Appl Intell
DOI 10.1007/s10489-007-0074-y

Detecting small group activities from multimodal observations

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

Abstract This article addresses the problem of detecting configurations and activities of small groups of people in an augmented 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. This article describes an unsupervised method based on the calculation of the Jeffrey divergence between histograms over observations. These histograms are generated from adjacent windows of variable size slid from the beginning to the end of a meeting recording. The peaks of the resulting Jeffrey divergence curves are detected using successive robust mean estimation. After a merging and filtering process, the retained peaks are used to select the best model, i.e. the best allocation of observation distributions for a meeting recording. These distinct distributions can be interpreted as distinct segments of group configuration and activity. To evaluate this approach, 5 small group meetings, one seminar and one cocktail party meeting have been recorded. The observations of the small groups meetings and 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. The authors measured the correspondence between detected segments and labeled group configurations and activities. The obtained results are promising, in particular as the method is completely unsupervised.

A short version of this article [6] obtained the Best Paper Award of the 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2006.

O. Brdiczka (✉) · J. Maisonnasse · P. Reignier · J.L. Crowley
INRIA Rhône-Alpes, 655 avenue de l'Europe, 38334 Saint Ismier Cedex, France
e-mail: brdiczka@inrialpes.fr

J. Maisonnasse
e-mail: maisonnasse@inrialpes.fr

P. Reignier
e-mail: reignier@inrialpes.fr

J.L. Crowley
e-mail: crowley@inrialpes.fr

1 Introduction

Automatic analysis of small group meetings is an emerging field of research for speech, video and multimodal technologies. In the last decade, many computerized spaces equipped with multiple sensor arrays for sensing human behavior have been demonstrated and used for experiments with new forms of man-machine interaction and computer mediated communication [4, 7, 11]. Such environments enable computer observation of human (inter)action. The analysis of (inter)actions of two and more individuals is of particular interest as it provides information about social context and relations and it further enables computer systems to follow and anticipate human (inter)action. The latter is a difficult task given the fact that human activity is situation dependent [17] and does not necessarily follow preestablished plans. A correct separation and recognition of human activities as well as a correct identification of human actors is thus a first step towards analyzing small group meetings.

The detection of small group configuration of the users is a useful first step for describing activity in meetings. In a physical environment, 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, i.e. the split and merge of small interaction groups, allows us to detect the appearance of new activities. It is assumed that a change in group configuration is strongly linked to a change in activity, at least to an interruption of the current activity. The fusion of several independent small groups provides important information for detecting a change of the current activity, on a local or global level. For example, people attending a seminar tend to form small groups discussing different topics before the seminar starts. When the lecturer arrives, these small groups merge and form a big group who listen to the lecture. In this example, the fusion of several small groups to one big group can be used to detect the beginning of a seminar. In the same manner, the split of the big group into several small groups can indicate a pause or the end of the lecture. The change in group configuration is thus a strong indicator of new activities as well as of activities that are linked to a particular group configuration (for example a lecture).

Many approaches for recognizing human (inter)actions from sensor data have been proposed in recent years [4, 5, 12–14, 20], with particular attention to applications in video surveillance [14, 20], workplace tools [12, 18] and group entertainment [4]. Most of the reported work has been based on visual information [14, 20] or audio information [5] using statistical models for learning and recognition (in particular Hidden Markov Models [5, 12, 14, 20]). Some projects have focused on supplying appropriate system services to the users [4, 18], while others focus on the correct classification of activities [13]. Most of the reported work has been concerned with the recognition of the activities of individuals who have been identified a priori. Very little work has been done on the analysis of group formation [1, 5].

The recognition of human activity based on speech events is often used in the context of group analysis. In general, the group and its members are defined in advance. The objective is then to use frequency and duration of speech contributions to recognize particular key actions executed by group members [12] or to analyse the type of meeting in a global manner [8]. However, the detection of dependencies between individuals and their membership in one or several groups is not considered. Analysing large amounts of data from recordings of interactions enables the reconstruction of social networks for a number of individuals [10]. The detection and analysis of conversations is then necessary. The automatic detection of conversations using mutual information [2], in order to determine who speaks and when, requires a significantly long duration for each conversation. Little work has been done on the analysis of changing small group configuration and activity. In [5], the authors have presented a user study of a system determining and supporting interaction groups in an audiospace. The system uses a naive Bayesian classifier to determine the interaction group configuration. However, the focus of the article is laid on the user study, no detection results are presented. In [5], a real-time detector for small group configurations of 4 participants has been proposed. This detector is based on a predefined Hidden Markov Model constructed upon conversational hypotheses (Fig. 1). The HMM takes speech activity events of meeting participants as input.

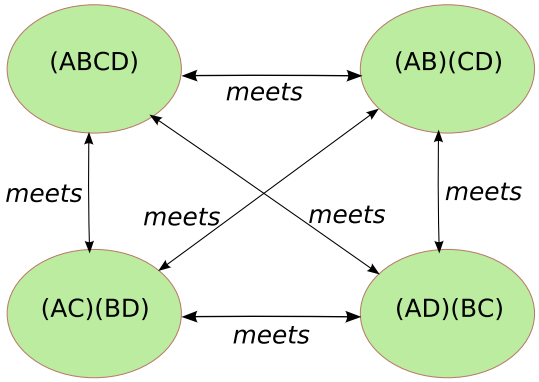


Fig. 1 Situation model describing possible group configurations for a meeting of 4 individuals A, B, C, D

