio does not correspond to the identified person derived from video. In this case human interaction is required to resolve the conflict with the help of the semi-automatic annotation tool.

VI. Conclusion

In this paper we propose audio-visual feature extraction modules of a semi-automatic annotation tool and discussed methods for localizing participants and person identity from video data as well as methods to extract voice activity, speaker identity and direction of arrival from audio data. Furthermore we show a method for temporal and spatial mapping between audio and video. At last we discussed the visualization of the extracted features which support efficient video and audio annotation of a huge meeting database.

VII. Acknowledgment

The project results have been developed in the MISTRAL Project, financed by the Austrian Research Promotion Agency (www.ffg.at) within the strategic objective FIT-IT under the project contract number 809264/9338.

References

[1] I. McCowan, . Gatica-Perez, S. Bengio, . Lathoud, . Barnard, and D. Zhang, “Automatic analysis of multimodal group actions in meetings,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[2] A. A. M. party Interaction, “Recognition of attentional cues in meetings,” Information Society, Tech. Rep., 2006.

[3] “Mistral project,” 2005-2006. [Online]. Available: http://www.mistral-project.at

[4] J. R. Smith and B. Lugeon, “A visual annotation tool for multimedia content description,” in *Proc. SPIE Photonics East, Internet Multimedia Management Systems*, 2000.

[5] M. P. I. for Psycholinguistics, “Elan, eudico linguistic annotator,” 2006. [Online]. Available: http://www.mpi.nl/tools/elan.html

[6] M. MRAS, “Microsoft research annotation system,” 2006. [Online]. Available: http://research.microsoft.com/research/pubs/view.aspx?tr˙id=194

[7] Ricoh, “Ricoh movie tool,” 2006. [Online]. Available: http://www.ricoh.co.jp/src/multimedia/MovieTool/

[8] Z. fuer Graphische Datenverarbeitung e. V., “Videto- video description tool,” 2006. [Online]. Available: http://www.zgdv.de/zgdv/zgdv/departments/zr4/Produkte/videto

[9] L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field, “Semi-automatic image annotation,” in *Proc. of Human Computer Interaction*, 2001.

[10] P. T. Brady, “A statistical analysis of on-off patterns in 16 conversations,” *Bell Syst. Tech. J.*, vol. 47(1), pp. 73–91, 1968.

[11] “The webpage of aurora group,” ETSI ES 202 211, 11 2003. [Online]. Available: http://www.etsi.org

[12] D. Tran and D. Sharma, *New Background Speaker Models and Experiments on the ANDOSL Speech Corpus*, Jan. 2004, vol. 3214.

[13] M. Skosan and D. Mashao, “Modified segmental histogram equalization for robust speaker verification,” *Pattern Recognition Letters*, vol. 27, no. 5, pp. 479–486, apr 2006.

[14] M. Képesi, T. V. Pham, G. Kubin, L. Weruaga, A. Juffinger, and M. Grabner, “Noise cancellation frontends for automatic meeting transcription,” in *EURONOISE 2006 Finland*, June 2006.

[15] C. J. van Rijsbergen., *Information Retrieval*, 2nd ed. Butterworths, 1979.

[16] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, vol. I, 2001, pp. 511–518.

[17] S. Avidan, “Ensemble tracking,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 494–501.

[18] N. Oza and S. Russell, “Online bagging and boosting,” in *Proceedings Artificial Intelligence and Statistics*, 2001, pp. 105–112.

[19] G. D. Hager and P. N. Belhumeur, “Efficient region tracking with parametric models of geometry and illumination,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, 1998.

[20] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.