Capturing Interactions in Meetings with Omnidirectional Cameras

Rainer Stiefelhagen[1], Xilin Chen[2], Jie Yang[2]
Interactive Systems Laboratories
[1]Universität Karlsruhe (TH), Germany
[2]Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

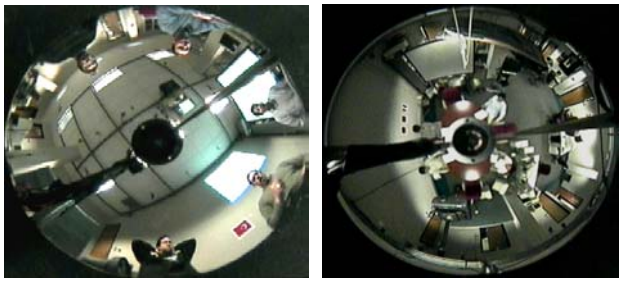
Human interaction is one of the most important characteristics of meetings. To explore complex human interactions in meetings we must understand them and their components in detail. In this paper we present our efforts in capturing human interactions in meetings using omnidirectional cameras. We present algorithms for person tracking, head pose estimation, and face recognition from omnidirectional images. We also discuss an approach for the estimation of who was talking to whom based on tracked head poses of the participants. Finally, we address the problem of activity modeling based on moving trajectories of people in a meeting room. We report experimental results to demonstrate the feasibility of the presented technologies and discuss future work.

1 Introduction

Meetings are an important part of daily life in governments, companies, universities, and other organizations. Most people find it impossible to attend all relevant meetings or to retain all the salient points raised in meetings. In the past few years, many researchers have been attempting to find various ways to lessen problems in meetings. Xerox has developed a media-enabled conference room equipped with cameras and microphones to capture audio-visual content. [Chiu 99]. In the NIST Smart Space Lab has set up another smart meeting room [Rosenthal 00]. At Microsoft research, some work has been conducted on capturing small group meetings using a ring camera [Rui 01]. The University of California, San Diego has also developed a meeting system equipped with several fixed calibrated cameras, some active cameras, and several omnidirectional cameras, and the system is able to track people in the room, recognizes their faces and is able to identify the current speaker [Mikic 00]. At the Interactive Systems Laboratory of Carnegie Mellon University we have been developing technologies for intelligent meeting room since 1997 [Waibel 98, Yang 99, Stiefelhagen 99, Waibel 03].

Meetings between people are events that encode a large amount of social and communicative information. To decode such information it requires understand multimedia information from multiple cues. One of important characteristics for a meeting is human interaction. In this research, we focus on how to use visual information to understand human interactions in meetings. In fact, we humans decode easily from visual scanning and observing our environment in a meeting. For example, in the same scene and from the same video stream, we identify simultaneously people, we understand what they are doing, why they are doing it, to whom and with whom they are interacting, what their mutual relationships are, what their social relationships, roles, and styles are, what their feelings, concern, interests are, how they are carrying out tasks over the period of time. For machine perception, however, human interactions have to be understood and described at multiple levels and in terms of multiple functionalities and perspectives. Loosely speaking, to understand human interactions the system must provide answers to the questions of the Who? Where? Why? When? To/With Whom, and How? the interaction happens.

In this paper, we present our efforts in capturing human interaction in a meeting using omnidirectional cameras. We can put an omnidirectional camera on the meeting table and/or mount it on the ceiling. Figure 1 shows an example of omnidirectional cameras in a meeting room. An omnidirectional camera has about 180 degree viewing angle. From the camera on a meeting table, the system is able to capture faces of all participants. By further processing captured faces, the system can obtain the information on who are in the meeting, where they are located, who is looking at whom. From the camera on the ceiling, the system can observe activities in a larger area around meeting table. The system, by virtue of being above the 'action,' is less prone to problems of occlusion. Although an omnidirectional camera has a limited resolution, we can till use it for capturing much useful information for modeling human interaction in meetings.



(a) On the table (b) On the ceiling

Figure 1. Omnidirectional cameras in a meeting room

The rest of this paper is organized as follows: In Section 2, we discuss robust tracking algorithms for tracking meeting participants, and their head poses using omnidirectional cameras. In section 3, we address problems on who, what (activity analysis), and who is talking to whom in a meeting. In section 4, we report experimental results on the described technologies. In Section 5 we conclude the paper and discuss the future work.

2 Person Tracking and Identification

Robust tracking meeting participants, their head poses, and their identities is essential for a system to provide information on who is in the meeting, where they located, what they are doing, who is talking to whom, and when interactions happen. In this section, we will discuss technologies for tracking people, their head poses, and recognizing their faces using omnidirectional cameras on the meeting table and ceiling of the meeting room.