For real applications, it is rarely practical to exactly specify all possible configurations of sub-groups (as in [5, 12]), especially when dealing with a variable or changing numbers of participants. In this article, an unsupervised method is proposed detecting small group meeting configurations and activities from a stream of multimodal observations. The obtained segmentation can be used as input for classification and detection of activities. The proposed method detects changes in small group configuration and activity based on measuring the Jeffrey divergence between adjacent histograms of observations. In [5], the authors showed that different meeting activities, and especially different group configurations, have particular distributions of speech activity. This can be extended to distributions of multimodal observations coming from multi-sensory input. These distributions are represented by histograms containing the frequency of these observations. To separate distinct distributions of observations, two adjacent windows are slid from the beginning to the end of the meeting recording, while constantly calculating the Jeffrey divergence between the histograms generated from the observations within these windows. The size of the sliding adjacent windows is varied generating several Jeffrey divergence curves. The peaks of the resulting curves are detected using successive robust mean estimation. The detected peaks are merged and filtered with respect to their height and window size. The retained peaks are finally used to select the best model, i.e. the best allocation of observation distributions for the given meeting recording.

The method has been tested on observation recordings of 7 meetings. Five speech activity recordings of short small group meetings with 4 participants, one speech activity

recording of a seminar with 5 participants and a audiovisual observation recording of a cocktail party meeting with 5 participants. The approach showed promising results for all meeting recordings.

2 Method

A novel method based on the calculation of the Jeffrey divergence between histograms of observations is presented. 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.

2.1 Observation distributions

In [5], the authors stated that the distribution of the different speech activity observations is discriminating for group configurations in small group meetings. It is further assumed that in small group meetings distinct group configurations and activities have distinct distributions of multimodal observations. The objective of the proposed approach is hence to separate these distinct distributions, in order to identify distinct small meeting configurations and activities. Because the observations are discrete and unordered (e.g. a 1-dimensional discrete code) and there is no a priori observation distribution, histograms are used 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.

$$J_{p,q} = \sum_{x \in X} p(x) \cdot \log \frac{p(x)}{\frac{p(x)+q(x)}{2}} + q(x) \cdot \log \frac{q(x)}{\frac{p(x)+q(x)}{2}} \quad (1)$$

The Jeffrey divergence [15] is a numerically stable and symmetric form of the Kullback–Leibler divergence between histograms. Equation (1) indicates the formula to calculate the Jeffrey divergence between two histograms p and q. The set X contains the bins of the histograms. The value $p(x)$ refers to the empirical probability of the observation associated to bin x.

The Jeffrey divergence may be used to separate different observation distributions by calculating the divergence between the histograms of two adjacent observation windows. With this approach, two adjacent observation windows are slid from the beginning to the end of the recorded meetings, and the Jeffrey divergence is computed for each position. The result is a divergence curve of adjacent histograms (Fig. 2).

The peaks of the Jeffrey divergence curve can be used to detect changes in the observation distribution of a meeting recording. 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.

As the observations are generated with a fixed sampling rate, an observation window size used for the calculation of a histogram corresponds to a temporal interval. Different window sizes cover thus the detection of activities with different durations. As there should not be a strong a priori concerning the duration of activities and group configurations, the method is applied to several different window sizes. The choice of these window sizes is fixed by the minimal duration of the activities that are expected. A minimal duration between 64 s and 4 min 16 s has been fixed for small group meetings, which corresponds to a window size of between 4000 and 16 000 audio observations.

2.2 Peak detection

To detect the peaks of the Jeffrey divergence curve, successive robust mean estimation is used. Robust mean estimation detects the dominant peak of the Jeffrey divergence curve. Successive robust mean estimation applies the robust mean estimation process several times to the curve in order to isolate all peaks. In the following, the robust mean estimation process and the associated equations will be detailed. Then, successive robust mean estimation will be described.

Robust mean estimation has first been used by Qian et al. [16] to locate the center position of a dominant face in skin color filtered images. The idea is to calculate iteratively a trimmed mean for the filtered pixels of the image. The trimmed mean converges towards the dominant skin color blob in the image.

Figure 3 describes the robust mean estimation process to detect the dominant peak of the Jeffrey divergence curve. The first step of robust mean estimation is to calculate global mean μ and standard deviation σ for the Jeffrey divergence curve using (2) and (3).

$$\mu = \frac{1}{\hat{J}} \sum_{t=t_{\mathrm{MIN}}}^{t_{\mathrm{MAX}}} t \cdot J_{h[t-\mathrm{size},t],h[t,t+\mathrm{size}]} \quad (2)$$

$$\sigma = \sqrt{\frac{1}{\hat{J}} \sum_{t=t_{\mathrm{MIN}}}^{t_{\mathrm{MAX}}} (t-\mu)^2 \cdot J_{h[t-\mathrm{size},t],h[t,t+\mathrm{size}]}} \quad (3)$$

$$\hat{J} = \sum_{t=t_{\mathrm{MIN}}}^{t_{\mathrm{MAX}}} J_{h[t-\mathrm{size},t],h[t,t+\mathrm{size}]} \quad (4)$$

$J_{h[t-\mathrm{size},t],h[t,t+\mathrm{size}]}$ refers to the Jeffrey divergence between the adjacent histograms of size *size* at time point t.

