# Modeling Focus of Attention for Meeting Indexing Based on Multiple Cues

Rainer Stiefelhagen, Jie Yang, *Member, IEEE*, and Alex Waibel, *Member, IEEE*

***Abstract*—A user's focus of attention plays an important role in human–computer interaction applications, such as a ubiquitous computing environment and intelligent space, where the user's goal and intent have to be continuously monitored. In this paper, we are interested in modeling people's focus of attention in a meeting situation. We propose to model participants' focus of attention from multiple cues. We have developed a system to estimate participants' focus of attention from gaze directions and sound sources. We employ an omnidirectional camera to simultaneously track participants' faces around a meeting table and use neural networks to estimate their head poses. In addition, we use microphones to detect who is speaking. The system predicts participants' focus of attention from acoustic and visual information separately. The system then combines the output of the audio- and video-based focus of attention predictors. We have evaluated the system using the data from three recorded meetings. The acoustic information has provided 8% relative error reduction on average compared to only using one modality. The focus of attention model can be used as an index for a multimedia meeting record. It can also be used for analyzing a meeting.**

***Index Terms*—Focus of attention, head pose estimation, human–computer interaction, meeting indexing, multimedia meeting record, multimodality.**

## I. Introduction

A person's focus of attention can be visually identified in certain circumstances. Participants in a meeting, for example, might look at the speaker while they are listening to the talk. When a user is editing a paper, he/she would direct his/her visual attention would direct toward a computer screen. Modeling and tracking a person's focus of attention is useful for many applications: Intelligent supportive computer applications could use information about a user's focus of attention to infer the user's mental status, his/her goals and cognitive load and adjust their own responses to the user accordingly. For multimodal human computer interaction, the user's focus of attention can be used to determine his/her message target. For example, in interactive intelligent rooms or houses [1], [2], focus of attention could be used to determine whether the user is to control the refrigerator, the TV set, or he/she is talking to another person in the room. In other words, the user's attention focus can be used to guide the environment's "focus" to the right application and to prevent responses generated from applications that have not been addressed. During social interaction, gaze serves for several functions which are not easily transmitted by auditory cues alone [3]. In computer mediated communication systems, such as virtual collaborative workspaces, detecting and conveying participants' gazes have several advantages: it can help the participants to determine who is talking or listening to whom, it can serve to establish joint attention during cooperative work, and it can facilitate turn taking among participants [4], [5]. In this paper, we are interested in modeling people's focus of attention in a meeting situation.

We are interested in meetings because they are one of the most common, important, and universally disliked events in our lives. Most people find it impossible to attend all relevant meetings or to retain all the salient points raised in meetings they do attend. Meeting records are intended to overcome these problems and extend human memories. Hand-recorded notes, however, have many drawbacks. Note-taking is time consuming, requires focus, and thus reduces one's attention to and participation in the ensuing discussions. For this reason, notes tend to be fragmentary and partially summarized, leaving one unsure exactly as to what was resolved, and why. At the Interactive Systems Lab of Carnegie Mellon University, we are developing a multimedia meeting recorder and browser to track and summarize discussions held in a specially equipped conference room [6]. The objective of the project is to provide a multimedia meeting record without using constraining devices such as headsets, helmets, suits, and buttons. The research issues include to identify: 1) who/what is the source of the message; 2) who or what is the target and object of the message (focus of attention); 3) what is the content of the message in the presence of jamming noise. The main components of the Meeting Browser are: a speech recognizer, a summarization module, a discourse component that attempts to identify the speech acts, a module for audio–visual identification of participants [7] and a module for tracking the participants' focus of attention.

In order to quickly retrieve information from such a multimedia meeting record, we can use various indexing methods. It is well known that visual communication cues, such as gesturing, looking at each other, or monitoring each others facial expressions, play an important role during face-to-face communication [3], [8]. Therefore, to fully understand an ongoing conversation, it is necessary to capture and analyze these visual cues in addition to spoken content. Once such visual cues can be tracked, they can be used to index and retrieve recorded meetings. Queries, such as "show me all parts of the meeting, where John was telling Mary something about

Manuscript receivedApril 12, 2001; revised October 29, 2001. This work was supported in part by the Defense Advanced Research Projects Agency under Contract DAAD17-99-C-0061, and by the National Science Foundation under Grant IIS-9980013.

R. Stiefelhagen is with the Institute for Logic, Complexity and Deduction Systems, University of Karlsruhe, Germany (e-mail: stiefel@ira.uka.de).

J. Yang and A. Waibel are with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Publisher Item Identifier S 1045-9227(02)04429-6.

the multimedia project" become possible. In addition, during playback of parts of a meeting, we could indicate at whom the speaker was looking.

In this research, we address the problem of tracking the visual focus of attention of participants in a meeting; i.e., tracking who is looking at whom during a meeting. Such information can be used to control interaction with a smart meeting room or to index and analyze multimedia meeting records.

In our system, an omnidirectional camera is used to capture the scene around a meeting table. Participants are detected and tracked in the panoramic image using a real-time face tracker. Furthermore, neural networks are used to compute head pose of each person simultaneously from the panoramic image. We then use a Bayesian approach to estimate a person's focus of attention from the computed head pose. We model the *a posteriori* probability that a person is looking at a certain target, given the observed head pose. Using this approach, we have achieved 74% accuracy in detecting the participants' focus of attention on three recorded meetings.

In addition to visual information, we have investigated whether a person's focus of attention can be predicted from other information. We have discovered that focus of attention is also correlated to sound sources in a meeting. We can estimate a person's focus of attention based on the information of who is talking at or was talking before a given moment. This is based on the idea that visual attention is *influenced* by external events such as noises, movements, or other person's speech. We have estimated probability distributions of where participants are looking during certain "speaking constellations." We can then use these distributions to predict the focus of attention using the sound information only. We have achieved 54% accuracy in predicting the participants' focus of attention on three recorded meetings. The accuracy of sound-based prediction can be significantly improved by also taking a history of speaker constellations into account. We have trained neural networks to predict focus of attention based on who was speaking during a short period of time. Using this approach, sound-based prediction could be increased to 63%.

Finally, the head pose based and the sound-based estimate are combined to obtain a multimodal estimation of the participants' focus of attention. By using both head pose and sound, we have achieved 76% accuracy in detecting the participants' focus of attention on the recorded meetings.

The novelty of this research lies in estimating focus of attention from multiple cues. To our knowledge, this is the first time that predicting a person's focus of attention based on who is talking has been reported.

