Unsupervised Segmentation of Small Group Meetings using Speech Activity Detection

Oliver Brdiczka, Patrick Reignier, Jérôme Maisonnasse

Laboratoire GRAVIR
INRIA Rhône-Alpes
655 Av. de l'Europe
38330 Montbonnot, France.

{brdiczka, reignier, maisonnasse}@inrialpes.fr

ABSTRACT

This paper addresses the problem of segmenting small group meetings in order to detect different group configurations in an intelligent environment. To track and understand human activity, we need to identify human actors and their interpersonal links. A small group can be seen as basic entity, within which individuals collaborate in order to achieve a common goal. The segmentation of a small group meeting into different small group configurations is an important issue to understand the dynamics of the meeting as well as to detect meeting activity. Our approach takes speech activity detection of individuals attending a meeting as input. The goal is to separate distinct distributions of speech activity observation corresponding to distinct group configurations and activities. We propose an unsupervised method based on the calculation of the Jeffrey divergence between histograms of speech activity 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 speech activity distribution allocation for a given meeting recording. These distinct distributions can be interpreted as distinct segments of group configuration and activity. To evaluate, we recorded 5 small group meetings. We measured the correspondence between detected segments and labeled group configurations and activities. The obtained results are promising, in particular as our method is completely unsupervised.

Categories and Subject Descriptors

I.2.10 **[ARTIFICIAL INTELLIGENCE]**: Vision and Scene Understanding - *Perceptual reasoning.*

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Ubiquitous computing, intelligent environment, speech activity detection, sliding window histograms, Jeffrey divergence curve, successive robust mean estimation.

1. INTRODUCTION

Ubiquitous computing [1] integrates computation into all-day environments. People are enabled to move around and interact with computers more and more naturally. One of the goals of ubiquitous computing is to enable devices to sense changes in the environment and to automatically adapt and act based on these changes. A main focus is laid on sensing and responding to human activity. Human actors need to be identified in order to perceive correctly their activity. In order that ubiquitous computer devices act and interact correctly with users, addressing the right user at the correct moment and perceiving his correct activity is essential.

The focus of this work is analyzing human (inter)action in 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 are strongly linked to changes in activity. The fusion of several independent small groups can be seen as 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 listening 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).

This paper proposes an unsupervised method for detecting changes in small group configuration and activity based on measuring the Jeffrey divergence between adjacent histograms. These histograms are calculated for a window sliding from the beginning to the end of the meeting and contain the frequency of (human) activity events. The peaks of the Jeffrey divergence curve are used to segment distinct distributions of activity events and to find the best model of activity event distributions for the given meeting. The method has been tested on speech activity detection events as sensor information for interacting individuals. We focus thus on verbal interaction. The evaluation has been done with speech activity recordings of 5 meetings of 4 individuals.

2. PREVIOUS AND RELATED WORK

Many approaches for the recognition of human activities in meetings have been proposed in recent years. Most works use supervised learning methods [2], [3], [4], [7]. Some projects focus on supplying appropriate services to the user [7], while others focus on the correct classification of meeting activities [3] or individual availability [4]. Less work has been conducted on unsupervised learning of meeting activities [10].

The recognition of human activity based on speech events is often used in the context of group analysis. The automatic detection of conversations using mutual information [1], in order to determine who speaks and when, needs an important duration of each conversation. 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 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. [2] shows that different meeting activities, and especially different group configurations, have particular distributions of speech activity. In our approach we will focus on an unsupervised method segmenting small group meetings into consecutive group configurations. Our objective is thus to separate automatically distinct distributions of speech activity in small group meeting recordings.

3. APPROACH

We present a novel approach based on the calculation of the Jeffrey divergence [5] between histograms of observations of speech activity. An automatic speech detector [1] generates discrete speech activity observations for several individuals attending a meeting (section 3.1). To separate distinct distributions of speech activity, we slide two adjacent windows 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. We vary the size of the sliding adjacent windows generating several Jeffrey divergence curves (section 3.2). The peaks of the resulting curves are detected using successive robust mean estimation (section 3.3). The detected peaks are merged and filtered with respect to their height and window size (section 3.4). The retained peaks are finally used to select the best model, i.e. the best speech activity distribution allocation for the given meeting recording (section 3.5).

ICMI'05, October 4–6, 2005, Trento, Italy.