2.1 Tracking People from an Omnidirectional Camera Mounted on Ceiling

An omnidirectional camera mounted on the ceiling prone to problems of occlusion. However, there are many challenges for robustly tracking people using such a camera, such as changing background and non-uniformed view. One of the fundamental problems in people tracking systems is to extract all objects in the scene from the background. Background subtraction has been widely used for this task. Many different models have been proposed to characterize the background, such as pixel interval estimation [Haritaoglu 89], single Gaussian Model [Karmann 90, Toyama 99, Wren 97], and Gaussian Mixture Model (GMM) [Friedman 97, Grimson 98]. All these models, however, cannot evolve over the time.

An omnidirectional camera mounted on the ceiling covers much larger area than that of a normal perspective camera. This makes the scene is more complex. In addition, many factors may cause the background to change during a meeting, such as:

1. Suddenly scene change, such as lights being turned on or off during a meeting, and at the system starting point;
2. Slow environment change, such as a meeting room with windows at different time of a day;
3. Fade in/fade out lighting change, such as the shadow of a moving object;
4. Partly background update, such as a chair has been moved during a meeting.

Therefore, a more sophisticated model is needed to handle such a dynamic environment. A good background model should have the capability of handling all or most of the above situations. A solution is to use an adaptive model. Instead of assuming the known background before tracking, we can build the background gradually by adaptation. From the mathematic point of view, a background can be considered as a field change over the time. The Markov Random Field is a natural way to describe the evolvement of the background. We have proposed to use MRF models to represent both foreground and background [Chen 02]. The basic assumption that supports this method is that the background is statistically stable. Unlike some of the previous methods, we do not assume that the background is known before the tracking process. The background model can be gradually generated during the tracking process. We describe the model in detail below.

A background can first be regarded as a 2D field with a limited support set, and evolve over time *t* as illustrated in Figure 2. An image within a sequence of images can be regarded as the background image covered with some objects and noises added.

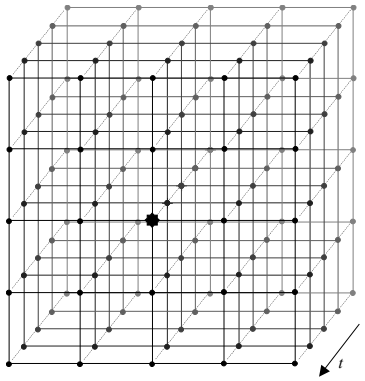


Figure 2. 2D background grid evolving along with time

Suppose the support set of the image is $\Lambda = \{(1,1), (1,2), \cdots, (1,n), (2,1), \cdots, (m,n)\}$ and m, n are the height and width of the image respectively. The support set of objects i at time t is $\Lambda_t^i \subset \Lambda$. The background support set at time t is $\Lambda_t = \Lambda \setminus \{\cup_i \Lambda_t^i\}$.

Assume that B_t is the ideal background image at time t, the object i in a 2-D image is O_t^i at time t, I_t is the observed image at time t. Therefore, the relationship among them is as equation (1).

$$I_t(\mathbf{X}) = \begin{cases} B_t(\mathbf{X}) + n_t(\mathbf{X}) & \mathbf{X} \in \Lambda_t \\ O_t^i(\mathbf{X}) + n_t(\mathbf{X}) & \mathbf{X} \in \Lambda_t^{O_t^i} \end{cases}, \quad (1)$$

where $n_t(\mathbf{X})$ is the noise at position $\mathbf{X}$ and time t. Therefore, the visual surveillance problem can be defined as: given an observed image $I_t(\mathbf{X})$, the background $B_{t-1}(\mathbf{X})$, and the object set $\{O_{t-1}^i\}$ at time t-

1, how can we obtain the best estimation of the background $B_t(\mathbf{X})$ and the object set $\{O_{t-1}^i\}$ at time t. This goal can be achieved by:

$$\arg\max_{B_t,\{O_t^i\}} \{P(B_t, \{O_t^i\} \mid I_t, B_{t-1}, \{O_{t-1}^j\})\}. \quad (2)$$