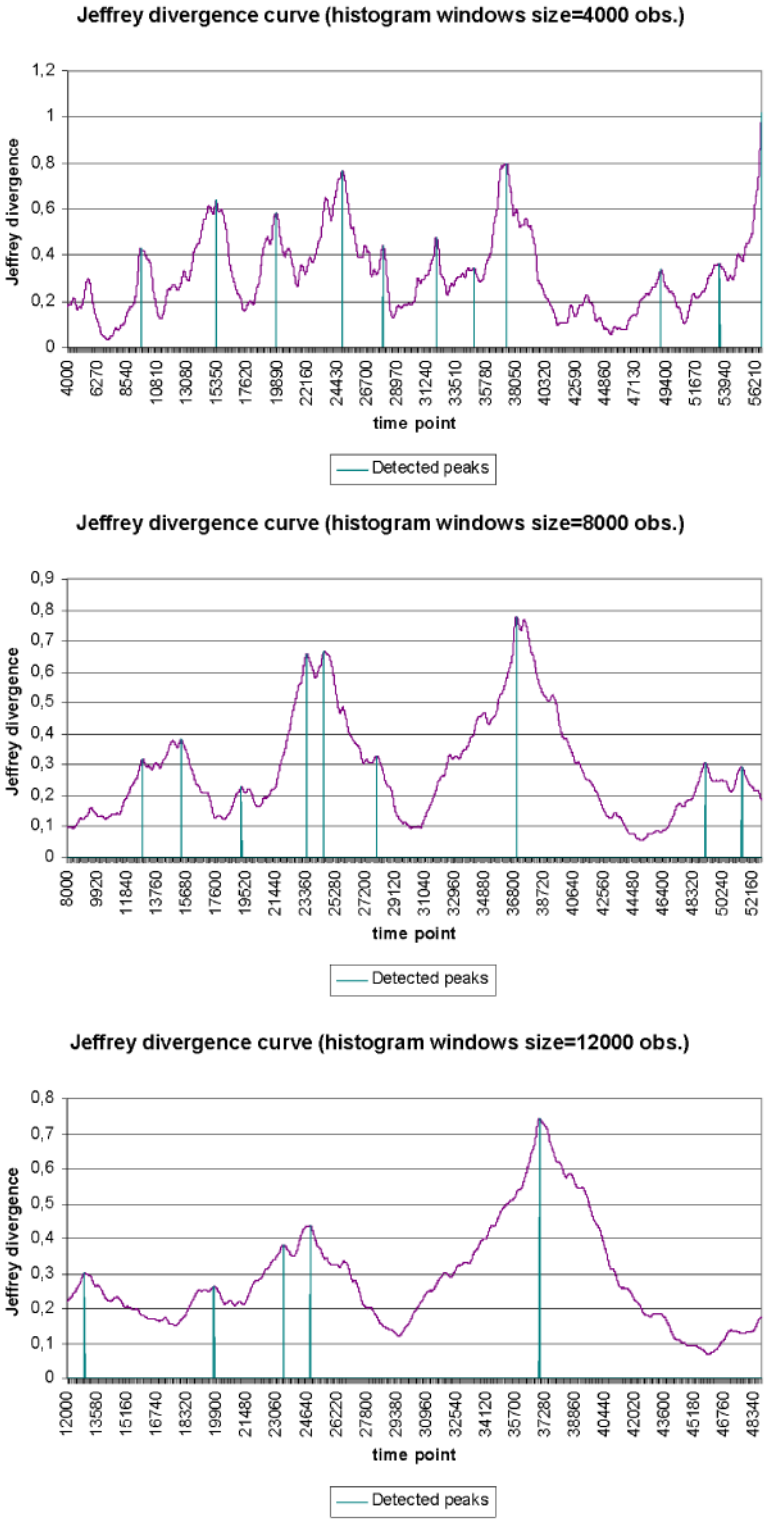


Fig. 2 Small Group Meeting 5: Jeffrey divergence between histograms of sliding adjacent windows of 4000, 8000 and 12 000 speech activity observations (64 s, 2 min 8 s and 3 min 12 s)

Both equations are normalized by the sum of all Jeffrey divergence values (see (4)). In the second and third step, a new trimmed mean $\mu(k+1)$ and deviation $\delta(k+1)$ are calculated based on the Jeffrey curve values within the (standard) deviation around the previous (global) mean. This process is repeated until the trimmed mean converges (Step 4). The maximum within the last interval is set to be the dominant peak of the Jeffrey divergence curve.

To detect all peaks of the Jeffrey divergence curve, the robust mean estimation process is successively applied

Step 1. Compute mean μ and standard deviation σ based on all the points of the Jeffrey curve.

Step 2. Let $\mu(0)=\mu$ and $\delta(0)=max(\sigma, mindev)$.

Step 3. Compute trimmed mean $\mu(k+1)$ and deviation $\delta(k+1)$ based on points within the interval $[\mu(k)-\delta(k), \mu(k)+\delta(k)]$.

Step 4. Repeat Step 3 until $|\mu(k+1)-\mu(k)| < \varepsilon$. Denote the converged mean as μ^* and the converged deviation δ^*.

Step 5. Set the dominant peak position p^* to the position of the maximum within the interval $[\mu^*-\delta^*, \mu^*+\delta^*]$.

Fig. 3 Robust mean estimation process detecting a dominant peak of the Jeffrey divergence curve

Step 1. Detect dominant peak p^* using robust mean estimation.

Step 2. Erase points within peak window $[\mu^*-\delta^*, \mu^*+\delta^*]$ from Jeffrey divergence curve.

Step 3. Repeat Steps 1 and 2 until $J_{h[p^*-size,\,p^*],\,h[p^*,\,p^*+size]} \leq \bar{J}$ with

$$\bar{J} = \frac{1}{t_{MAX}-t_{MIN}+1} \sum_{t=t_{MIN}}^{t_{MAX}} J_{h[t-size,t],\,h[t,t+size]} \,.$$

Fig. 4 Successive robust mean estimation process detecting the peaks of the Jeffrey divergence curve

(Fig. 4). After each robust mean estimation, the found dominant peak is erased (Step 2). This process is repeated while the heights of isolated peaks are above the average height of the curve (Step 3).