3.1 Speech Activity Detection

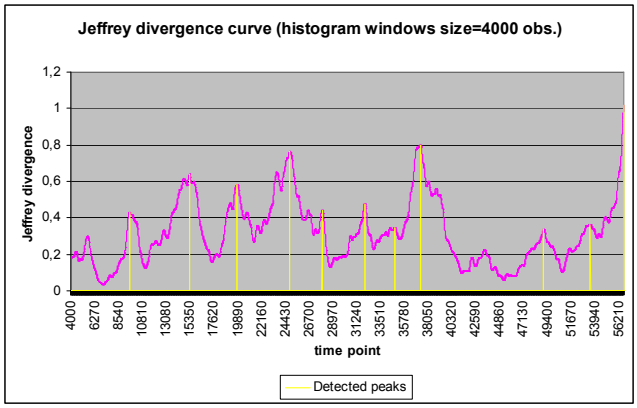
Our approach is based on speech activity detection of individuals attending a meeting. We are recording the speech of each individual using lapel microphones. We admit the use of lapel microphones in order to minimize detection errors. An automatic speech detector [1] parses the audio channels of the different lapel microphones and detects which individual stops and starts speaking. The observations used in our approach are a discretization of speech activity events sent by this detector. One observation is a vector containing a binary value (speaking, not speaking) for each individual that is recorded. This vector is transformed to a 1-dimensional discrete code used as input. The automatic speech detector has a sampling rate of 62.5 Hz, which corresponds to the generation of an observation every 16 milliseconds.

3.2 Speech Activity Distributions

In [2], the authors stated that the distribution of the different speech activity observations is discriminating for group configurations in small group meetings. Thus we assume that in small group meetings distinct group configurations and activities have distinct distributions of speech activity 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 of speech activity are discrete and unordered (a 1-dimensional discrete code) and we do not want to admit any a priori distribution, we use histograms to represent speech activity 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 speech activity distributions, we calculate the Jeffrey divergence [5] between the histograms of two adjacent observation windows. The Jeffrey divergence 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).



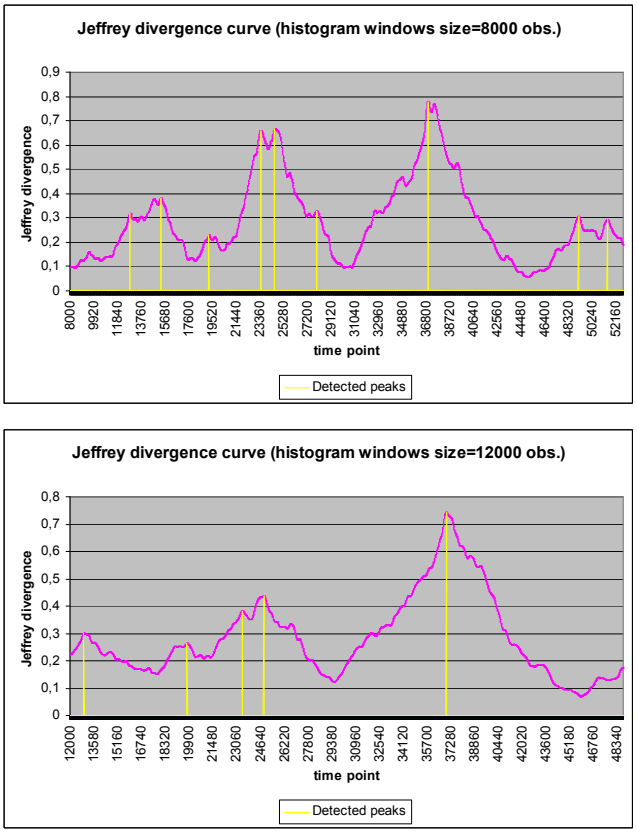


Figure 1. Meeting 5: Jeffrey divergence between histograms of sliding adjacent windows of 4000, 8000 and 12000 observations (64sec, 2min 8sec and 3min 12sec)

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 (sizes of 4000, 6000, 8000, 10000, 12000, 14000 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 speech activity distribution of the small meeting recording.

3.3 Peak Detection

To detect the peaks of the Jeffrey divergence curve, we use successive robust mean estimation. Robust mean estimation has been used in [6] to locate the center position of a dominant face in skin color filtered images. Figure 2 describes the robust mean estimation process to detect the dominant peak of the Jeffrey curve. Mean and standard deviation are calculated for the Jeffrey curve using the formulas of figure 3. Note that time is discrete within these formulas as we have discrete observations and that mean and peak position correspond to a time point and the standard deviation to a time interval. To detect all peaks of the Jeffrey divergence curve, we apply the robust mean estimation process successively (figure 4).

Step 1. Compute mean μ and standard deviation σ based on all the points of the Jeffrey curve (figure 3).

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

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 $[\mu^*-\delta^*, \mu^*+\delta^*]$.