The remainder of this paper is organized as follows: In Section II, we introduce the idea of modeling a person's focus of attention by observing a person's gaze as well as monitoring relevant stimuli in the scene. In Section III, we introduce the approach to estimate head poses of participants using neural networks. In Section IV, we discuss methods to model the probability distributions of whom a person is looking at based on his/her head pose. In Section V, we present two different approaches to predict a person's focus of attention by monitoring who is speaking. We provide details how focus of attention can be predicted by knowledge about who is currently speaking, and how prediction accuracy can be improved by taking the history of speakers into account. We also address combination of audio- and head pose-based focus predictions, and illustrate experimental results. In Section VI, we present an application of our model to the meeting browser. Information about the participants' focus of attention is tracked and is integrated as a component in the meeting browser. The meeting browser can then be used to index meeting transcriptions and summaries with visual cues. In Section VII we summarize the paper.

## II. Modeling Focus of Attention

The idea of this research is to track at whom or what the participants are paying attention to during the course of a meeting. Gaze is a good indicator of a person's attention during social interaction. When humans pay attention to someone, they usually orient themselves toward the person of interest so as to have it in the center of their visual field and also to signal that they are paying attention to the other person [9], [10].

Although the eyes are the primary source to detect a person's gaze during social interaction, gaze is not limited to information from the eyes. The perception of someone else's direction of attention also depends on the direction of their head, body posture and other gestures, such as pointing gestures. All theses cues are likely to be processed automatically by observers and all make contributions to the perceptions of another person's attention [11]. In fact it has been shown that head orientation strongly influences the perception of gaze, even when the eyes are visible [12].

In our approach we aim to estimate a person's focus of attention, based on his head orientation. To map a person's head orientation onto the focussed object in the scene, a model of the scene and the interesting objects in it are needed. In the case of a meeting scenario, clearly the participants around the table are likely targets of interest. Therefore, our approach to tracking at whom a participant is looking is the following.

1) Detect all participants in the scene.
2) Estimate each participant's head orientation.
3) Map each estimated head orientation to its likely targets using a probabilistic framework.

Compared to directly classifying a person's focus of attention target—based on images of the person's face for example—our approach has the advantage, that different numbers and positions of participants in the meeting can be handled. If the problem was treated as a multiclass classification problem, and a classifier such as a neural network was trained to directly learn the focus of attention target from the facial images of a user, then the number of possible focus targets would have to be known in advance. Furthermore, with such an approach it would be difficult to handle situations were participants sit at different locations than during collection of the training data.

Objects which draw a person's attention can be external stimuli such as pictures, sounds, etc., or internal stimuli such as thoughts and attempts to retrieve information from memory [13]. Clearly, visual attention is influenced by external stimuli, such as noises, movements, or speech of other persons. There is evidence, for example, that two or more people will orient themselves toward each other as soon as they begin to interact.

And it has been argued that there is an orientation reflex to the source of a sound, causing interactors to line up the visual and auditory channel; i.e., to look at the face which is the source of the sound [14] (cf. [15]).

Another approach to estimate at whom or what a person is paying attention to, could therefore be, to monitor external events in the meeting environment, such as sounds, utterances, gestures, persons entering the room etc., and try to make a prediction of the participants' focus of attention based on these external events.

Following this idea, we have also tried to predict at whom a person is looking, based on who is speaking at the moment and based on the temporal sequence of speakers.

## III. Estimating Head Pose Using Neural Nets

In this section we present an approach to estimate head poses of participants from panoramic images using neural networks.

The main advantage of using neural networks to estimate head pose as compared to using a model-based approach is its robustness: With model-based approaches to head pose estimation [16]–[18], head pose is computed by finding correspondences between facial landmarks points (such as eyes, nostrils, lip corners) in the image and their respective locations in a head model. Therefore these approaches rely on tracking a minimum number of facial landmark points in the image correctly, which is a difficult task and is likely to fail. On the other hand, the neural-network-based approach does not require tracking detailed facial features. Instead, the whole facial region is used for estimating the user's head pose.

In our approach, we are using neural networks to estimate pan and tilt of a person's head, given automatically extracted and preprocessed facial images as input to the neural net. Schiele and Waibel [19] demonstrated a neural-network-based head pose tracking system. The system estimated only head rotation in pan direction for one person. Rae *et al.* [20] describe a user dependent neural-network-based system to estimate the pan and tilt of a person. In their approach, color segmentation, ellipse fitting, and Gabor-filtering on a segmented face are used for preprocessing. They reported an average accuracy of 9° for pan and 7° for tilt for one user with a user dependent system.

In our previous work on estimating participant's focus in meetings [21], we have used separate cameras to zoom in on each of the participants in order to obtain the input images for pose estimation. Using these high-resolution images, we achieved an accuracy of 7° for pan and 8° for tilt on a user independent test set in recent experiments.

In this research, we use an omnidirectional camera to capture the participants, and track faces in the panoramic image. Compared to using multiple cameras to capture all participants this has the advantage that only one video-stream has to be captured, which eliminates the need for camera calibration, synchronization and camera control such as zooming on different participants. While the facial images extracted from the panoramic view are of considerably lower resolution than images taken with close up views, we could still obtain good accuracy using our approach.

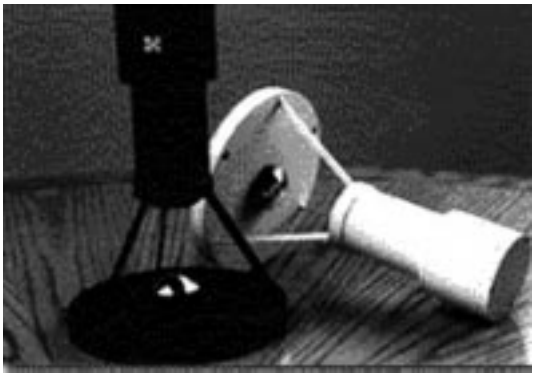


Fig. 1. The panoramic camera used to capture the scene.

Fig. 2. Meeting scene as captured with the panoramic camera.

### A. Capturing the Scene

In our system, an omnidirectional camera put on top of the conference table is used to capture the scene. Fig. 1[1] shows a picture of the panoramic camera system. The camera is located in the top cylinder and is focusing on a parabolic mirror on the bottom plate. Through this mirror almost a whole hemisphere of the surrounding scene is visible. Fig. 2 shows the view of a meeting scene as it is captured with this camera. As the topology of the mirror and the optical system are known, it is possible to compute panoramic views of the scene as well as perspective views at different angles of the panoramic view [22]. Fig. 3 shows the rectified panoramic image (with faces marked) of the camera view depicted in Fig. 2.

### B. Using Color and Motion for Face Detection

To detect and track faces in the panoramic view, a statistical skin color model in the normalized color space is used. The color distribution is initialized so as to find a variety of face colors and is gradually adapted to the faces actually found. The input image is searched for pixels with skin colors. Connected regions of skin-colored pixels in the camera image are considered as possible faces. Since humans rarely sit perfectly still for a long time, motion detection is used to reject outliers that might be

[1]Image courtesy of CycloVision Technologies, Inc.