2.3 Merging and filtering peaks from different window sizes

Peak detection is conducted for a fixed histogram window size, i.e. the size of the adjacent observation windows used for calculating the histograms needs to be specified for the successive robust mean estimation process (Sect. 2.2).

Peak detection using successive robust mean estimation (Sect. 2.2) is conducted for Jeffrey curves with different histogram window sizes. The window size refers to the observation window used for calculating the histograms. Figure 2 shows example Jeffrey curves for three different observation window sizes. Some peaks are detected for several curves, while others are specific for one particular window size. In order to determine which peaks to choose for segmenting the multimodal observation recording, peaks appearing at several window sizes are first merged and then filtered locally with respect to their window size and peak height. Local filtering refers to a selection of peaks based on their properties.

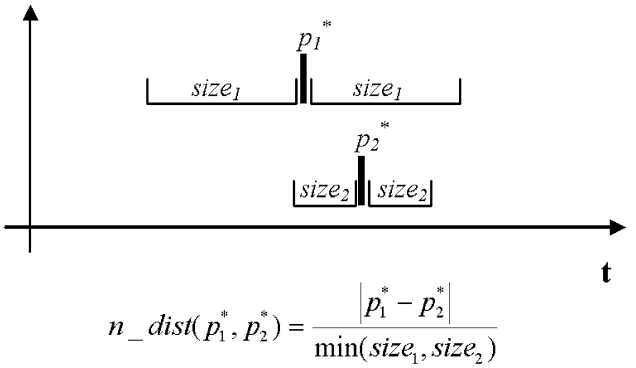


Fig. 5 Normalized distance *n_dist* between two peaks p_1^*, p_2^* of Jeffrey curves with different window sizes $size_1$, $size_2$

2.3.1 Merging peaks

To merge peaks appearing at several histogram window sizes, the distance between these peaks needs to be calculated. Figure 5 proposes a normalized distance measure between peaks of different window sizes. The (temporal) distance between two peaks is normalized by the minimum of the involved histogram window sizes. The resulting normalized distance measures the degree of overlap between the histogram windows.

To merge two peaks, the histogram windows on both sides of the peaks must overlap, i.e. the normalized distance must be less than 1.0. The position of the resulting merged peak is determined by the position of the highest peak that has been merged.

2.3.2 Filtering peaks

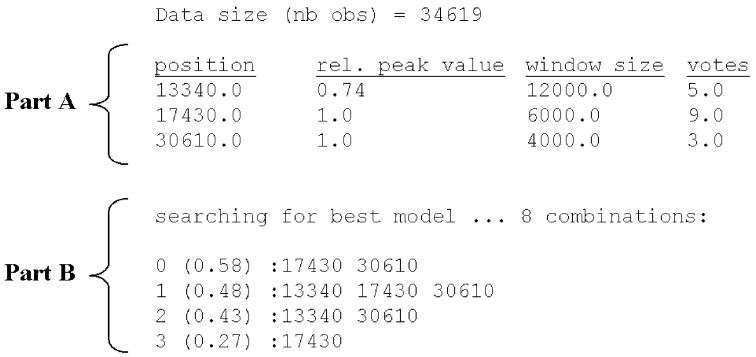
The resulting peaks are filtered by measuring peak quality. Relative peak height and number of votes are introduced as quality measures. The relative peak height is the Jeffrey curve value of the peak point normalized by the maximum value of the Jeffrey curve (with the same window size). A peak needs 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.

Merging and filtering operate on the positions and features of the detected peaks, i.e. in a local context. In order to determine the best allocation of observation distributions for a given recording, the best combination of the peaks retained by the merging and filtering process must be searched. This global search process is called *model selection*.

2.4 Model selection

Model selection is a global search process that aims at determining the best allocation of observation distributions for

Fig. 6 Small Group Meeting 1: Output of the algorithm



a given recording. The input is the list of peaks retained by the merging and filtering process. The output is the combination that maximizes the divergence between the distinct observation distributions of the recording. It is assumed that the best allocation of observation distributions corresponds to maximizing the average divergence between the observation distributions.

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

Figure 6 shows an example output of the model selection algorithm. Part A of the figure indicates the resulting peaks of the merging and filtering process. Four peaks have been retained, which means that $4 * 4 = 16$ possible peak combinations must be examined. Part B lists the eight best peak combinations (sorted by descending average Jeffrey divergence) that have been found by the model selection process. Model 0 would have been selected, corresponding to a segmentation of the recording at positions 17430, 30610 and an average Jeffrey divergence between the three segments of 0.58.

3 Evaluation and results

To evaluate the approach, 5 short small group meetings (Sect. 3.2), one seminar (Sect. 3.3) and a cocktail party meeting (Sect. 3.4) have been recorded. The group configurations and activities of these meetings have been hand labeled. The result of the proposed approach is the peak combination separating best the activity distributions for each meeting recording. The intervals between the peaks are interpreted as segments of distinct group configuration and activity. The *asp*, *aap* and Q measures (described in Sect. 3.1) are used for the evaluation of these segments with regard to the labeled group configurations and activities.

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

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

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*

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

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

Fig. 7 Average segment purity (*asp*), average activity purity (*aap*) and the overall criterion Q

3.1 Segmentation quality measure

This subsection defines the measures that are used to determine the quality of the automatic segmentations obtained for the evaluation data. The timestamps and durations of the (correct) group configurations and activities have been hand labeled. As the proposed method is unsupervised, the direct correspondence between detected segments and hand labeled activities cannot be measured (e.g. by using confusion matrices) because the unsupervised segmentation process does not assign any labels to the found segments. In order to measure segmentation quality, three measures proposed by Zhang et al. [21] are used: average segment purity (*asp*), average activity purity (*aap*) and the overall criterion Q (Fig. 7).