Figure 2. Robust mean estimation process detecting a dominant peak of the Jeffrey divergence curve

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

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

with

$J_{h[t-size,t],\, h[t,t+size]}$ = Jeffrey divergence between adjacent histograms of size *size* at time point *t*

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

Figure 3. Mean and standard deviation formulas for the Jeffrey divergence curve

Step 1. Detect dominant peak p^* using robust mean estimation (figure 2)

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

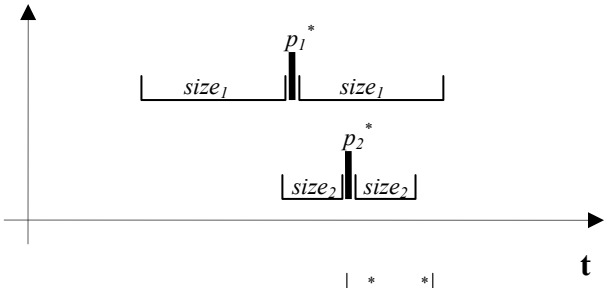
Step 3. Repeat Steps 1 and 2 until $J_{h[p^*-size,p^*],\, h[p^*,p^*+size]} \leq \bar{J}$ with

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

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

3.4 Merging and Filtering Peaks from different Window Sizes

Peak detection using successive robust mean estimation (section 3.3) is conducted for Jeffrey curves with histogram window sizes of 4000, 6000, 8000, 10000, 12000, 14000 and 16000 observations. 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.



$$n\_dist(p_1^*, p_2^*) = \frac{|p_1^* - p_2^*|}{\min(size_1, size_2)}$$

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

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 (figure 5). 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 speech activity distributions, i.e. to search for the best model for a given meeting.

3.5 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. We accept the peak combination that maximizes the average divergence between the peak histograms as the best model for the given meeting. Figure 6 shows the output of the model selection algorithm.

Data size (nb obs) = 60694

position	rel. peak height	window size	votes
11800	0.98	8000.0	11.0
30930	1.0	4000.0	12.0
39700	1.0	6000.0	21.0
47830	0.63	4000.0	8.0

searching for best model ... 16 combinations:

0 (0.29569755320228747) :30930 39700

1 (0.26898563670964654) :11800 30930 39700

2 (0.23388392790545345) :30930 39700 47830

3 (0.2248771197707075) :11800 30930 39700 47830

4 (0.18113869241523864) :11800 39700

5 (0.16278942504749674) :11800 30930 47830

6 (0.15893069429274265) :11800 47830

7 (0.1568547077635417) :11800 39700 47830

…

Figure 6. Meeting 3: Output of the model selection algorithm. The best combination has the highest average Jeffrey divergence indicated between the parentheses.

4. EVALUATION AND RESULTS

The result of our approach is the peak combination separating best the speech 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 5 small group meetings. 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* measures proposed in [10].

Figure 7. Small group meeting of 4 individuals recorded for the experiments

4.1 Experiments

To evaluate our approach, we recorded 5 small group meetings of 4 individuals (figure 7). The number and order of group configurations, i.e. who will speak with whom, was fixed in advance for the experiments. 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.

4.2 Evaluation measures

To evaluate, 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.

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

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

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*

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

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

Figure 8. Average segment purity (*asp*), average activity purity (*aap*) and the overall criterion *Q*

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 8). 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. The *Q* criterion is an overall evaluation criterion combining *asp* and *aap*, where larger *Q* indicates better overall performance.

4.3 Results

Figures 9-13 show the labeled group configurations for each small group meeting as well as the segments detected by our approach. Table 1 indicates the *asp*, *aap* 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. However, our approach worked well for meeting recording 2, while the segmentation of meeting recording 3 is mediocre. The overall results of our approach are very good; the average *Q* value is 0.80. By excluding meeting 3, we even obtain a *Q* value of 0.87.

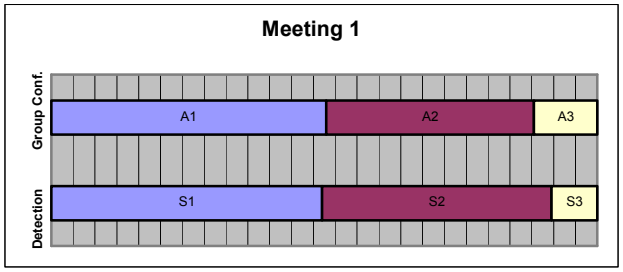


Figure 9. Group configurations and their detection for Meeting 1 (Q=0.93, meeting duration=9min 14sec).

