Extracting Activities from Multimodal Observation

Oliver Brdiczka, Jérôme Maisonnasse, Patrick Reignier, James L. Crowley

INRIA Rhône-Alpes, Montbonnot, France
{brdiczka, maisonnasse, reignier, crowley}@inrialpes.fr

This paper addresses the extraction of small group configurations and activities in an intelligent meeting environment. The proposed approach takes a continuous stream of observations coming from different sensors in the environment as input. The goal is to separate distinct distributions of these observations corresponding to distinct group configurations and activities. In this paper, we explore an unsupervised method based on the calculation of the Jeffrey divergence between histograms over observations. The obtained distinct distributions of observations can be interpreted as distinct segments of group configuration and activity. To evaluate this approach, we recorded a seminar and a cocktail party meeting. The observations of the seminar were generated by a speech activity detector, while the observations of the cocktail party meeting were generated by both the speech activity detector and a visual tracking system. We measured the correspondence between detected segments and labelled group configurations and activities. The obtained results are promising, in particular as our method is completely unsupervised.

Introduction

The focus of this work is analyzing human (inter)action in intelligent meeting environments. In these environments, users are collaborating in order to achieve a common goal. Several individuals can form one group working on the same task, or they can split into subgroups doing independent tasks in parallel. The dynamics of group configuration and activity need to be tracked in order to supply reactions or interactions at the most appropriate moment. Changes in group configuration need to be detected to identify main actors, while changes in activity within a group need to be detected to identify activities.

This paper proposes an unsupervised method for extracting small group meeting configurations and activities from a stream of multimodal observations. The method detects changes in small group configuration and activity based on measuring the Jeffrey divergence between adjacent histograms. These histograms are calculated for a window slid from the beginning to the end of a meeting recording and contain the frequency of observations coming from multi-sensory input. The peaks of the Jeffrey divergence curve between these histograms are used to segment distinct distributions of multimodal observations and to find the best model of observation distributions for the given meeting. The method has been tested on observations coming from a speech activity detector as well as a visual tracking system. The evaluation has been done with recordings of a seminar and a cocktail party meeting.

Previous and Related Work

Many approaches for the recognition of human activities in meetings have been proposed in recent years. Most work use supervised learning methods [2], [4], [5], [8], [9]. Some projects focus on supplying appropriate services to the user [8], while others focus on the correct classification of meeting activities [4] or individual availability [5]. Less work has been conducted on unsupervised learning of meeting activities [10]. To our knowledge, little work has been done on the analysis of changing small group configuration *and* activity. In [2] a real-time detector for changing small group configurations has been proposed. This detector is based on speech activity detection and either trained with recorded meetings or defined by hand based on conversational hypotheses. In [2], we showed that different meeting activities, and especially different group configurations, have particular distributions of speech activity. Detecting group configuration or activity (as in [2], [4], [5]) requires, however, a predefined set of activities or group configurations. New activities or group configurations with a different number of individuals cannot be detected and distinguished with these approaches. The approach proposed in this paper focuses on an unsupervised method segmenting small group meetings into consecutive group configurations and activities. These configurations and activities are distinguished by their distributions, but not labelled or compared. The method can thus be seen as a first step within a classification process identifying (unseen) group configurations and activities in meetings.

Approach

We present a novel approach based on the calculation of the Jeffrey divergence between histograms of observations. These observations are a discretization of events coming from multi-sensory input. The observations are generated with a constant sampling rate depending on the sampling rates of the sensors.

Observation Distributions

In [2], we stated that the distribution of the different speech activity observations is discriminating for group configurations in small group meetings. We assume further that in small group meetings distinct group configurations and activities have distinct distributions of multimodal observations. The objective of our approach is hence to separate these distinct distributions, in order to identify distinct small meeting configurations and activities.

As our observations are discrete and unordered (e.g. a 1-dimensional discrete code) and we do not want to admit any a priori distribution, we use histograms to represent observation distributions. A histogram is calculated for an observation window (i.e. the observations between two distinct time points in the meeting recording) and contains the frequency of each observation code within this window.

