Estimating Focus of Attention Based on Gaze and Sound

Rainer Stiefelhagen
Interactive Systems
Laboratories
University of Karlsruhe
Germany
stiefel@ira.uka.de

Jie Yang
Interactive Systems
Laboratories
Carnegie Mellon University
Pittsburgh, PA
yang+@cs.cmu.edu

Alex Waibel
Interactive Systems
Laboratories
Carnegie Mellon University
Pittsburgh, PA
ahw@cs.cmu.edu

ABSTRACT

Estimating a person's focus of attention is useful for various human-computer interaction applications, such as smart meeting rooms, where a user's goals and intent have to be monitored. In work presented here, we are interested in modeling focus of attention in a meeting situation. We have developed a system capable of estimating participants' focus of attention from multiple cues. 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, and 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% error reduction on average compared to using a single modality.

Keywords

Focus of Attention, Gaze Tracking, Meeting Analysis, Intelligent Environments

1. INTRODUCTION

In the last few years there has been a growing interest in building computerized intelligent environments, which aim at supporting humans during various tasks and situations. Research projects include the "digital office" [4], "intelligent house," which adapts illumination and heating to a user's needs [11], "intelligent classroom", which automatically takes notes and provides students with relevant web pages [1], and "smart conferencing rooms", which aim to support cooperative work and help to document and analyze the activities that occur in meetings [5, 15].

For many of these applications, tracking a user's focus of attention would be helpful: in an interactive living room, for example, it would be helpful to know whether a user is trying to control his/her VCR by voice, or whether he/she is talking to someone else in the room. To enable interaction with an intelligent conference room, it would be interesting to know whether a user is focusing on the whiteboard or on another person while talking.

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 omni-directional 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 par-

PUI 2001 Orlando, FL, USA

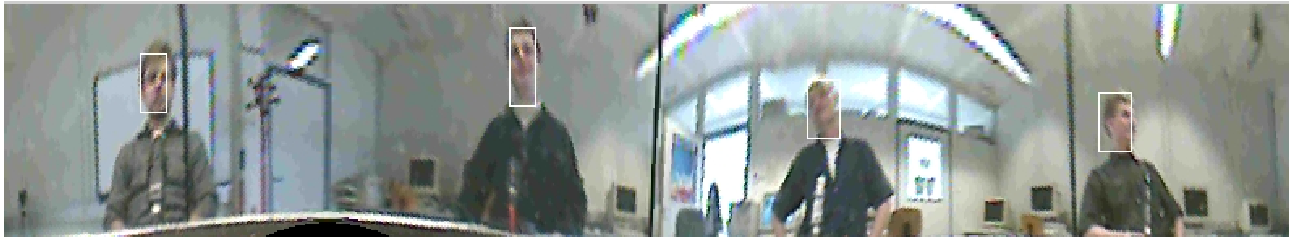


Figure 1: Panoramic view of the scene around the table. Faces are automatically detected and tracked.

ticipants' focus of attention. By using both head pose and sound, we have achieved 76 % accuracy in detecting the participants' focus of attention on three 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 framework presented in this paper can be applied not only to intelligent meeting rooms, but also in other intelligent computerized environments.

The remainder of this paper is organized as follows: In Section 2 we describe how we detect and track participants simultaneously in the view of an omni-directional camera. In Section 3 we introduce an approach to estimate head poses of participants using neural networks. In Section 4 we discuss methods to model the probability distributions of whom a person is looking at based on his head pose. In Section 4.3 we present the idea of predicting a person's focus of attention by monitoring who is speaking. We provide details of how focus of attention can be predicted by knowledge, 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 provide experimental results. In Section 5 we summarize the paper.

2. TRACKING FACES IN A PANORAMIC VIEW

To capture the participants of a meeting, we are using an omnidirectional camera set in the middle of the conference table. Compared to using multiple cameras to capture all participants, as described in our previous work [14], 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.

From the view of the camera, we can compute a panoramic view of the whole scene, as well as perspective views of each user. Figure 1 shows the rectified panoramic image (with faces marked) that is computed from the camera view. To detect and track faces in the panoramic camera view, a statistical skin color model consisting of a two-dimensional Gaussian distribution of normalized skin colors 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 [16]. To detect faces, 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. In addition, some heuristics are used to distinguish hands from faces; see [14] for details.

Once a face is found in the panoramic view, a perspective view of the person can be computed, and the face can again be detected in the perspective view using the face detector. Perspective views of two participants are shown in Figure 2. The automatically detected faces are marked in with boxes. Faces extracted from such perspective views are later used to estimate each participant's head pose with neural nets.