Fig. 3. Panoramic view of the scene around the conference table. Faces are automatically detected and tracked (marked with white rectangles).

caused due to noise in the image or skin colored objects within the image. Only regions with a response from the color-classifier and some motion during a period of time are considered as faces. In addition, some geometric constraints are applied to distinguish (skin-colored) hands from faces. For more detail, the interested reader is referred to [23].

### C. Data Collection

We collected training data from 14 users. During data collection, the user was automatically tracked in the panoramic view, and a perspective view as depicted in Fig. 4 was generated. To determine true head pose, the users had to wear a head band with a sensor of a Polhemus pose tracker attached to it. Using the pose tracker, the head pose with respect to a magnetic transmitter could be collected in real time. The user was asked to randomly look around in the room and the perspective images of the user were recorded together with the pose sensor readings.

### D. Preprocessing of Images

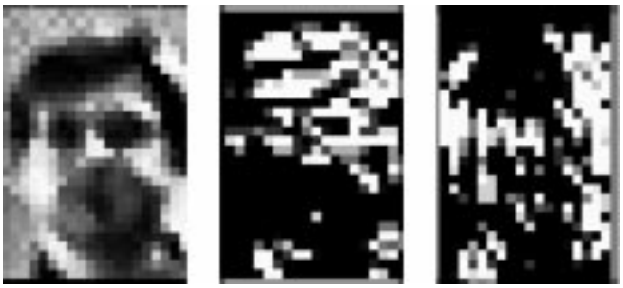
We have investigated two different image preprocessing methods as input to the neural networks for head pose estimation: 1) using normalized gray-scale images of the user's face as input and 2) applying edge detection to the facial images before feeding them into the networks. To find and extract faces in the collected images, we use the color-based face detector described in Section III-B.

In the first preprocessing approach, histogram normalization is applied to the gray-scale face images. No additional feature extraction is performed. The normalized gray-scale images are down-sampled to a fixed size of $20 \times 30$ pixels and are then used as input to the networks. Histogram normalization defines a mapping of gray levels $p$ into gray levels $q$ such that the distribution of $q$ matches a certain target distribution (e.g., a uniform distribution). This mapping stretches contrast and usually improves the detectability of many image features [24]. Histogram normalization is also helpful to get some illumination invariance.

In the second approach, a horizontal and a vertical edge operator plus thresholding is applied to the facial gray-scale images. The resulting edge images are down-sampled to $20 \times 30$ pixels and are both used as input to the neural networks. Fig. 5 shows the preprocessed images of a user's faces. The normalized gray-scale image and the horizontal and vertical edge images of a user's face are depicted.

(a) (b)

Fig. 4. Training samples. (a) The perspective images are generated from the panoramic view. (b) Head pose labels are collected with a magnetic field pose tracker.



(a) (b) (c)

Fig. 5. Preprocessed images. (a) Normalized gray-scale. (b) Horizontal edge. (c) Vertical edge image.

### E. Neural-Network Architecture

We have trained separate neural networks to estimate head pose in pan and tilt directions. For each network, a multilayer perceptron architecture with one output unit (for pan or tilt) and one hidden layer with 20 to 60 hidden units. The input retina varied between $20 \times 30$ units and $3 \times 20 \times 30$ units depending on the different types of input images. Output activations for pan and tilt are normalized to vary between zero and one. Neural networks are trained using standard backpropagation.

### F. Experimental Results

We divided the data set of 12 users (of the 14 users in the whole data set) into a training set consisting of 6080 images, a cross-evaluation set of size 760 images and a test set with a size of 760 images. The images of the remaining two users were kept as a user independent test set. As input to the neural networks, three different approaches were evaluated:

1) using histogram normalized gray-scale images as input;
2) using horizontal and vertical edge images as input;
3) using both normalized gray-scale plus the edge images as input.

The neural networks were trained on the training data set and the cross-evaluation set was used to determine when to stop training. The performance of the networks was then evaluated on the test set containing images of the 12 persons that were also in the training set (multiuser case). On the multiuser test set, we obtained the best performance using both, normalized gray-scale images and edge images as input. A mean error of 7.8° for pan and 5.4° for tilt was obtained with the best networks. Using only the gray-scale images as input, the results decreased to a mean error of 9.4° for pan and 6.9° for tilt. With edge images as input, a mean error of only 10.8° for pan and 7.1° for tilt could be achieved.

*1) User Independent Results:* To determine how well the neural nets can generalize to new users, we have also evaluated the networks on the two new users whose data have not been in the training set. On the two new users the best result for pan estimation, which was 9.9° mean error, was obtained using normalized gray-scale images plus edge images as input. The best result for tilt-estimation measured was 9.1° mean error and was obtained using only normalized gray-scale images as input. Table I summarizes the results on the multiuser and the user-independent test sets.

*2) Adding Artificial Training Data:* In order to obtain additional training data, we have artificially mirrored all of the images in the training set, as well as the labels for head pan. As a result, the available amount of training data could be doubled without having the effort of additional data collection. Having more training data should especially be helpful in order to get better generalization on images from new, unseen users. Indeed, after training with the additional data, we obtained an average error of 9.5° for pan and 9.8° for tilt on the two new users using the gray-scale and the edge images as input. On the multiuser test set the mean pose estimation error significantly decreased to 3.1° for pan and 2.5° for tilt. Table II shows the results on the multiuser test set as well as the new user test set for the different preprocessing approaches.

*3) Discussion:* From experiment results, we have observed that using only edge images as input leads to poorer head pose estimations in both pan and tilt directions, as compared to using only gray-scale images as input. Furthermore, using both gray-scale and edge images as input leads to the best results in both pan and tilt directions in most cases. However, on the test set of new users, using only gray-scale images as input leads to slightly better results for the estimation of head tilt (up/down) as compared to using both gray-scale and edge images. Moreover, adding artificial training data improves estimation results both on the multiuser test set and on the new users.

TABLE I
AVERAGE ERROR IN DEGREES (PAN/TILT) ON A MULTIUSER AND A USER-INDEPENDENT TEST SET

| Net Input | Multi-user Test Set | New Users |
|---|---|---|
| Gray-scale | 9.4 / 6.9 | 11.3 / **9.1** |
| Edges | 10.8 / 7.1 | 13.3 / 10.8 |
| Edges + Gray-scale | **7.8 / 5.4** | **9.9** / 10.3 |

TABLE II
RESULTS USING ADDITIONAL ARTIFICIAL TRAINING DATA. RESULTS ON THE MULTIUSER TEST SET AND ON THE TWO NEW USERS ARE SHOWN FOR THE DIFFERENT PREPROCESSING APPROACHES. THE MEAN ERROR IN DEGREES OF PAN/TILT IS SHOWN

| Net Input | Multi-user Test Set | New Users |
|---|---|---|
| Gray-scale | 5.5 / 4.1 | 10.4 / **9.3** |
| Edges | 5.6 / 3.5 | 12.2 / 10.3 |
| Edges + Gray-scale | **3.1 / 2.5** | **9.5** / 9.8 |

## IV. MODELING FOCUS BASED ON HEAD ROTATION

In our approach, we first estimate a persons head orientation—as described in Section III—and then estimate at whom a person was looking at, based on his estimated head rotation.