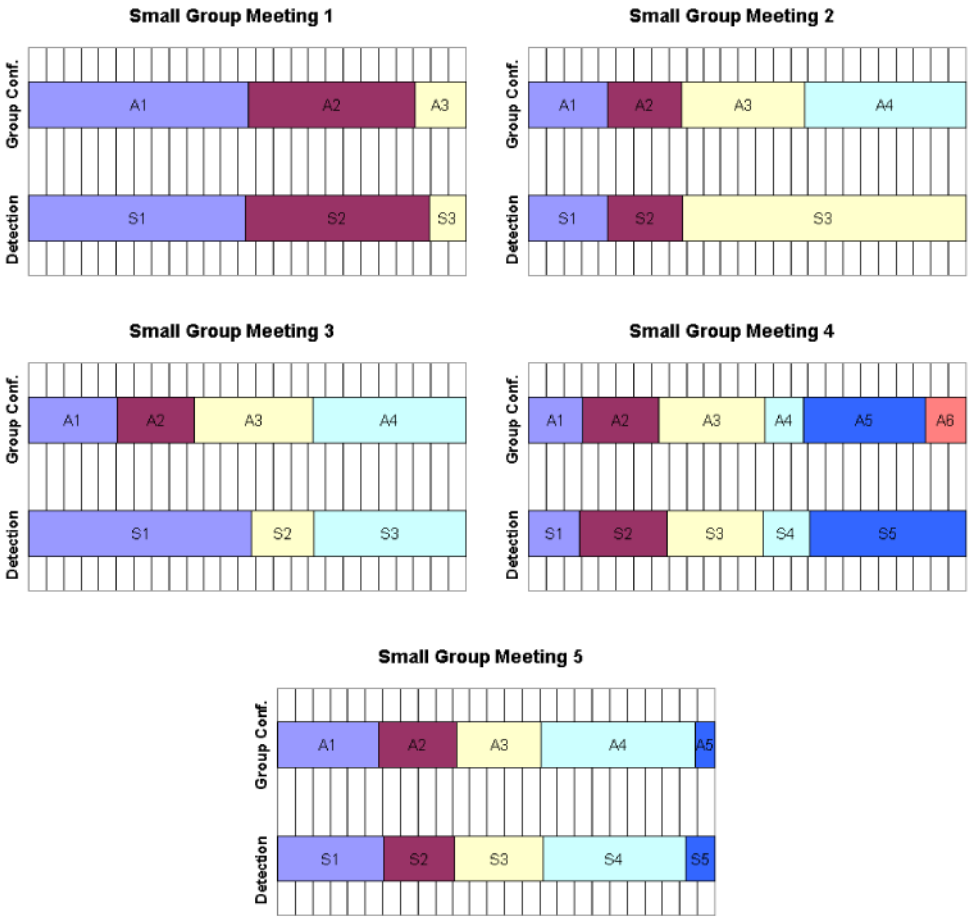


Fig. 8 Meeting 1-5: group configurations and their detection

asp, *aap* and *Q* measure the quality of the segmentation based on purity of the found segments and labeled segments. The *asp* measures the purity of one segment with regard to the labeled activities, i.e. the *asp* indicates how well one segment is limited to only one activity. The *aap* measures the purity of one activity with regard to the detected segments, i.e. the *aap* indicates to which extent one labeled activity corresponds to only one detected segment. The *Q* criterion is an overall evaluation criterion combining *asp* and *aap*.

asp, *aap* and *Q* values are comprised between 0 and 1, where larger values indicate better quality. In the ideal case (one segment for each labeled activity), $asp = aap = 1$ and $Q = 1$.

3.2 Short small group meetings

Five short meetings (duration: between 9 min 14 s and 16 min 12 s) with 4 participants have been recorded. The speech of each individual was recorded using a lapel microphone. The use of lapel microphones has been admitted in order to minimize correlation errors of speech activity of different individuals, i.e. speech of individual *A* is detected as speech of individual *B*. A real-time speech activity detector [6, 19] generated binary observation values (speaking, not speaking) for each individual that is recorded. These binary observations were combined to a 1-dimensional discrete observation code. The generated code comprises 2^n different values, where n is the number of recorded individuals. For the five short meetings with 4 individuals, the resulting observation code has $2^4 = 16$ values comprised between 0 and 15. The automatic speech detector has a sampling rate of 62.5 Hz, which corresponds to the generation of one observation every 16 milliseconds.

The individuals formed different interaction groups during the meetings. The possible configurations of these interaction groups were (*ABCD*), (*AC*)(*BD*), (*AD*)(*BC*) and (*AB*)(*CD*), where *A, B, C, D* refer to the 4 individuals. As verbal interaction concerns at least two individuals, groups containing only one individual are excluded. For the experiments, The number and order of group configurations, i.e. who will speak with whom, was fixed in advance. The timestamps and durations of the group configurations were, however, not predefined and changed spontaneously. The individuals were free to move and to discuss any topic.