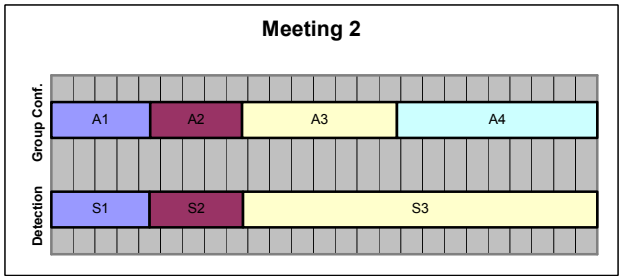


Figure 10. Group configurations and their detection for Meeting 2 (Q=0.81, meeting duration=10min 14sec).

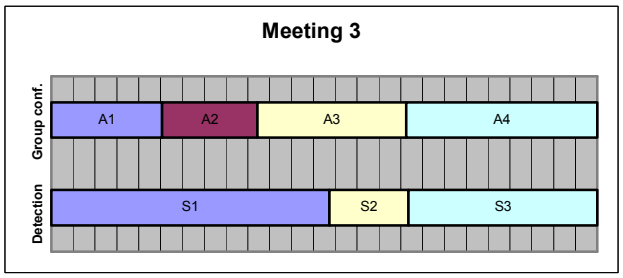


Figure 11. Group configurations and their detection for Meeting 3 (Q=0.51, meeting duration=16min 11sec).

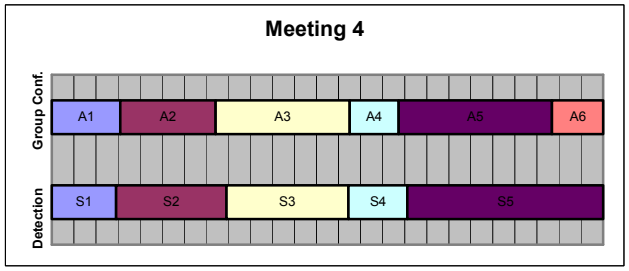


Figure 12. Group configurations and their detection for Meeting 4 (Q=0.84, meeting duration=14min 47sec).

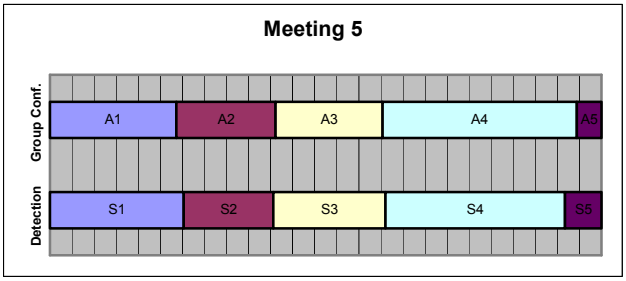


Figure 13. Group configurations and their detection for Meeting 5 (Q=0.92, meeting duration=16min 12sec).

Table 1. *asp*, *aap* and *Q* values for the recorded meetings.

	asp	aap	Q
Meeting 1	0.93	0.92	0.93
Meeting 2	0.67	0.99	0.81
Meeting 3	0.42	0.62	0.51
Meeting 4	0.78	0.91	0.84
Meeting 5	0.92	0.91	0.92
Average	0.74	0.87	0.80
Average w/o Meeting 3	0.83	0.94	0.87

5. CONCLUSION

We proposed an unsupervised method for segmenting small group meeting configurations and activities. This method is based on the calculation of the Jeffrey divergence between histograms of observations of speech activity. The peaks of the Jeffrey divergence curve are used to separate distinct distributions of speech 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. The obtained results are promising, in particular as our method is completely unsupervised.

Further meeting recordings need to be done in order to apply and evaluate our method on more and subtler meeting activities. These meeting activities will include activity changes within a group configuration.

Our method can help obtaining a first segmentation of a meeting. The detected segments can then be used as input for further classification tasks like meeting comparison, meeting activity recognition etc.

Future work will concern the test of our method on further meeting information. Speech activity detection is not sufficient to disambiguate all situations. Further information like head orientation, pointing gestures or interpersonal distances seem to be good indicators. Thus a multimodal approach needs to be envisaged. The method can easily be extended to such an approach as we only need to upgrade the observation codes used for the generation of the histograms.

6. REFERENCES

[1] Basu S., *Conversational Scene Analysis,* Ph.D. Thesis. MIT Department of EECS. September, 2002.

[2] Brdiczka, O., Maisonnasse, J., and Reignier, P., *Automatic Detection of Interaction Groups,* Proc. Int'l Conf. Multimodal Interfaces, 2005 (to appear).

[3] 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, vol. 27, no. 3, pp. 305-317, March 2005.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Vaufreydaz, D., *IST-2000-28323 FAME: Facilitating Agent for Multi-Cultural Exchange (WP4)*, European Commission project IST-2000-28323, October 2001.

[9] Weiser, M., *Ubiquitous Computing: Definition 1*, http://www.ubiq.com/hypertext/weiser/UbiHome.html, March 1996.

[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.