Note that the object set at time t can be different from the object set at time t-1. Consider that the objects are independent to the background, we have:

$$\begin{aligned} &\arg\max_{B_t,\{O_t^i,\Lambda_t^i\}} \{P(B_t, \{O_t^i, \Lambda_t^i\} \mid I_t, B_{t-1}, \{O_{t-1}^j, \Lambda_{t-1}^j\})\} \\ &= \arg\max_{B_t,\Lambda_t} \{P(B_t, \Lambda_t \mid I_t, B_{t-1}, \{O_{t-1}^j, \Lambda_{t-1}^j\})\} \\ &\cdot \arg\max_{\{O_t^i,\Lambda_t^i\}} \{P(\{O_t^i, \Lambda_t^i\} \mid I_t, B_{t-1}, \{O_{t-1}^j, \Lambda_{t-1}^j\})\} \end{aligned} \quad (3)$$

The first item is the estimated background, and the second one is the estimation of the objects. If we apply a first order MRF model, the tracking problem in Equation (3) can be formulated as minimizing the following Equation (4).

$$\arg\min_{B_t,\{O_t^i,\Lambda_t^i\}} \left\{ \begin{aligned} &\sum_{\mathbf{X}\in\Lambda_t} E_N(\mathbf{X}) + E_P(\mathbf{X})/T \\ &+ \sum_i \sum_{\mathbf{X}\in\Lambda_t^i} E_N(\mathbf{X}, \dot{\mathbf{X}}_t^i) + E_P(\mathbf{X}, \dot{\mathbf{X}}_t^i)/T_i \end{aligned} \right\}, \quad (4)$$

where T and T_i are update speed factors, and

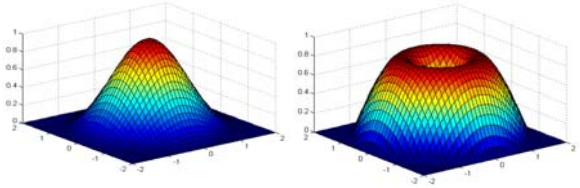
$$E_N(\mathbf{X}) = \frac{[B_t(\mathbf{X}) - I_t(\mathbf{X})]^2}{\sigma^2},$$

$$E_P(\mathbf{X}) = \left[\frac{\partial B_t(\mathbf{X})}{\partial t}\right]^2 + \left[\frac{\partial^2 B_t(\mathbf{X})}{\partial t \partial x}\right]^2 + \left[\frac{\partial^2 B_t(\mathbf{X})}{\partial t \partial y}\right]^2$$

$$E_N(\mathbf{X}, \dot{\mathbf{X}}_t^i) = \begin{cases} \dfrac{[B_t(\mathbf{X}) - B_{t-1}(\mathbf{X} - \dot{\mathbf{X}})]^2}{\sigma^2} & \text{iff. } \mathbf{X} \in \Lambda_t^i \text{ and } \mathbf{X} - \dot{\mathbf{X}} \in \Lambda_{t-1}^i \\ \left[\dfrac{\max(B_t(\mathbf{X}), B_{t-1}(\mathbf{X} - \dot{\mathbf{X}}))}{\sigma}\right]^2 & \text{otherwise} \end{cases}$$

$$E_P(\mathbf{X}, \dot{\mathbf{X}}_t^i) = [B_t(\mathbf{X}) - B_{t-1}(\mathbf{X} - \dot{\mathbf{X}})]^2 + \left[\frac{\partial B_t(\mathbf{X})}{\partial x} - \frac{\partial B_{t-1}(\mathbf{X} - \dot{\mathbf{X}})}{\partial x}\right]^2 + \left[\frac{\partial B_t(\mathbf{X})}{\partial y} - \frac{\partial B_{t-1}(\mathbf{X} - \dot{\mathbf{X}})}{\partial y}\right]^2$$

One of the basic assumptions for this model is that the pixel mesh is unified. Therefore an adaptation between the unified mesh and omni-view is needed if we apply this model for tracking people in an omnidirectional camera. There are two different ways of adaptation: the first is to convert the image into a uniformed view, such as transforming it into a paranormal view. But the system will loss some useful information near the center area; and the second is to adapt the model to the image. Based on the characteristic of the camera we used, we can compensate the non-linearity based on the mapping shown in Figure 3. Figure 3(a) is the factor in direction X and Y and Figure 3(b) is the factor in direction Z which is the optical axis direction. We assume that the object movement is only in direction X and Y in most cases. The captured scene in the image is located within a circle, whose radius is 2f, and f is the focus length of the paraboloid. The object's dimension in direction X and Y will be maximum when the object is located on the optical axis, and will disappear at the circle. The object's dimension in direction Z will be invisible at the center and the circle, and will reach maximum at the radius 0.8629*f*. We can estimate changes in size for each object using such a match factor map.