Using *a priori* knowledge about the size of the table and assuming that participants are located close to the table, it is possible to compute the approximate two-dimensional (2-D) location of each participant from the positions of the faces found in the panoramic image. A first solution to find out at whom a person $S$ is looking could be, to use the measured head pose of $S$ and look which target person $T_i$ sits nearest the position to which $S$ is looking. Gaze is, however, not only determined by head pose, but also by the direction of eye gaze. People do not always completely turn their heads toward the person at which they are looking. Instead, they also use their eye gaze direction.

We have therefore developed a Bayesian approach to estimate at which target a person is looking, based on his observed head rotation. More precisely, we wish to find $P(\text{Focus}_S = T|x_S)$, the probability that a person $S$ is looking toward a certain target person $T$, given the person's observed horizontal head rotation $x_S$, which is the output of the neural network for head pan estimation. Using Bayes formula, this can of be decomposed to

$$P(\text{Foc.}_S = T|x_S) = \frac{p(x_S|\text{Foc.}_S = T)P(\text{Foc.}_S = T)}{p(x_S)} \quad (1)$$

where $x_s$ denotes the head pan of person $S$ in degrees and $T$ is one of the other persons around the table.

Using this framework, given a pan observation $x_s$ for a person $S$—as estimated by the neural network for head pan estimation—it is then possible to compute the posterior probabilities $P(\text{Focus}_S = T_i|x_S)$ for all targets $T_i$ and choose the one with highest posterior probability as the focus of attention target in the current frame.

In order to compute $P(\text{Focus}_S = T|x_S)$, it is however necessary to estimate the class-conditional probability density function $p(x_S|\text{Focus}_S = T)$, the class prior $P(\text{Focus}_S = T)$ 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 rotations of a person over time.

One possibility to find the class-conditional pdf and the prior would be to adjust them on a training set of similar meetings. This, however, would require training data for any possible number of participants at the table and for any possible combination of the participants' locations around the table. Furthermore, adapting on different meetings and different persons would probably not model a certain person's head turning

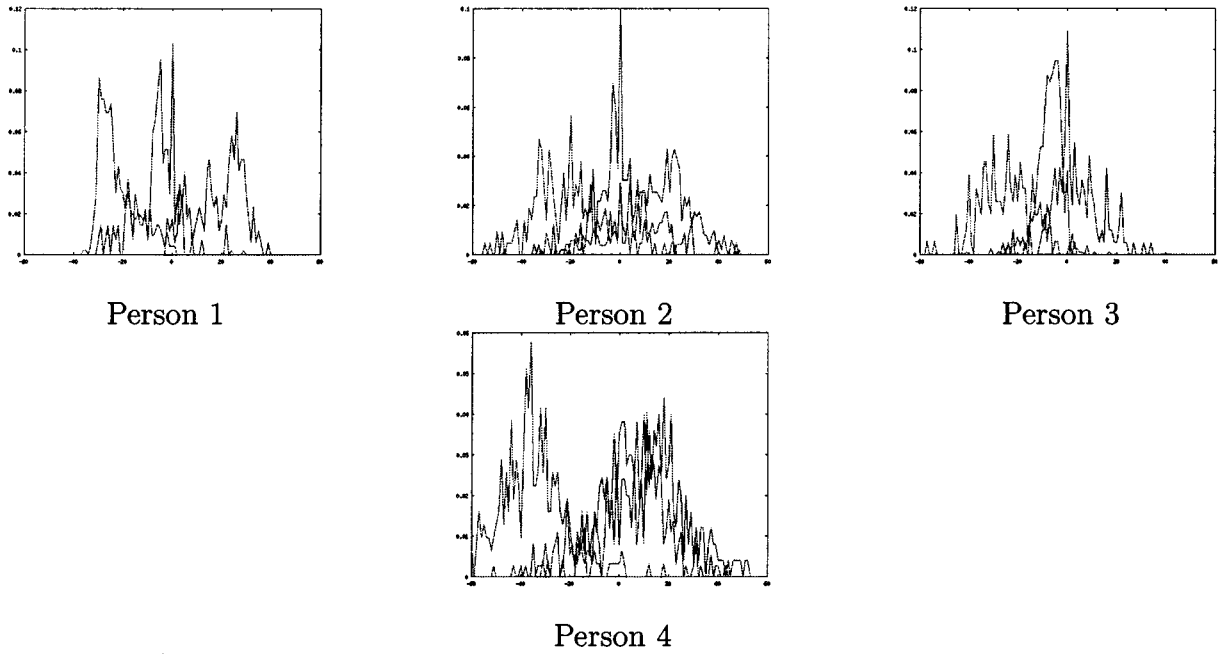



Fig. 6. Class-conditional head pan distributions of four persons when looking to the person to their left, to their right or to the person sitting opposite.

style very well, nor would the priors necessarily be the same in different meetings. In our meeting recordings we observed that some people turned their head more than others and some people made stronger use of their eye-gaze and turned their head less when looking at other people. Fig. 6 shows the head pan distributions of four participants in one of our recorded meetings. The head rotation of the user was estimated with the neural nets. It can be seen, for example, for Person 1, the three class-conditionals are well separated, whereas for Person 3 or Person 4, the peaks of some distributions are much closer to each other, and a higher overlap of the distributions can be observed.

We have therefore developed an unsupervised learning approach to find the head pan distributions of each participant when looking at the others.

### *A. Unsupervised Adaptation of Model Parameters*

In our approach, we assume that the class-conditional head pan distributions, such as depicted in Fig. 6, can be modeled as Gaussian distributions. Then, the distribution of all head pan observations from a person $p(x)$ will result in a mixture of Gaussians

$$p(x) \approx \sum_{j=1}^{M} p(x|j)P(j) \tag{2}$$

where the individual component densities $p(x|j)$ are given by Gaussian distributions $N_j\left(\mu_j, \sigma_j^2\right)$.

In our approach, the number of Gaussians $M$ is set to the number of other participants at the table, because we assume that these are the most likely targets that the person has looked at during the meeting, and because we want to find the individual Gaussian components that correspond to looking at these target persons.