To separate different observation distributions, we calculate the Jeffrey divergence between the histograms of two adjacent observation windows. The Jeffrey divergence

[6] is a numerically stable and symmetric form of the Kullback-Leibler divergence between histograms. We slide two adjacent observation windows from the beginning to the end of the recorded meetings, while constantly calculating the Jeffrey divergence between these windows. The result is a divergence curve of adjacent histograms (Figure 1).

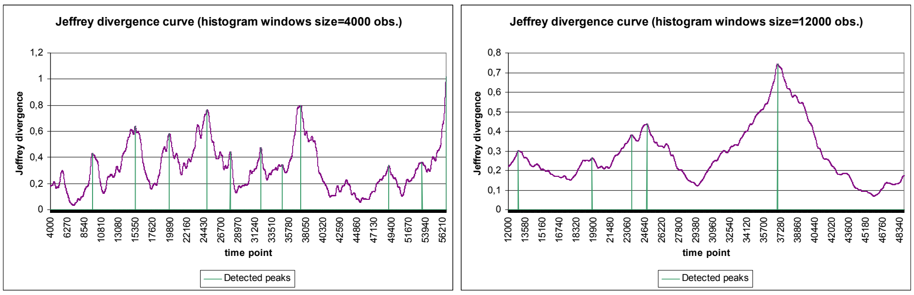


Figure 1. Jeffrey divergence between histograms of sliding adjacent windows of 4000, and 12000 observations (64sec and 3min 12sec)

The peaks of the curves indicate high divergence values, i.e. a big difference between the adjacent histograms at that time point. The size of the adjacent windows determines the exactitude of the divergence measurement. The larger the window size, the less peaks has the curve. However, peaks of larger window sizes are less precise than those of smaller window sizes. Thus we parse the meeting recordings with different window sizes (e.g. in recording of the seminar: window sizes of between 4000 and 16000 observations, which corresponds to a duration between 64sec and 4min 16sec for each window). The peaks of the Jeffrey divergence curve can then be used to detect changes in the observation distribution of the meeting recording.

Peak Detection

To detect the peaks of the Jeffrey divergence curve, we use successive robust mean estimation. Robust mean estimation has been used in [7] to locate the center position of a dominant face in skin color filtered images. Mean and standard deviation are calculated repeatedly in order to isolate a dominant peak. To detect all peaks of the Jeffrey divergence curve, we apply the robust mean estimation process successively to the Jeffrey divergence values.

Merging and Filtering Peaks from different Window Sizes

Peak detection using successive robust mean estimation is conducted for Jeffrey curves with different histogram window sizes. A global peak list is maintained containing the peaks of different window sizes. Peaks in this list are merged and filtered with respect to their window size and peak height.

To merge peaks of Jeffrey curves with different histogram window sizes, we calculate the distance between these peaks normalized by the minimum of the histogram window sizes. The distance is hence a fraction of the minimum window size measuring the degree of overlap of the histogram windows. To merge two peaks, the histogram windows on both sides of the peaks need to overlap, i.e. the normalized distance needs to be less than 1.0.

We filter the resulting peaks by measuring peak quality. We introduce the relative peak height and the number of votes as quality measures. The relative peak height is the Jeffrey curve value of the peak point normalized by the maximum value of the Jeffrey curve (with the same window size). A peak needs to have a relative peak height between 0.5 and 0.6 to be retained. The number of votes of a peak is the number of peaks that have been merged to form this peak. A number of 2 votes are necessary for a peak to be retained.

The small number of peaks resulting from merging and filtering is used to search for the best allocation of observation distributions, i.e. to search for the best model for a given meeting.

Model Selection

To search for the best model for a given meeting recording, we examine all possible peak combinations, i.e. each peak of the final peak list is both included and excluded to the (final) model. For each such peak combination, we calculate the average Jeffrey divergence of the histograms between the peaks. As we want to separate best the distinct observation distributions of a meeting, we accept the peak combination that maximizes the average divergence between the peak histograms as the best model for the given meeting recording.