Figure 8 shows the labeled group configurations for each small group meeting as well as the segments detected by the proposed approach. Table 1 indicates the *asp*, *aap*

Table 1 *asp*, *aap* and *Q* values for the small group meetings

	Duration	asp	aap	Q
Meeting 1	9 min 14 s	0.94	0.93	0.93
Meeting 2	10 min 14 s	0.68	0.99	0.82
Meeting 3	16 min 11 s	0.66	0.86	0.75
Meeting 4	14 min 47 s	0.78	0.91	0.85
Meeting 5	16 min 12 s	0.93	0.92	0.92
Average		0.80	0.92	0.85

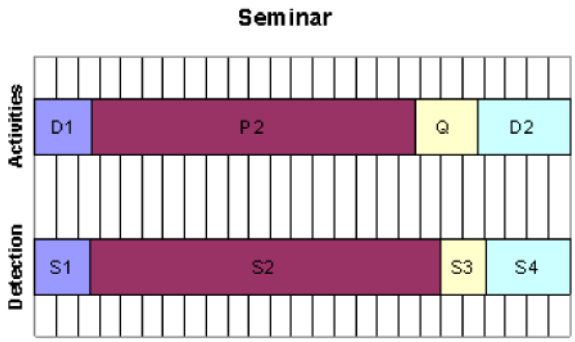


Fig. 9 Seminar: activities and their detection

and *Q* values for each meeting as well as the average of these values for all meetings. Unlike meeting recordings 1, 4 and 5, recordings 2 and 3 contain numerous wrong speech activity detections caused by correlation errors and microphone malfunctions, which explains lower *asp* and *Q* values. However, the overall results of the proposed approach are very good; the average *Q* value is 0.85.

3.3 Seminar

A seminar (duration: 25 min 2 s) with 5 participants has been recorded. As for the small group meetings, the speech of the participants was recorded using lapel microphones. The automatic speech detector provided the speech activity observations (speaking, not speaking) for each individual. These observations were combined to an observation code with 2^5 different values comprised between 0 and 31. The activities during the seminar were "discussion in small groups" (D1), "presentation" (P), "questions" (Q) and "discussion in small groups" (D2). Figure 9 shows the labeled activities for the seminar as well as the segments detected by the proposed approach. Table 2 indicates the *asp*, *aap* and *Q* value. The results of the automatic segmentation are very good; the obtained *Q* value is 0.90.

3.4 Cocktail party meeting

A cocktail party meeting (duration: 30 min 26 s) with 5 participants has been recorded in the entrance hall of INRIA Rhône-Alpes. The recording was multimodal, including audio and video information. The speech of the participants was recorded using headset microphones. A wide-angle camera filmed the scene and a visual tracking system provided targets corresponding to individuals or small groups (Fig. 10 top). The proposed method has been applied to the audio, video and audiovisual information of the recording.

The audio of each individual has been recorded using lapel microphones. As for the small group meetings and the seminar, the audio channels of the different lapel microphones have been analyzed by a speech activity detector providing binary speech activity observations (speaking, not speaking) for each individual. These binary values are combined to an audio observation code ($2^5 = 32$ values between 0 and 31) generated every 16 milliseconds.

Table 2 *asp*, *aap* and *Q* values for the seminar

	Duration	asp	aap	Q
Seminar	25 min 2 s	0.88	0.91	0.90

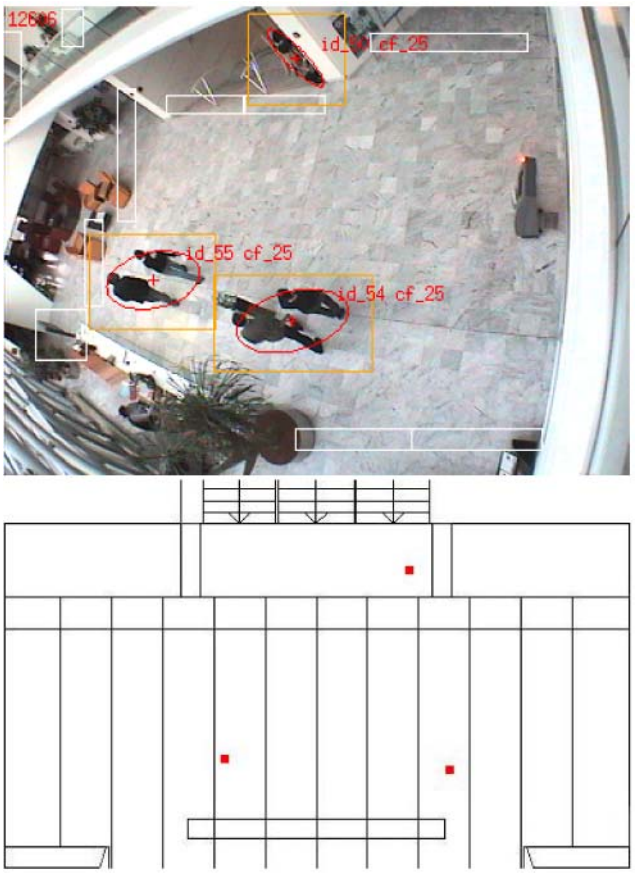


Fig. 10 Wide-angle camera image of INRIA Rhône-Alpes entrance hall with three targets being tracked (*top*) and the corresponding target positions on the hall map after applying a homography (*bottom*). White rectangles in the camera image (*top*) indicate the detection zones used by the visual tracker for creating new targets

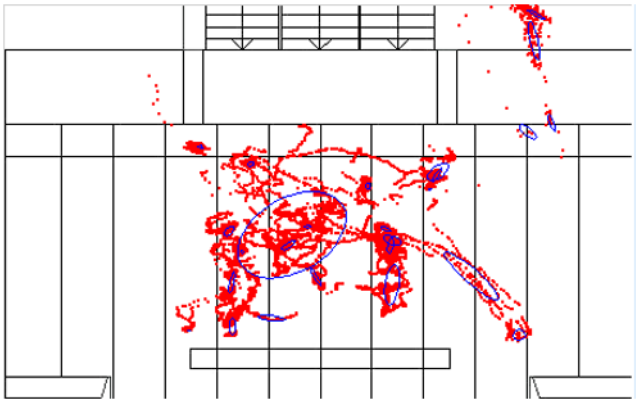


Fig. 11 Cocktail party: positions of all targets on the hall map (*red dots*) and 27 position clusters isolated by EM algorithm (*small blue ellipses*)