The model parameters of the mixture model can then be adapted so as to maximize the likelihood of the pan observations given the mixture model. This is done using the expectation-maximization algorithm by iteratively updating the parameter values using the following update equations [25]:

$$\mu_j^{\text{new}} = \frac{\sum_n P^{\text{old}}(j|x^n)x^n}{\sum_n P^{\text{old}}(j|x^n)} \tag{3}$$

$$\left(\sigma_j^{\text{new}}\right)^2 = \frac{1}{d}\frac{\sum_n P^{\text{old}}(j|x^n)\left\|x^n - \mu_j^{\text{new}}\right\|^2}{\sum_n P^{\text{old}}(j|x^n)} \tag{4}$$

$$P(j)^{\text{new}} = \frac{1}{N}\sum_n P^{\text{old}}(j|x^n). \tag{5}$$

To initialize the means $\mu_j$ of the mixture model, kmeans clustering was performed on the pan observations.

After adapting the mixture model to the data, the individual Gaussian components can be used as an approximation of the class-conditionals $p(x|\text{Focus} = T)$, and the priors of the mixture model $P(j)$ can be used to approximate the focus priors $P(\text{Focus} = T)$ of our model, described in (1). Furthermore, the individual Gaussian components can be assigned to corresponding target persons based on their relative position around the table.

Fig. 7 shows an example of the adaptation on pan observations from one user. In Fig. 7(a) the distribution of all head pan observations of the user is depicted together with the Gaussian mixture that was adapted as described above. Fig. 7(b) depicts the real class-conditional head pan distributions of that person, together with the Gaussian components taken from the Gaussian mixture model depicted in Fig. 7(a). As can be seen, the Gaussian components provide a good approximation of the real class-conditional distributions of the person. Note that the real class-conditional distributions are just

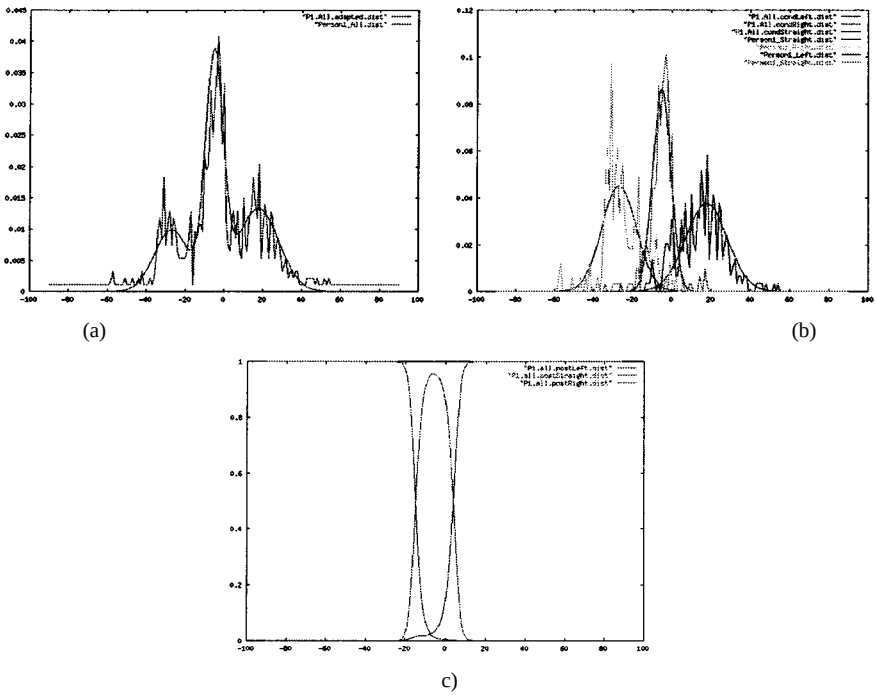


Fig. 7. (a) The distribution $p(x)$ of all head pan observations for a person. Also the adapted mixture of three Gaussians is plotted. (b) True and estimated class-conditional distributions of head pan $x$ for the same person, when looking to three different targets. The adapted Gaussians, are taken from the adapted Gaussian mixture model depicted in (a). (c) The posterior probability distributions $P(\mathrm{Focus}|x)$ for resulting from the found mixture of Gaussians.

depicted for comparison and are of course not necessary for the adaptation of the Gaussian components. Fig. 7(c) depicts the posterior probability distribution resulting from the adapted class-conditionals and class priors.

### B. Meetings for Evaluation

To evaluate our system, several meetings were 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 each participant had one microphone in front of him to capture his speech. Using this setup, we recorded audio streams for each of the participants plus the panoramic view of the scene simultaneously to harddisk. The three recorded meetings varied from 5 min and 30 s to 8 min and 30 s and contained between 870 to 1280 video frames.

In each frame of the recorded meetings, we labeled for each of the participants at whom he was looking. These lables 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 was not 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 assigned to each video frame. These labels could be assigned by listening to the audio streams.

### C. Experimental Results

We have evaluated this approach on three evaluation meetings. In each meeting, the faces of the participants were automatically tracked, and head pan was estimated using the neural-network-based approach. For each of the four participants in each meeting, the class-conditional head pan distribution $p(x|\mathrm{Focus})$, the class-priors $P(\mathrm{Focus})$ and the observation distributions $p(x)$ were automatically adapted to compute the posterior probabilities $P(\mathrm{Focus} = T_i|x)$ for each person. In each frame the target with the highest posterior probability was chosen as the focus of attention target of the person. For the 12 users in the three meetings, the correct focus target could be detect on average in 73.9% of the frames. Table III shows the average results on the three meetings.

TABLE III
PERCENTAGE OF CORRECT ASSIGNED FOCUS TARGETS BASED ON COMPUTING $P(\mathrm{Focus}|\mathrm{head\ pan})$

| | $P(\mathrm{Focus}|\mathrm{Gaze})$ |
|---|---|
| Meeting A (4 participants) | 68.8 % |
| Meeting B (4 participants) | 73.4 % |
| Meeting C (4 participants) | 79.5 % |
| Average | 73.9 % |

## V. PREDICTING FOCUS FROM SOUND

As we have argued before, visual attention is influenced by external stimuli. We have, therefore, investigated whether it is possible to predict a person's focus of attention based on audio information.

In our first experiment to predict focus from sound we analyzed at whom the four participants in the recorded meetings were looking during certain "speaking" conditions. Here, "speaking" was treated as a binary variable; i.e., each of the four participants, was either labeled as "speaking" or "not speaking" in each video frame. Now, using this binary "speaking" variable and having four participants, there exist $2^4$ possible "speaking"

conditions in each frame, ranging from none of the participants is speaking to all of the participants are speaking.

Table IV summarizes at whom participants in our three meetings were looking, based on who was speaking. In the first row, the speaking condition is represented as the binary vector $\vec{A}$, with entry $a_s$ indicating whether the subject $S$ himself ("self") was speaking, the second entry $a_L$ indicating whether the person to the subject's left was speaking, the third entry $a_C$ indicating whether the person opposite ("center") to $S$ was speaking, and entry $a_R$ indicating whether the person to its right was speaking. For each person and each case we counted how often the subjects looked to the right, looked straight or looked to the person to their right. For example, when only the person to the subject's left was speaking (entry "0 1 0 0"), in 59% of the cases the subject was looking to the left person (the speaker), in 28% of the cases he was looking straight to the opposite person and in 11% of the cases he was looking to the person to his right.