Evaluation and Results

The result of our approach is the peak combination separating best the activity distributions of a given meeting recording. We interpret the intervals between the peaks as segments of distinct group configuration and activity. To evaluate our approach, we recorded a seminar and a cocktail party meeting. The group configurations and activities of these meetings have been labeled. For the evaluation of the detected segments, we use the *asp*, *aap* and *Q* measure.

Evaluation measures

For the evaluation, we dispose of the timestamps and durations of the (correct) group configurations and activities. However, classical evaluation measures like confusion matrices can not be used here because the unsupervised segmentation process does not assign any labels to the found segments. Instead, we use three measures proposed in [10] to evaluate the detection results: average segment purity (*asp*), average activity purity (*aap*) and the overall criterion *Q* (Figure 2). The *asp* is a measure of how well a segment is limited to only one activity, while the *aap* is a measure of how well one activity is limited to only one segment. In the ideal case (one segment for each activity), *asp* = *aap* = 1. The *Q* criterion is an overall evaluation criterion combining *asp* and *aap*, where larger *Q* indicates better overall performance.

$$asp = \frac{1}{N}\sum_{i=1}^{N_s} p_{i\bullet} \times n_{i\bullet} \quad , \qquad aap = \frac{1}{N}\sum_{j=1}^{N_a} p_{\bullet j} \times n_{\bullet j} \quad ,$$

$$Q = \sqrt{asp \times aap} \quad .$$

with

n_{ij} = *total number of observations in segment i by activity j*

$n_{i\bullet}$ = *total number of observations in segment i*

$n_{\bullet j}$ = *total number of observations of activity j*

N_a = *total number of activities*

N_s = *total number of segments*

N = *total number of observations*

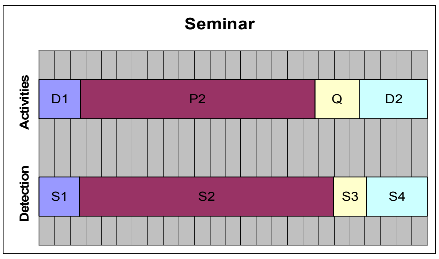
$$p_{\bullet j} = \sum_{i=1}^{N_s} \frac{n_{ij}^2}{n_{\bullet j}^2}$$

$$p_{i\bullet} = \sum_{j=1}^{N_a} \frac{n_{ij}^2}{n_{i\bullet}^2}$$

Figure 2. Average segment purity (*asp*), average activity purity (*aap*) and the overall criterion *Q*

Seminar

We recorded a seminar with 5 participants. The speech of the participants was recorded using lapel microphones. A speech activity detector was executed on the audio channels of the different lapel microphones. One observation was a vector containing a binary value (speaking, not speaking) for each individual that is recorded. This vector was transformed to a 1-dimensional discrete code used as input. Our automatic speech detector has a sampling rate of 62.5 Hz, which corresponds to the generation of one observation every 16 milliseconds.



	asp	aap	Q
Seminar	0.88	0.91	0.90

Figure 3. Activities and their detection for the seminar (meeting duration = 25min 2sec).

The activities during the seminar were discussion in small groups (D1), presentation (P), questions (Q) and discussion in small groups (D2). Figure 3 shows the labeled activities for the seminar and the segments detected by our approach as well as the corresponding *asp*, *aap* and *Q* values. The results of the automatic segmentation are very good; we obtain a *Q* value of 0.90.