(a) factor in X-Y direction (b) factor in Z direction

Figure 3. Horizontal and vertical match factor map

2.2 Simultaneous Head Pose Tracking

In order to simultaneously track head poses of meeting participants, we use an omnidirectional camera to capture the scene around a meeting table. In the panoramic view of the meeting scene (see Figure 4 for an example) we then detect the participants' faces by searching for skin-colored regions and use some heuristics to distinguish skin-colored hands from faces [Stiefelhagen 2000].

Figure 4. Panoramic view of a meeting scene

For each detected participant a rectified (perspective) view is computed (see Figure 5). Faces extracted from these views are then used to estimate each participant's head pose.

Figure 5. Perspective views of three participants

We use neural networks to estimate head pan and tilt from such facial images [Stiefelhagen 00]. In our approach, preprocessed facial images are used as input to the neural networks, and the networks are trained so as to estimate the horizontal pan) or vertical (tilt) head orientation of the input images. Separate networks were trained to estimate head pan and tilt. These networks contained one hidden layer, and one output unit that encodes the head orientation in degrees. By training multi-user networks on images from twelve users we achieved average estimation errors as low as three degrees for pan and tilt. On images from new users, head orientation could be estimated with an average error of 10 degrees for pan and tilt. More details can for be found in [Stiefelhagen 2000]

2.3 Face Recognition from a Panoramic View

Face recognition has been an active research area in the last two decades. The progress in this area can be found in review papers [Chellappa 95, Samal 92] and the proceedings of the last five international conferences on Automatic Face and Gesture Recognition. A major challenge for face recognition from a panoramic view is difficulties in face alignment due to low resolution of images. For a holistic template matching approach, e.g., Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA), facial features such as eyes are commonly used for aligning faces. In a panoramic view of the omnidirectional camera that we used for meeting capturing (640x480 pixels), we cannot robustly detect facial features for alignment. We employ a detection-based method for face alignment.

We detect faces using a PCA based method with different scales in the panoramic image. To train the face subspace, we use 400 faces image crop from training sequence. Figure 6 is the first 24 eigenfaces of the subspaces. In the detection process, we project the candidate area to the space and use these projection values to reconstruct a new image. We then measure the distance between the reconstructed image and original one. For a non-face image, the distance is larger. Figure 7 is an example of face and non-face and their reconstructions.

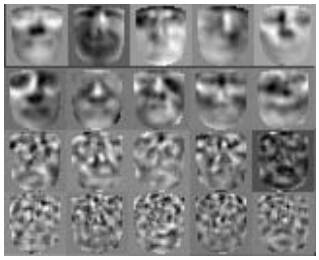


Figure 6. The first 24 eigenfaces for face detection

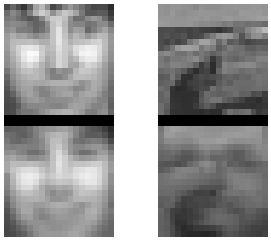


Figure 7. Face, non-face and their reconstructions (The first row is the original images, and the second row is the reconstruction ones)

To obtain an accurate position of a face, we use a two-steps method. In the first step, we determine a rough position of the face, and then we search at a sub-pixel level to obtain the optimal position and orientation in second step.

Continuously identifying people in a meeting room is a challenging task [Yang 1999]. We have previously developed technologies for face recognition in a meeting room [Yang 99, Yang 00, Gross 00a, Gross 00b]. Under a constrained scenario such as a few people around a meeting table and faces are captured from an omnidirectional camera, PCA or *eigenfaces* [Turk 91] can perform well. Since we have used PCA for the face alignment, we use PCA for face recognition. We will present the evaluation result in next section.

3 Tracking and Modeling Interaction

In the previous sections we have discussed technologies that can help answering questions such as, "Who is in the meeting?" (person identification), "Where are they in the room?" (person locating & tracking), and "Where did someone look?" (focus of attention). Given answers to (or better: hypotheses about) these basic questions, it is possible to speculate about meeting actions and interactions, and the ways in which they are performed.

3.1 From Head Pose to Focus of Attention

Knowing who is talking to whom is an important cue both for the understanding and indexing of meetings as well as for videoconferencing applications. In our research we have addressed the problem of tracking who is looking at whom during meetings. There are two contribution factors of where a person looks: head orientation and eye orientation. In this work head orientation is considered as a sufficient cue to detect a person's direction of attention. Relevant psychological literature offers a number of convincing arguments for this approach (e.g. [Emery 00, Argyle 76, Cranach 71]) and the feasibility of this approach has previously been demonstrated experimentally [Stiefelhagen 02a, Stiefelhagen 02b]. Our approach to tracking at whom participants look, i.e. their focus of attention, is the following: 1) Detect all participants in the scene, 2)

estimate each participant's head orientation and 3) map each estimated head orientation to its likely targets using a probabilistic framework.