Overall it can be seen that if there was only one speaker, subjects most often looked to that speaker (percentages are indicated in bold font in Table IV for that person). This also holds for cases were there was only one *additional* speaker when the subject itself was speaking. The last row of Table IV indicates in which direction subjects looked on average, regardless of speaking conditions. It can be seen that there is a bias toward looking straight; i.e., regardless who was speaking, in 44% of the cases the person opposite has been looked at, whereas the persons sitting to the side have been looked at in only 26% of the cases.

The entries of Table IV can be directly interpreted as the the probability that a subject $S$ was looking to a certain person $T$, based on the binary audio-observation vector $\vec{A}$:

$$P(\text{Focus}|\text{Sound}) = P(\text{Focus}_S = T_j|\vec{A})$$

where $T_j$, with $j \in \{\text{"left," "straight," "right"}\}$ denote the possible persons to look at, and where

$$\vec{A} = (a_{\text{self}}, a_{\text{left}}, a_{\text{center}}, a_{\text{right}})$$

denotes the audio-observation vector with binary components $a_i$, indicating whether the subject it *self*, the person to his *right*, *left*, or the person opposite (*center*) to the subject was speaking.

The probability $P(\text{Focus}|\text{Sound})$ can be used directly to predict at whom a participant is looking in a frame, based on who was speaking during that video frame. In each frame, for each subject $S$ the person $T_i$ was chosen as the focus of person $S$, which maximized $P(\text{Focus}_S = T_i|\vec{A})$.

By using only the speaker labels to make a sound-based focus prediction, the correct focus of each participants could be predicted with an average accuracy of 54% on three evaluation meetings.

### A. Combining Gaze and Sound to Predict Focus

In Section IV it was shown, how we can determine the probability $P(\text{Focus}|\text{Sound})$; i.e., the probability that a person is looking toward a certain other person, based on the information, of whom is currently speaking. By choosing in each frame the target person $T_i$ which maximized $P(\text{Focus}_S = T_i|\vec{A})$ as the focus of person $S$, a focus prediction accuracy of 54% could be achieved.

TABLE IV
TABLE SUMMARIZES, HOW OFTEN PEOPLE LOOKED TO PARTICIPANTS IN CERTAIN DIRECTIONS, DURING THE DIFFERENT SPEAKING CONDITIONS. THE SPEAKING CONDITION IS REPRESENTED IN THE FIRST ROW (SEE TEXT)

| $\vec{A} = (a_S a_L a_C a_R)$ | Left | Straight | Right |
|---|---|---|---|
| 0 0 0 0 | 0.26 | 0.49 | 0.23 |
| 0 0 0 **1** | 0.11 | 0.27 | **0.60** |
| 0 0 **1** 0 | 0.12 | **0.74** | 0.11 |
| 0 0 1 1 | 0.07 | 0.49 | 0.40 |
| 0 **1** 0 0 | **0.59** | 0.28 | 0.11 |
| 0 1 0 1 | 0.35 | 0.24 | 0.37 |
| 0 1 1 0 | 0.33 | 0.60 | 0.05 |
| 0 1 1 1 | 0.21 | 0.41 | 0.38 |
| 1 0 0 0 | 0.24 | 0.48 | 0.25 |
| 1 0 0 **1** | 0.09 | 0.34 | **0.53** |
| 1 0 **1** 0 | 0.18 | **0.61** | 0.18 |
| 1 0 1 1 | 0.08 | 0.59 | 0.30 |
| 1 **1** 0 0 | **0.60** | 0.24 | 0.11 |
| 1 1 0 1 | 0.29 | 0.44 | 0.26 |
| 1 1 1 0 | 0.35 | 0.56 | 0.08 |
| **1** 1 1 1 | 0.50 | 0.50 | 0.00 |
| all cases | 0.26 | 0.44 | 0.26 |

In Section IV we showed, how to compute $P(\text{Focus}_S = T_i|x_S)$, the posterior probability, that a person $S$ is looking toward person $T_i$, based on his estimated head rotation $x_S$. There, by again choosing in each frame the target person $T_i$ which maximized $P(\text{Focus}_S = T_i|x_S)$ as the focus of person $S$, we achieved correct focus prediction in 73.9% of the frames.

These two independent predictions of a person's focus—$P(\text{Focus}|\text{Sound})$ and $P(\text{Focus}|\text{gaze})$—can be combined in a straightforward way to obtain a prediction of a person's focus which is based on both, the observation, who is speaking, and based on the estimation of the person's head rotation. The combined result can be obtained by computing the weighted sum of both predictions

$$p(\text{Focus}) = (1-\alpha)P(\text{Focus}|\text{Gaze}) + \alpha P(\text{Focus}|\text{Sound}).$$

We have evaluated the combined prediction results on our meetings for different values of $\alpha$, ranging from 0.0 to 1.0. On the three meetings, the optimal values of $\alpha$ ranged from 0.3 to 0.6. By setting $\alpha$ to 0.6, good results could be achieved on all meetings. Using this multimodal prediction, an accuracy of 74.8% was achieved on the three meetings (see Table V). Compared to the prediction accuracy of 73.9% using gaze only, this corresponds to a relative error reduction of 3.4%.

While the presented combination of head pose- and sound-based prediction is done heuristically by choosing a weighting parameter, we expect that by using more advanced and adaptive fusion methods, better combination results will be obtained. Appropriate fusion methods to be investigated could be to train neural networks for fusion of the two modalities, to determine the weighting parameters using error information of the two models or to investigate other feature dependent combinations methods [26]–[29].

TABLE V
FOCUS-PREDICTION USING GAZE ONLY SOUND ONLY AND PREDICTION USING BOTH GAZE AND SOUND

| | Gaze only | Sound only | Combined |
|---|---|---|---|
| Meeting A | 68.8 | 57.7 % | 69.7 % |
| Meeting B | 73.4 | 57.6 % | 75.3 % |
| Meeting C | 79.5 | 46.9 % | 79.5 % |
| Average | 73.9 | 54.1 % | 74.8 % |

### B. Using the Sound History to Predict Focus

In Section IV, information about who is speaking is used to predict $p(\text{Focus}|\text{Sound})$; i.e., how likely it is for a person to look at one of the others based on who is speaking. The prediction, however, is only based the audio-observation $\vec{A}^t$ corresponding to the current video frame at time $t$.

This has several drawbacks: By using only audio-information from one frame, no temporal information is taken into account for the prediction. Temporal information, however might be very important.

A straightforward extension is, to use a history of audio-events $A^t, A^{t-1}, \ldots, A^{t-N}$ to predict the probability that a person $S$ is looking toward one of the others; i.e., to estimate $P(\text{Focus}|A^t, A^{t-1}, \ldots, A^{t-N})$.