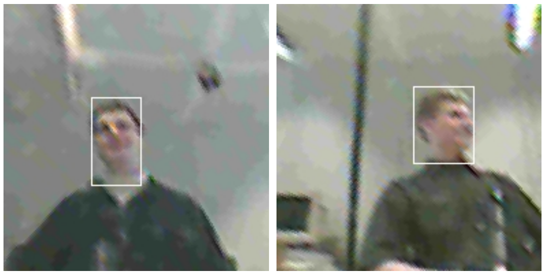


Figure 2: Perspective views of two participants.

3. ESTIMATING HEAD POSE USING NEURAL NETS

In our work, we are using neural networks to estimate head pan and tilt from facial images. Compared to model-based approaches [7, 13, 9], such an appearance-based approach has the advantage that no facial landmark points, such as eyes, nostrils or lip-corners, have to be found in order to compute head pose.

In our previous work on estimating participant's focus in meetings [14], 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 degrees for pan and 8 degrees for tilt on a user independent test set in recent experiments.

In the work presented here, however, only the perspective images generated from the panoramic view were used as input for the neural nets for pose estimation. 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.

3.1 Data Collection

To train and evaluate the neural networks, data from 14 users was collected. All the users were male and four of them had glasses. Hair-styles ranged from almost bold to shoulder-long hair. None of the subjects had a beard.

During data collection, the user was automatically tracked in the panoramic view and a perspective view of the user was generated; see Figure 2. To determine true head pose, the user had to wear a head band with a Polhemus pose tracker sensor attached to it. Using this pose tracker, the head pose with respect to a magnetic field 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.

3.2 Preprocessing of Images

To locate and extract the faces from the perspective user views, the skin-color-based face detector [16] was again used. We have then investigated two different image preprocessing methods as input to the neural nets for pose estimation: 1) Using normalized grayscale images of the user's face as input and 2) applying edge detection to the images before feeding them into the nets.

In the first preprocessing approach, histogram normalization is applied to the grayscale face images as a means of normalizing against different lighting conditions. No additional feature extraction is performed. The normalized grayscale images are down-sampled to a fixed size of 20x30 pixels and then are used as input to the nets. In the second approach, a horizontal and a vertical edge operator plus thresholding is applied to the facial grayscale images. The resulting edge images are down-sampled to 20x30 pixels and are both used as input to the neural nets. Since our previous experiments showed that we obtain the best results when combining the histogram-normalized and the edge images as input to the neural nets , we are only presenting results using this combination of preprocessed images as input to the neural net here. Figure 3 shows the preprocessed images of a user's faces.

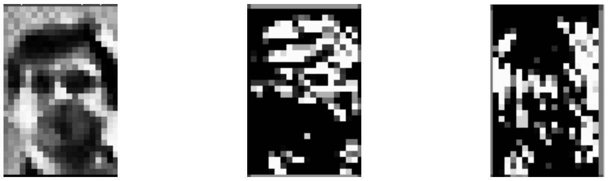


Figure 3: Preprocessed images: normalized grayscale image, horizontal and vertical edge image (from left to right)

3.3 Neural Net Architecture, Training and Results

We have trained separate nets to estimate head pan and tilt. For each net, a multi-layer perceptron architecture with one output unit (for pan or tilt), one hidden layer with 20 to 60 hidden units and an input retina of 20x90 units for the three input images of size 20x30 pixels. Output activations for pan and tilt were normalized to vary between zero and one. Training of the neural net was done using standard back-propagation.

3.3.1 Results

To train a multi-user neural network, we divided the data set of 12 users 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. After training, we achieved a mean error of 7.8 degrees for pan and 5.4 degrees for tilt on the test set.

To determine how well the neural nets can generalize to new users, we have also evaluated the multi-user system on two new users, that were not present in the training set. On the two new users we obtained an average error of 9.9 degrees for pan and 10.3 degrees for tilt. These results demonstrate that the neural networks can generalize also to faces of new users.

3.3.2 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 data could be doubled without the effort of additional data collection. Having more training data should be especially helpful in order to get better generalization on images from new, unseen users. Indeed, after training with the additional data, we achieved an average error of only 9.5 degrees for pan and 9.8 degrees for tilt on the two new users. Table 1 summarizes the results.

	multi-user	user-independent
basic data	7.8 / 5.4	9.9 / 10.3
+ artificial data	3.1 / 2.5	9.5 / 9.8

Table 1: Average error in degrees (pan/tilt) for multi-user and user-independent head pose estimation.

4. MODELING FOCUS OF ATTENTION

Gaze is a good indicator of a person's attention on external objects. When humans pay attention to an external object, they usually orient themselves towards the object of interest so as to have it in the center of their visual field.

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 elses 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 [12]. In fact it has been shown that head orientation strongly influences the perception of gaze, even when the eyes are visible [10].