We have developed a Bayesian approach to estimate at which target a person is looking, based on his observed head orientation [Stiefelhagen 01, Stiefelhagen 02]. More precisely, we wish to find $P(Focus_s = T_i \mid x_s)$, the probability that a subject s is looking towards a certain target person T_i, given the subject's observed horizontal head orientation x_s, which is the output of the neural network for head pan estimation. Using Bayes formula, this can of be decomposed into

$$P(Focus_S = T_i \mid x_S) = \frac{p(x_S \mid Focus_S = T_i) \cdot P(Focus_S = T_i)}{p(x_S)}, \quad (5)$$

where x_s denotes the head pan of person s in degrees and T_i is one of the other persons around the table.

In order to compute $P(Focus_s \mid x_s)$, it is necessary, to learn the class-conditional probability density function $p(x_s \mid Focus_s = T)$, the class prior $P(Focus_s = T)$ and $p(x_s)$ for each person. Finding $p(x_s)$ is trivial and can be done by just building a histogram of the observed head orientations of a person over time.

We have developed an unsupervised learning approach to find the class-conditional head pan distributions of each participant. In our approach, we assume that the class-conditional head pan distributions can be modeled as Gaussian distributions. Then, the distribution $p(x)$ of all head pan observations from a person will result in a mixture of Gaussians,

$$p(x) \approx \sum_{j=1}^{M} p(x \mid j) P(j), \quad (6)$$

where the individual component densities $p(x \mid j)$ are given by Gaussian distributions $N(\mu_i, \sigma_i^2)$. The number of Gaussians M is set to the number of other participants that are detected around the table. The parameters of the mixture model can be adapted so as to maximize the likelihood of the pan observations given the mixture model. This can be done using the *EM* algorithm (for further details see [Stiefelhagen 01]).

After adaptation of the mixture model (6), we use the resulting individual Gaussian components as an approximation of the class-conditionals $p(x \mid Focus = T)$ of our focus of attention model described in equation (5). We furthermore use the priors of the mixture model as the focus priors $P(Focus = T)$. To assign the individual Gaussian components and the priors to their corresponding target persons, the relative position of the participants around the table are used. Figure 8 depicts the adaptation result for one user. One the left side, the true class-conditional head pan distributions are depicted together with the learned class-conditionals. On the right side, the resulting learned posterior distributions are shown.

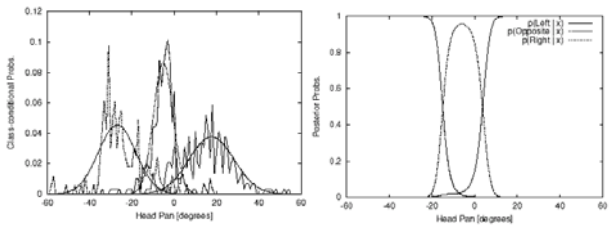
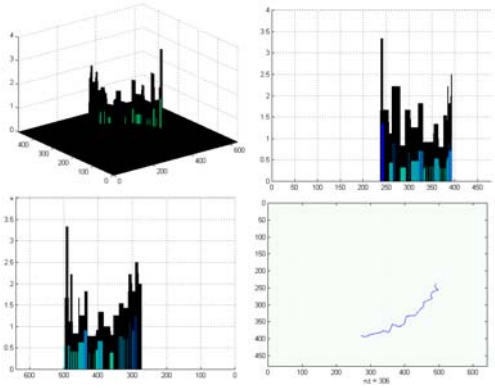


Figure 8. Learned class-conditional head pan distributions (left) and resulting posteriors (right).

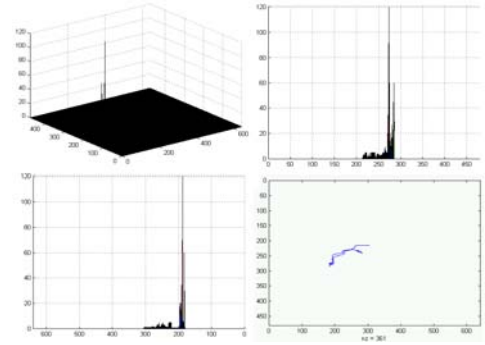
3.2 Activity and Scene Modeling from Moving Trajectories

Although an omnidirectional camera has a limited resolution, we can use it for analyzing some simple activities and interactions. The basic idea is to analyze human activities and interactions using moving trajectories. We define a hierarchical behavior model. At the lowest level of the model, it contains essential information such as moving or stopping and sitting or standing, which can be observed from the tracking sequence. At a higher level we can distinguish some different activities, such as working alone or having a meeting, etc., which can't be observed directly. These activities can be observed via tracking moving trajectories of people in a scene. For example, we can define a meeting as two or more trajectories coming from same or different directions and staying in the scene for a period of time.