In this work, we have chosen to use a neural network to predict $P(\text{Focus}|A^t, A^{t-1}, \ldots, A^{t-N})$. We have trained one neural network to estimate the probabilities that a person is looking to the person to his right, to his left, and to the person opposite to himself, based on a history of ten audio-observations. As audio-observations, we have again chosen the binary audio-observation vector $\vec{A} = (a_{\mathbf{S}}, a_{\mathbf{L}}, a_{\mathbf{C}}, a_{\mathbf{R}})$, described in Section IV.

To evaluate the performance of the audio-history-based prediction, we have trained networks round-robin; i.e., the neural nets were trained on data from two out of the three meetings and were evaluated on the remaining third meeting.

Using the audio-history based prediction of focus, an average prediction accuracy of 63.5% on the three meetings could be achieved. Compared to the 54.1% achieved with the prediction based on a single audio-frame, this is a relative error reduction of 20%. The audio-based prediction results are summarized in Table VI.

Again we can compute a combined, gaze- and sound-based prediction, by computing the weighted sum of $P(\text{Focus}|\text{Gaze})$ and $P(\text{Focus}|\text{Sound})$

$$P(\text{Foc.}) = (1-\alpha)P(\text{Foc.}|\text{Gaze}) + \alpha P(\text{Foc.}|A^t, \ldots, A^{t-N}).$$

By setting $\alpha$ to 0.5, we achieved an average accuracy of 75.9% on the three meetings.

Table VII summarizes the results we obtained by using sound-only based focus prediction, gaze-only based focus estimation and combined estimation.

## VI. INTEGRATING FOCUS OF ATTENTION MODELING INTO A MEETING BROWSER

We have integrated a component to track people's focus of attention into the "meeting browser"—a system to track and summarize meetings [6], [30], [31]. The meeting browser is a system designed to automatically review and search recordings of meetings. The browser is implemented in Java and includes video capture of individuals in the meeting, as pictured in Fig. 8. The main components of the meeting browser are: 1) a speech recognizer; 2) a summarization module; 3) a discourse component that attempts to identify speech acts; 4) a module for audio–visual identification of participants [7]; and 5) a module for tracking the participants' focus of attention.

The meeting browser is part of a multimodal intelligent meeting room. The goal of this project is not only to provide a tool to record and transcribe spoken content of the meetings, but to also detect who participated in the meeting and who was talking when and to whom. For the data acquisition in the meeting room, we used several microphones, a panoramic camera as described in Section III-A and several cameras around the table to capture close-up views of the participants.

With the components described in this paper, it is possible to detect the number and positions of participants in a meeting as well as to track which person at the table each of the participants look at. Together with the components for person and speaker identification, which are described in detail in [7], it is furthermore possible to determine who these participants are and who the speaker of a certain utterance was (speaker ID). Given all these cues for indexing of the meetings, it is then possible to formulate queries such as: "show me all parts, where John was telling Mary something about the multimedia project." In addition, during playback of parts of the meeting, we could indicate at whom the speaker was looking during his speech. For example, Fig. 9 shows an example where the gaze tracking component detected and indicated that the person was looking at the participant to her left and at the one to her right, respectively. Finally, we could even use this data to analyze meetings in many ways. One such usage could be to calculate how much of the time someone was speaking or how much of the time person X was addressing person Y.

TABLE VI
FOCUS-PREDICTION USING ONE FRAME AND TEN FRAMES OF SPEAKER INFORMATION. NEURAL NETWORKS WERE TRAINED TO PREDICT $P(\text{Focus}|A^t, A^{t-1}, \ldots, A^{t-9})$

| | $P(\text{Focus}\|A^t)$ | $P(\text{Focus}\|A^t, \ldots, A^{t-9})$ |
|---|---|---|
| Meeting A | 57.7 % | 63.0 % |
| Meeting B | 57.6 % | 67.2 % |
| Meeting C | 46.9 % | 60.2 % |
| Average | 54.1 % | 63.5 % |

TABLE VII
FOCUS-PREDICTION USING GAZE ONLY SOUND ONLY AND PREDICTION USING BOTH. SOUND-BASED FOCUS PREDICTION IS DONE WITH A NEURAL NETWORK, USING TEN FRAMES OF SPEAKER INFORMATION AS INPUT

| | Gaze only | Sound only | Combined |
|---|---|---|---|
| Meeting A | 68.8 | 63.0 % | 71.4 % |
| Meeting B | 73.4 | 67.2 % | 77.1 % |
| Meeting C | 79.5 | 60.2 % | 80.5 % |
| Average | 73.9 | 63.5 % | 75.9 % |

Fig. 8. Meeting browser with video capture.

(a) (b)

Fig. 9. Examples in which the attention model indicates that the person is looking to the participant to the left and right, respectively.

## VII. Conclusion

We have presented a system to estimate visual focus of attention of participants in a meeting from multiple cues. The participants are simultaneously tracked in a panoramic view and their head poses are estimated using neural networks. For each participant, probability distributions of looking toward other participants are estimated from head poses using an unsupervised learning approach. These distributions are then used to predict focus of attention given a head pose. The accuracy of such predication is 74% accurate in detecting the participants' focus of attention on our test data.

Furthermore, we have demonstrated how focus of attention can be predicted based on knowledge of who is currently speaking, and how this audio-based prediction can be improved by taking the history of utterances into account. On the recorded meetings, participants' focus of attention has been predicted correctly in 63% of the frames by using audio information only.

Finally, we have shown how the audio- and the video-based predictions can be fused to get a more accurate and robust estimation of participants' focus of attention. By using both head pose and sound, focus of attention could be detected in 76% of the frames in recorded meetings.

Other application areas of tracking focus of attention include: multimodal human computer interfaces, computer supported collaborative work, and interactive intelligent environments.

## Acknowledgment

The authors would like to thank the many colleagues in Interactive Systems Lab in Pittsburgh and in Karlsruhe for participating in experiments during data collection and for various fruitful discussions during the work on this project. Thanks also to Prof. Dillmann's group at the University of Karlsruhe for use of their Polhemus tracker several times for data collection. The neural networks used in this research were trained using the Stuttgart Neural Net Simulator tool [32]. The authors also wish to thank the reviewers for their helpful comments to the manuscript of this paper.