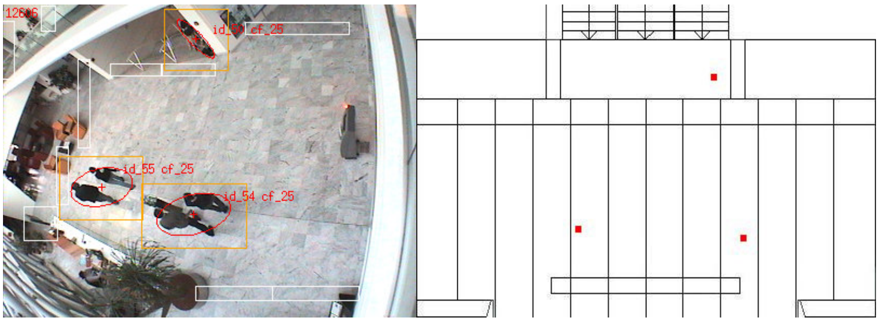


Figure 4. Wide-angle camera image of INRIA Rhône-Alpes entrance hall with one individual and two small groups being tracked (left) and corresponding positions on the hall map after applying a homography (right).

Cocktail Party Meeting

We recorded a cocktail party meeting with 5 participants in the entrance hall of INRIA Rhône-Alpes. The speech of the participants was recorded using headset microphones. As for the seminar, a speech activity detector provided the speech activity observations for each individual. A wide-angle camera filmed the scene and a visual tracking system [3] based on background subtraction provided targets corresponding to individuals or small groups (Figure 4 left). We used a homography to calculate the positions of these targets on the hall map (Figure 4 right). The split and merge of the targets made it difficult to track small interaction groups directly, in particular when interaction groups are near to each other.

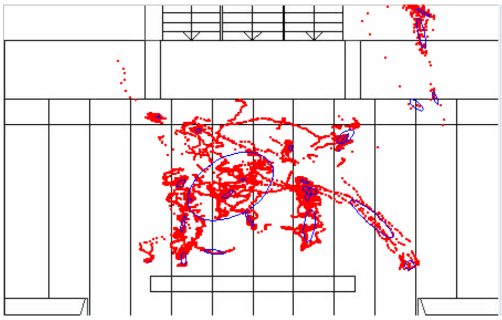
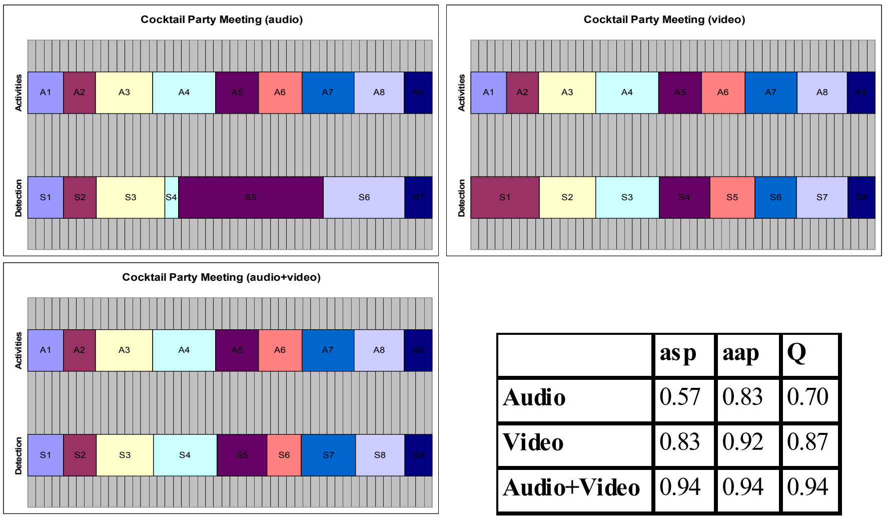


Figure 5. Detected positions of the small groups for the cocktail party recording and the clusters learned by EM algorithm.

To build up a visual model for the changing interaction groups in the scene, we applied a multidimensional EM clustering algorithm [1] to the positions on the hall map as well as the angle and the ratio of first and second moment of the bounding ellipses of all targets. The EM algorithm identified 27 clusters for the cocktail party recording. Figure 5 indicates the positions of all targets as well as the clusters learned by EM on the hall map.