We have tested the idea on the limited dataset. We collected 1 hour of video data from the omnidirectional camera mounted on the ceiling of our meeting room. We input the video into the people tracking system and obtained moving trajectories of people. We then used a time-spatial window to analyze individual trajectories. Five seconds' duration was used as the time window. The trajectory within this time window formed a spatial window. The time overlap window was used for each clip, and the overlap time was 2.5s. If the object stayed at a spot for a period of time, the moving trajectories would be accumulated into histograms. These histograms could have different patterns corresponding to different activities. We could then infer human activities from the histograms. Figure 9 shows some patterns for different activities. The top-left in each group is the spatial histogram. The top-right and bottom-left are horizontal and vertical view of the histogram, and the bottom-right is the top view of the trajectory.



(a) Walking through clip



(b) Looking for something clip

Figure 9. Examples of activity patterns

4 Experimental Results

We have performed various experiments to evaluate the technologies. We present some experimental results in this section.

4.1 Experiments for People Tracking

We have tested the system's ability to initialize the background model with objects in the scene, track and monitor an object with changes in size and lighting, and track multiple objects. Figure 10 is an example of background building with an object in the scene. The first row of images is the images correspondence to the beginning and the end of background setup. The other two rows are the under building background. At the beginning of the background setup, some black areas within the circle are under construction, and we can dynamically update the background while tracking the object and obtain a complete background when the object is out of the scene.

Figure 11 is an example of multi-object tracking. Only the top left one with the background, the others are only moving objects. Although one of the objects passes through the blind zone as the frame 155, it is continuously tracked by the system. In this example, we track both stable and moving object at the same time. It can be seen that the size and lighting of the moving object in the tracking sequence change with positions.

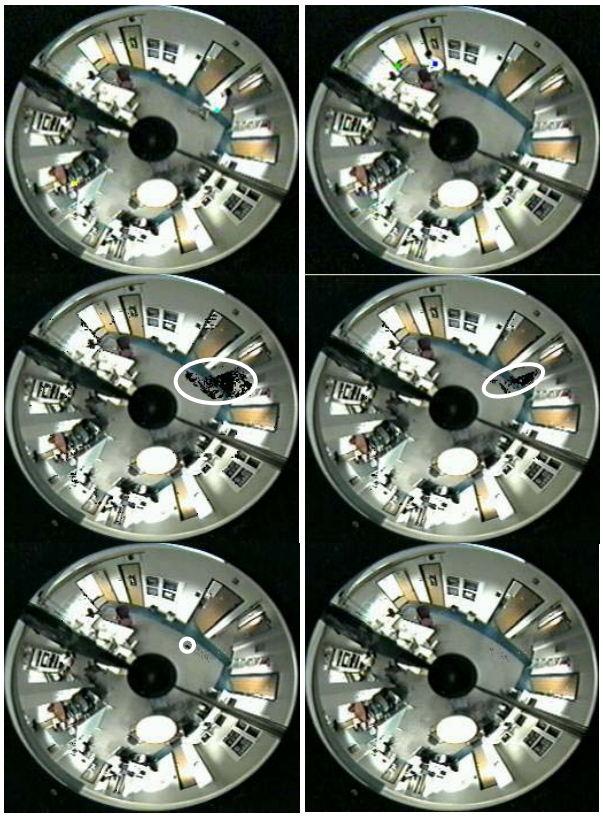
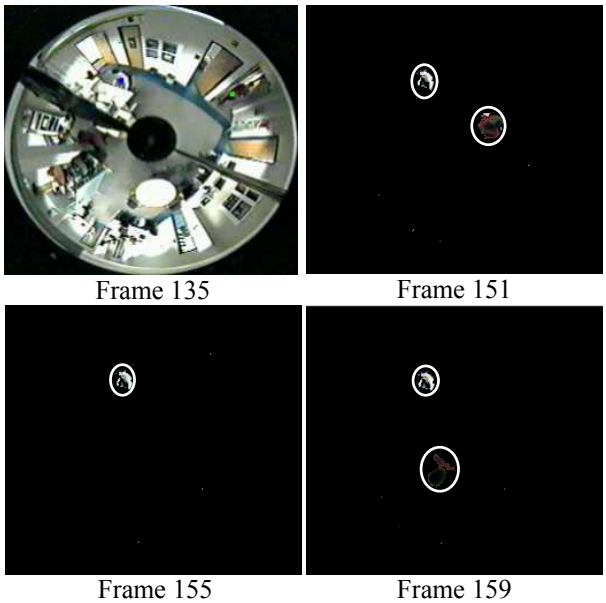