The visual tracking system [9] is based on background subtraction and creates and tracks targets based on the video images of the wide-angle camera. The detected targets may correspond to individuals or small groups. The split and merge of these targets made it difficult to track small interaction groups directly, in particular when interaction groups are near to each other. In order to generate visual observation codes, the positions of the targets need to be discretized. First, the targets tracked by the visual tracking system have been mapped on the hall map using a homography. A homography defines a relation between two figures, such that to any point in one figure corresponds one and only one point in the other, and vice versa. Figure 10 bottom shows the three points on the hall map corresponding to the three targets currently tracked by the visual tracking system (Fig. 10 top).

Then, a multidimensional EM clustering algorithm [3] has been applied to all target positions on the hall map as well as the angle and the ratio of first and second axis of the bounding ellipses of all targets. The EM algorithm was initially run with a high number of possible clusters, while constantly eliminating those with too weak contribution to the whole model. 27 clusters were identified for the cocktail party recording. Figure 11 indicates the positions of all targets (red dots on the hall map) as well as the clusters learned by EM (small blue ellipses on the hall map). Finally, the visual observations are generated based on the dominant position clusters in the current video frame. The dominant position clusters are the clusters of the EM model with the highest probability of having generated the targets in the current video frame. The number of visual observations is limited to the number of clusters (here: 27). The appearance of a dominant cluster in a video frame is counted as one observation, thus augmenting the frequency of this cluster in the histograms. The tracking system has a frame rate of 16 frames per second, which corresponds to the generation of visual observation codes every 62.5 milliseconds.

Table 3 asp, aap and Q values for the cocktail party

	Duration	asp	aap	Q
Audio	30 min 26 s	0.57	0.83	0.70
Video	30 min 26 s	0.83	0.92	0.87
Audio+Video	30 min 26 s	0.94	0.94	0.94

The histograms of the proposed approach are calculated for the audio observations coming from the speech activity detector as well as for the visual observations coming from the visual tracker. The fusion is done by simply summing the Jeffrey divergence values of the audio observation histograms and the visual observation histograms. Summing the Jeffrey divergence values of the histograms from different modalities is an easy and efficient way to fuse multimodal information because no data conversions or additional fusion calculations are necessary.

The participants formed different interaction groups during the cocktail party meeting. The interaction group configurations were labeled. Figure 12 shows the labeled group configurations as well as the segments detected by the proposed approach. The approach has been applied to the speech detector observations (Fig. 12 top left), the visual model observations (Fig. 12 top right), and both the speech detector and the visual model observations (Fig. 12 bottom). Table 3 indicates the corresponding asp, aap and Q values. The results of the audio segmentation were very good in the beginning of the cocktail party, but degraded afterwards due to less regulation in speech contributions of the participants and correlation errors of the microphones.

The results of the visual segmentation are very good because of the fact that participants forming an interaction group tend to separate from other interaction groups in the environment. However, distinct interaction groups do not always separate in the environment, which leads to detection errors in the beginning of the meeting. 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.

4 Conclusions

This chapter proposed an approach for detecting small group configurations and activities from multimodal observations. The approach is based on an unsupervised method for segmenting meeting observations coming from multiple sensors. The Jeffrey divergence between histograms of meeting activity observations is calculated. 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. The correspondence between the detected