## References

[1] M. Mozer, "The neural network house: An environment that adapts to its inhabitants," in *Proc. Intell. Environments 1998 AAAI Spring Symp.*, 1998, SS-98-92, pp. 110–114.
[2] M. H. Coen, "Design principles for intelligent environments," in *Proc. Intell. Environments 1998 AAAI Spring Symp.*, 1998, SS-98-92, pp. 37–43.
[3] M. Argyle, *Social Interaction*. London, U.K.: Methuen, 1969.
[4] R. Vertegaal, "The gaze groupware system: Mediating joint attention in multiparty communication and collaboration," in *Proc. ACM CHI'99 Conf. Human Factors Comput. Syst.*, Pittsburgh, PA, 1999.
[5] H. Ishii and M. Kobayashi, "Clearboard: A seamless medium for shared drawing and conversation with eye contact," in *Proc. ACM CHI'92 Conf. Human Factors Comput. Syst.*, 1992, pp. 525–532.
[6] M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel, "Multimodal meeting tracker," in *Proc. RIAO 2000: Content-Based Multimedia Inform. Access*, Paris, France, April 2000.
[7] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel, "Multimodal people id for a multimedia meeting browser," in *Proc. ACM Multimedia '99*, 1999.
[8] C. Goodwin, *Conversational Organization: Interaction Between Speakers and Hearers*. New York: Academic, 1981.
[9] N. J. Emery, "The eyes have it: The neuroethology, function and evolution of social gaze," *Neurosci. Biobehavioral Rev.*, vol. 24, pp. 581–604, 2000.
[10] J. Ruusuvuori, "Looking means listening: Coordinating displays of engagement in doctor-patient interaction," *Social Sci. Medicine*, vol. 52, pp. 1093–1108, 2001.
[11] D. I. Perret and N. J. Emery, "Understanding the intentions of others from visual signals: Neurophysiological evidence," *Cahiers de Psychologie Cognitive*, vol. 13, pp. 683–694, 1994.
[12] S. R. H. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? Cues to the direction of social attention," *Trends Cognitive Neurosci.*, vol. 4, no. 2, 2000.
[13] D. Gopher, "Chapter attention," in *The Blackwell Dictionary of Cognitive Psychology*. Oxford, U.K.: Basil Blackwell, 1990, pp. 23–28.
[14] A. R. Diebold Jr., "Chapter anthropology of the comparative psychology of communicative behavior," in *Animal Communication—Techniques of Study and Results of Research*. Bloomington, IL: Indiana Univ. Press, 1968.
[15] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge, U.K.: Cambridge Univ. Press, 1976.
[16] A. H. Gee and R. Cipolla, "Non-intrusive gaze tracking for human–computer interaction," in *Proc. Mechatron. Machine Vision Practice*, 1994, pp. 112–117.
[17] R. Stiefelhagen, J. Yang, and A. Waibel, "A model-based gaze tracking system," in *Proc. IEEE Int. Joint Symp. Intell. Syst.*, 1996, pp. 304–310.
[18] T. S. Jebara and A. Pentland, "Parametrized structure from motion for 3D adaptive feedback tracking of faces," in *Proc. Comput. Vision Pattern Recognition*, 1997.
[19] B. Schiele and A. Waibel, "Gaze tracking based on face-color," in *Int. Workshop Automatic Face and Gesture Recognition*, 1995, pp. 344–348.
[20] R. Rae and H. J. Ritter, "Recognition of human head orientation based on artificial neural networks," *IEEE Trans. Neural Networks*, vol. 9, pp. 257–265, Mar. 1998.
[21] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing," in *Proc. ACM Multimedia '99*, 1999, pp. 3–10.
[22] S. Baker and S. K. Nayar, "A theory of catadioptric image formation," in *Proc. 6th Int. Conf. Comput. Vision, ICCV'98*, Bombay, India, Jan. 1998, pp. 35–42.

[23] J. Yang and A. Waibel, "A real-time face tracker," in *Proc. WACV*, 1996, pp. 142–147.
[24] D. H. Ballard and C. M. Brown, *Computer Vision*. Englegwood Cliffs, NJ: Prentice-Hall, 1982.
[25] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.
[26] D. J. Miller and L. Yan, "Critic-driven ensemble classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 47, pp. 2833–2844, Oct. 1999.
[27] K. Woods, W. P. Kegelmeyer Jr., and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 405–410, Apr. 1997.
[28] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 226–239, Mar. 1998.
[29] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 993–1001, Dec. 1990.
[30] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, "Meeting browser: Tracking and summarizing meetings," in *Proc. Broadcast News Transcription Understanding Workshop*, D. E. M. Penrose, Ed., Lansdowne, Virginia, February 8–11, 1998, pp. 281–286.
[31] R. Gross, M. Bett, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel, "Toward a multimodal meeting record," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2000.
[32] The Stuttgart Neural Network Simulator [Online]. Available: http://www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html

**Rainer Stiefelhagen** received the diploma degree in computer science from the University of Karlsruhe, Germany, in 1996. From 1995 to 1996 he was a Visiting Researcher at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Since 1996, he has been working as a Research Assistant at the University of Karlsruhe, while pursuing his Ph.D. degree. In his Ph.D. research, he has investigated methods for visual tracking of gaze and head orientation, and developed a system to monitor meeting participants' focus of attention. He has also been working on other research topics in multimodal human computer interaction, such as audio-visual speechreading and gaze aware human computer interfaces. His research interests include human-computer interaction, multimodal interfaces, intelligent environments, computer vision, and pattern recognition.

**Jie Yang** (S'93–M'93) received the Ph.D. degree from the University of Akron in 1994.

He is currently a Research Computer Scientist at School of Computer Science in Carnegie Mellon University. He pioneered hidden Markov model for human performance modeling in his Ph.D. dissertation research. He joined the Interactive Systems Laboratories in 1994, where he has been leading research efforts to develop visual tracking and recognition systems for multimodal human computer interaction. He developed adaptive skin color modeling techniques and demonstrated software-based real-time face tracking system in 1995. He has involved developments of many multimodal systems such as gaze-based interface, lipreading system, image-based multimodal translation agent, and multimodal people ID. His current research interests include multimodal interfaces, computer vision, and pattern recognition.

**Alex Waibel** (S'79–M'80) received the B.S. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1979, and the M.S. and Ph.D. degrees in computer science from Carnegie Mellon University, Pittsburgh, PA, in 1980 and 1986, respectively.

He is a Professor of Computer Science at Carnegie Mellon University and at the University of Karlsruhe, Germany. He directs the Interactive Systems Laboratories at both Universities with research emphasis in speech recognition, handwriting recognition, language processing, speech translation, machine learning, and multimodal and multimedia interfaces. At Carnegie Mellon, he also serves as Associate Director of the Language Technology Institute and as Director of the Language Technology Ph.D. program. He was also one of the founding members of the CMU's Human Computer Interaction Institute (HCII) and continues on its steering committee. Dr. Waibel was one of the founders of C-STAR, the international consortium for speech translation research, and currently serves as its chairman. He codirects Verbmobil, the German national speech translation initiative.

Dr. Waibel received the IEEE Best Paper Award in 1990 for his work on the time-delay Neural Networks, and the research prize for technical communication in 1994 for his work on speech translation systems the Alcatel SEL.