Figure 10. Background evolutions with time changing



Frame 135 Frame 151

Frame 155 Frame 159

Figure 11. An example of multi-object tracking

4.2 Experiments on Face Recognition

In our experiments, we test the average response time for detections and recognition accuracy. We define the response time as the time from a participant sit on his/her chair to the time when his/her face is detected. The average response time is 1.52s, which is based on 32 participant' sequences.

In consideration that people in meeting room can be tracked, the recognized people can be attach an ID on their continuously trajectories, we define the recognition rate as the following:

$$\gamma = \frac{n}{N},$$

where n is the number of correctly recognized people with a duration T, from the time when they sit on their chair, and N is the total number of the participants.

We have 10 classes in our experiment, and the recognition results with different time duration are shown Figure 12.

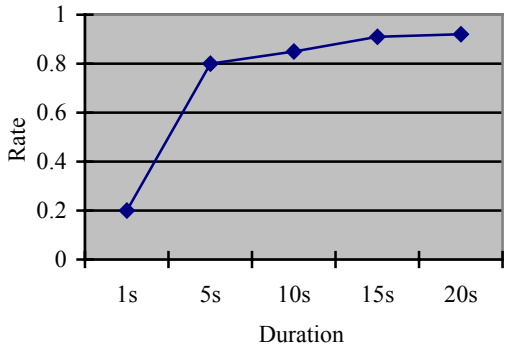


Figure 12. The recognition accuracy over the time

4.3 Experimental Results for Focus of Attention Tracking

We evaluated our approach on several meetings that we recorded. In each of the meetings four participants were sitting around a table and were discussing a freely chosen topic. Video was captured with the panoramic camera and audio was recorded using several microphones. In each frame we manually labeled at whom each participant was looking. These labels could be one of *"Left", "Right"* or *"Straight"*, meaning a person was looking to the person to his left, to his right, or to the person at the opposite. If the person wasn't looking at one of these targets; e.g., the person was looking down on the table or was staring up to the ceiling, the label *"Other"* was assigned. In addition, labels indicating whether a person was speaking or not, were manually assigned for each participant and each video frame. Table 1 shows the evaluation results on the four recorded meetings. In the table, the average accuracy on the four participants in each meeting is indicated.

Table1. Percentage of correctly assigned focus targets.

Meeting	A	B	C	D	Avg.
Accuracy	68.8%	73.4%	79.5%	69.8%	72.9%

For the evaluation the faces of the participants were automatically tracked. Head pan was then computed using the neural network-based system to estimate head orientation. For each of the meeting participants, the class-conditional head pan distribution $p(x|Focus)$, the class-priors $P(Focus)$ and the observation distributions $p(x)$ were adapted as described in the previous section, and the posterior probabilities $P(Focus = T_i \mid x)$ for each person were computed. During evaluation, the target with the highest posterior probability was then chosen as the focus of attention target of the person in each frame. For the evaluation, we manually marked frames where a subject's face was occluded or where the face was not correctly tracked. These frames were not used for evaluation. Face occlusion occurred in 1.6% of the captured images. Occlusion sometimes happened, when a user covered his face with his arms or with a coffee mug for example; sometimes a face was occluded by one of the posts of the camera. In another 4.2% of the frames the face was not correctly tracked. We also did not use frames where a subject did not look at one of the other persons at the table. This happened in 3.8 % of the frames. Overall 8.2% of the frames were not used for evaluation since at least one of the above indications was given.

5 Conclusions

We have presented our efforts in capturing interactions in meetings using omnidirectional cameras. We discussed approaches that provide answers to the questions Who (Face Recognition), What (Activity Classification), Where (Person Tracking) and To/With Whom (Focus of Attention) in meeting situations and we presented some experimental results. In our future work, we aim at improving the robustness of the individual technologies. In addition, we plan to combine the use of omnidirectional cameras, perspective cameras, and actively controlled cameras for the analysis of meetings.

Acknowledgements

We would like to thank our colleagues in the Interactive Systems Laboratories for their helpful discussions and technical support. This research was partially supported by the European Union within the IST project FAME under grant No. IST-2000-28323, the National Science Foundation (USA) under Grant No. IIS-0121560, and the Department of Defense (USA) through award number N41756-03-C4024.