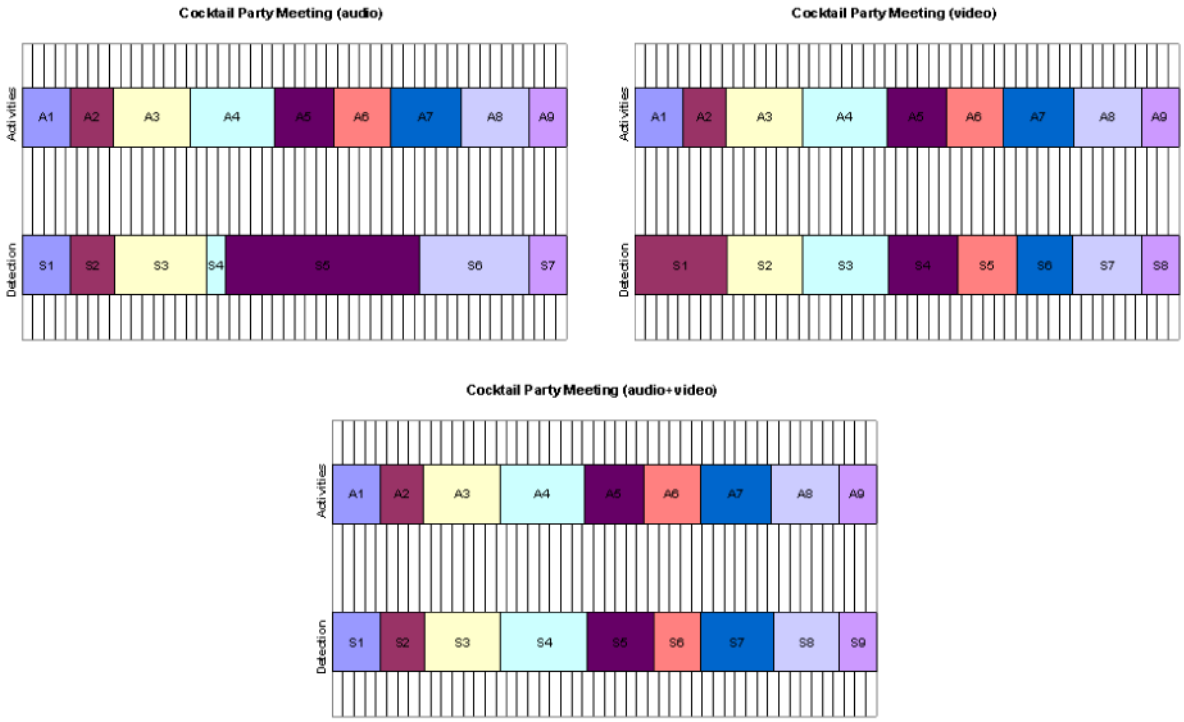


Fig. 12 Cocktail party: group configurations and their detection based on audio, video and audiovisual data

segments and labeled group configurations and activities has been measured for 7 small group recordings. The obtained results are promising, in particular as the method is completely unsupervised.

The fact that the proposed 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. The method then provides a first segmentation of a meeting, separating distinct group configurations and activities. These detected segments can be used as input for learning and recognizing meeting situations and to build up a context model for a meeting. Additional meeting information will then be necessary to disambiguate all situations. Head orientation, pointing gestures or interpersonal distances seem to be good indicators. As described for the cocktail party meeting, the proposed approach can easily be extended to integrate further meeting information coming from different sensors.

Acknowledgements The anonymous reviewers are thanked for making very useful comments that helped improving the quality of this article.

References

1. Aoki PM, Romaine M, Szymanski MH, Thornton JD, Wilson D, Woodruff A (2003) The mad hatter's cocktail party: a social mobile audio space supporting multiple simultaneous conversations. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 425–432
2. Basu S (2002) Conversational scene analysis. PhD thesis, MIT Department of EECS, Cambridge, MA
3. Bilmes JA (1998) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report ICSI-TR-97-021, University of Berkeley
4. Bobick A, Intille S, Davis J, Baird F, Pinhanez C, Campbell L, Ivanov Y, Schutte A, Wilson A (1999) The KidsRoom: a perceptually-based interactive and immersive story environment. In: Presence (USA), vol 8, pp 369–393
5. Brdiczka O, Maisonnasse J, Reignier P (2005) Automatic detection of interaction groups. In: Proceedings of the international conference multimodal interfaces, pp 32–36, October 2005
6. Brdiczka O, Vaufreydaz D, Maisonnasse J, Reignier P (2006) Unsupervised segmentation of meeting configurations and activities using speech activity detection. In: Maglogiannis I, Karpouzis K, Bramer M (eds) IFIP international federation of information processing. Artificial intelligence applications and innovations, vol 204. Springer, Boston, pp 195–203
7. Brumitt B, Meyers B, Krumm J, Kern A, Shafer SA (2000) EasyLiving: technologies for intelligent environments. In: Proceedings of the international conference on handheld and ubiquitous computing, pp 12–29
8. Burger S, MacLaren V, Yu H (2002) The ISL meeting corpus: the impact of meeting type on speech style. In: Proceedings of the international conference on spoken language processing, pp 301–304
9. Caporossi A, Hall D, Reignier P, Crowley JL (2004) Robust visual tracking from dynamic control of processing. In: Proceedings of the international workshop on performance evaluation for tracking and surveillance, pp 23–32

10. Choudhury T, Pentland A (2004) Characterizing social interactions using the sociometer. In: Proceedings NAACOS 2004, June 2004
11. Le Gal Ch, Martin J, Lux A, Crowley JL (2001) Smartoffice: design of an intelligent environment. IEEE Intell Syst 16(4): 60–66
12. McCowan I, Gatica-Perez D, Bengio S, Lathoud G, Barnard M, Zhang D (2005) Automatic analysis of multimodal group actions in meetings. IEEE Trans Pattern Anal Mach Intell 27(3): 305–317
13. Muehlenbrock M, Brdiczka O, Snowdon D, Meunier J-L (2004) Learning to detect user activity and availability from a variety of sensor data. In: Proceedings of the IEEE international conference on pervasive computing and communications, March 2004, pp 13–22
14. Oliver N, Rosario B, Pentland A (2000) A Bayesian computer vision system for modeling human interactions. IEEE Trans Pattern Anal Mach Intell 22(8): 831–843
15. Puzicha J, Hofmann Th, Buhmann J (1997) Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In: Proceedings of the international conference on computer vision and pattern recognition, pp 267–272
16. Qian RJ, Sezan MI, Mathews KE (1998) Face tracking using robust statistical estimation. In: Proceedings workshop on perceptual user interfaces, San Francisco
17. Suchman L (1987) Plans and situated actions: the problem of human–machine communication. Cambridge University Press, Cambridge
18. Stiefelhagen R, Steusloff H, Waibel A (2004) CHIL—computers in the human interaction loop. In: Proceedings of the international workshop on image analysis for multimedia interactive services
19. Vaufreydaz D (2001) IST-2000-28323 FAME: facilitating agent for multi-cultural exchange (WP4). European Commission project IST-2000-28323, October 2001
20. Zaidenberg S, Brdiczka O, Reignier P, Crowley JL (2006) Learning context models for the recognition of scenarios. In: Maglogiannis I, Karpouzis K, Bramer M (eds) IFIP international federation of information processing. Artificial intelligence applications and innovations, vol 204. Springer, Boston, pp 86–97
21. Zhang D, Gatica-Perez D, Bengio S, McCowan I, Lathoud G (2004) Multimodal group action clustering in meetings. In: Proceedings of the international workshop on video surveillance & sensor networks