The observations are provided by the automatic speech detector and by the visual model built up by EM. The observations provided by the visual model are the dominant clusters given the targets in the current video frame, i.e. the clusters of the model with the highest probability of having generated the targets. The tracking system has a frame rate of 16 frames per second, which corresponds to the generation of an observation every 62.5 ms. The histograms of our approach are calculated for the observations coming from the speech activity detector as well as from the visual model. The fusion is done by simply summing the Jeffrey divergence values of the speech detector and visual model histograms.


Cocktail Party Meeting (audio)

Cocktail Party Meeting (video)

Cocktail Party Meeting (audio+video)

	asp	aap	Q
Audio	0.57	0.83	0.70
Video	0.83	0.92	0.87
Audio+Video	0.94	0.94	0.94

Figure 6. Group configurations and their detection for the cocktail party (meeting duration=30min 26sec).

The participants formed different interaction groups during the cocktail party meeting. The interaction group configurations were labeled. Our approach has been applied to the speech detector observations, the visual model observations, and both the speech detector and the visual model observations. Figure 6 shows the labeled group configurations and the detected segments as well as the corresponding *asp*, *aap* and *Q* values. The results of the segmentation of both audio and video are very good, outperforming the separate segmentations. The *Q* value of the video and audio segmentation is 0.94.

Conclusion

We proposed an approach for extracting small group configurations and activities from multimodal observations. The approach is based on an unsupervised method for segmenting meeting observations coming from multiple sensors. We calculate the Jeffrey divergence between histograms of meeting activity observations. The peaks of the Jeffrey divergence curve are used to separate distinct distributions of meeting activity observations. These distinct distributions can be interpreted as distinct segments of group configuration and activity. We measured the correspondence between the detected segments and labeled group configurations and activities for a seminar and a cocktail party recording. The obtained results are promising, in particular as our method is completely unsupervised.

The fact that our method is unsupervised is especially advantageous when analyzing meetings with an increasing number of participants (and thus possible group configurations) and a priori unknown activities. Our method then provides a first segmentation of a meeting, separating distinct group configurations and activities. These detected segments can be used as input for further classification tasks like meeting comparison or meeting activity recognition.

References

1. Bilmes, J. A., *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, Technical Report, University of Berkeley, 1998.
2. Brdiczka, O., Maisonnasse, J., and Reignier, P., *Automatic Detection of Interaction Groups*, Proc. Int'l Conf. Multimodal Interfaces, 2005.
3. Caporossi, A., Hall, D., Reignier, P., Crowley, J.L., *Robust visual tracking from dynamic control of processing*, Proc. Int'l PETS Workshop, 2004.
4. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D., *Automatic Analysis of Multimodal Group Actions in Meetings*, IEEE Trans. on Pattern Analysis and Machine Intelligence, March 2005.
5. Muehlenbrock, M., Brdiczka, O., Snowdon, D., and Meunier, J.-L., *Learning to Detect User Activity and Availability from a Variety of Sensor Data*, Proc. IEEE Int'l Conference on Pervasive Computing and Communications, March 2004.
6. Puzicha, J., Hofmann, Th., and Buhmann, J., *Non-parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval*. Proc. Int'l Conf. Computer Vision and Pattern Recognition, 1997.
7. Qian, R. J., Sezan, M. I., and Mathews, K. E., *Face Tracking Using Robust Statistical Estimation*, Proc. Workshop on Perceptual User Interfaces, San Francisco, 1998.
8. Stiefelhagen, R., Steusloff, H., and Waibel, A., *CHIL - Computers in the Human Interaction Loop*, Proc. Int'l Workshop on Image Analysis for Multimedia Interactive Services, 2004.
9. Zaidenberg, S., Brdiczka, O., Reignier, P., Crowley, J.L., *Learning context models for the recognition of scenarios*, Proc. IFIP Conf. on Artificial Intelligence Applications and Innovations, 2006 (to appear).
10. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., and Lathoud, G., *Multimodal Group Action Clustering in Meetings*, Proc. Int'l Workshop on Video Surveillance & Sensor Networks, 2004.