References

[Argyle 76] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[Chen 02] X. Chen and J. Yang, Towards monitoring human activities using an omnidirectional camera, Proceedings of ICMI 2002, 2002.

[Chellappa 95] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. Proceedings of the IEEE, 83(5):705-740, 1995.

[Chiu 99] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox. Meeting Capture in a Media Enriched Conference Room. In Proceedings of the Second International Workshop on Cooperative Buildings (CoBuild'99). pp. 79-88, 1999.

[Cranach 71] M. von Cranach. The role of orienting behaviour in human interaction. In A. H. Esser, editor, *Environmental Space and Behaviour*. Plenum Press, New York, 1971.

[Emery 00] N. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604, 2000.

[Froedan 97] N. Friedman and S. Russell, Image segmentation in video sequences: A probabilistic approach, presented at the 13th Conf. Uncertainty in Artificial Intelligence, 1997.

[Grimson 98] W. E. L. Grimson, C. Stauffer, and R. Romano, "Using adaptive tracking to classify and monitor activities in a site," Proceedings of CVPR 1998, pp. 22–29, 1998.

[Gross 00a] R. Gross, J. Yang, A. Waibel, "Face Recognition in a Meeting Room," Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'2000).

[Gross 00b] R.Gross, J. Yang, A. Waibel, "Growing Gaussian Mixture Model for Pose Invariant Face Recognition," International Conference on Pattern Recognition (ICPR 2000), Barcelona, Spain, Sept. 2000.

[Haritaoglu 98] I. Haritaoglu, D. Harwood, and L. S. Davis. W^4 - a real time system for detection and tracking people and their parts. In Proceedings of International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998.

[Karmann 90] K.-P. Karmann, A. V. Brandt, and R. Gerl, Moving object segmentation based on adaptive reference images, in Signal Processing V: Theories and Application. Amsterdam, The Netherlands: Elsevier, 1990.

[Mikic 00] I. Mikic, K. Huang, M. Trivedi, "Activity monitoring and summarization for intelligent environments", Workshop on Human Motion, 2000

[Rosenthal 00] L. Rosenthal, V. Stanford. NIST Information Technology Laboratory Pervasive Computing Initiative. IEEE Ninth International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, June 2000, NIST; USA.

[Rui 01] Y. Rui, A. Gupta and J. J. Cadiz. Viewing meeting captured by an omni-directional camera. Human Factors in Computing Systems CHI 2001, pp. 450-457, Seattle Washington, 2001.

[Samal 92] A. Samal and P. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. Pattern Recognition, 25(1):65-77, 1992.

[Stiefelhagen 99] R. Stiefelhagen, J Yang, A. Waibel, "Modeling Focus of Attention for Meeting Indexing," Proceedings of ACM Multimedia 1999.

[Stiefelhagen 00] R. Stiefelhagen, J. Yang, and A. Waibel. Simultaneous tracking of head poses in a panoramic view. In *International Conference on Pattern Recognition*, volume 3, pages 726–729, September 2000.

[Stiefelhagen 01] R. Stiefelhagen, J. Yang, and A. Waibel. Estimating focus of attention based on gaze and sound. In *Workshop on Perceptive User Interfaces (PUI'01)*, Orlando, Florida, November 2001.

[Stiefelhagen 02][6] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, July 2002.

[Stiefelhagen 02a] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems (CHI2002)*, Minneapolis, April 2002.

[Stiefelhagen 02b] R. Stiefelhagen. Tracking focus of attention in meetings.In *International Conference on Multimodal Interfaces*, pages 273–280, Pittsburgh, PA, October 2002. IEEE.

[Toyama 99] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In Proc. 7th Int. Conf. on Computer Vision, pp. 255–261, 1999.

[Turk 1991] M. Turk and A. Pentland. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1):72--86, 1991.

[Waibel 98] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting browser: Tracking and summarizing meetings. Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pages 281–286, Lansdowne, Virginia, February. 8-11 1998.

[Waibel 03] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, SMaRT: the smart meeting room task at ISL, Proceedings of ICASSP 2003.

[Wren 97] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, Pfinder: Real-time tracking of human body, IEEE Trans. Pattern Anal. Machine Intell., vol. 19, pp. 780–785, July 1997.

[Yang 99] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, A. Waibel, "Multimodal People ID for a Multimedia Meeting Browser," Proceedings of ACM Multimedia 99, pp.159-168.

[Yang 00] J. Yang, H. Yu, W. Kunz, "An Efficient LDA Algorithm for Face Recognition," International Conference on Automation, Robotics, and Computer Vision (ICARCV'2000), 2000.