The idea of this research is to track at whom or what the participants are paying attention to during the course of a

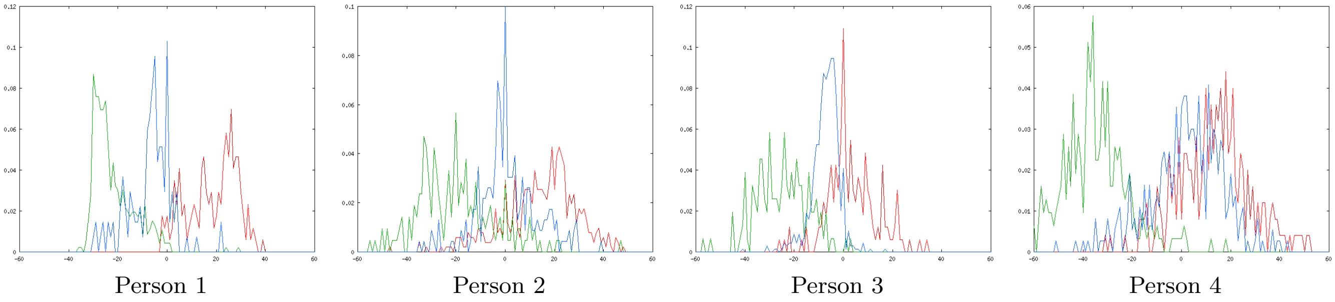


Figure 4: Head pan distributions of four persons in a meeting.

meeting. 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 and 3) map each estimated head orientation to its likely targets using a probabilistic framework.

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 [8]. Clearly, visual attention is influenced by external stimuli, such as noises, movements or speech of other persons. There is some evidence, for example, that two or more people will orient themselves towards 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 [6] (cf. [2]).

Therefore, another approach to estimate at whom or what a person is paying attention to, could 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.

In the remainder of this section, we will first describe how we model focus based on head pose estimations. Then the approach to estimate a person's focus based on sound; e.g., information about who is/was speaking, is described. Finally we'll present results obtained by combining head-pose-based and sound-based focus estimation.

4.1 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 minutes and 30 seconds to 8 minutes and 30 seconds 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 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 assigned to each video frame. These labels could be assigned by listening to the audio streams.

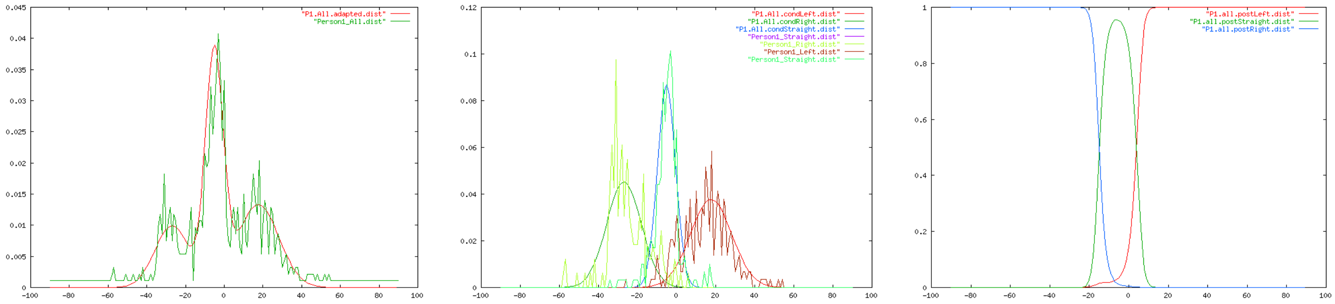
4.2 Modeling Focus Based on Head Rotation

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 2D location of each participant from the positions of the faces found in the panoramic image.

A first, straightforward 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 $\mathbf{T}_i$ sits nearest the position to which S is looking.

Gaze is not only determined by head pose, however, 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. In our meeting recordings we observed that some people turned their heads more than others, who relied more on eye movements instead and less head turning when looking at other people. Figure 4 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 and a higher overlap of the distributions can be observed.

Motivated by these observations, we have developed a Bayesian approach to estimate at which target a person is



a) b) c)

Figure 5: 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 $P(\textbf{Focus}|x)$ for resulting from the found mixture of Gaussians

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 towards a certain target person T , given the person's observed horizontal head rotation x_S. 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 for a person S, it is then possible to compute the posterior probabilities $P(\text{Focus}_S|T_i)$ 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 $P(\text{Focus}_S = T|x_S)$, however, it is 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 style very well, nor would the priors necessarily be the same in different meetings.

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

4.2.1 *Unsupervised Adaptation of Model Parameters*

In our approach, we assume that the class-conditional head pan distributions, such as depicted in Figure 4, 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,

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

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

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 [3]:

$$\begin{aligned}
\mu_j^{new} &= \frac{\sum_n P^{old}(j|x^n)x^n}{\sum_n P^{old}(j|x^n)} && (3)\\
(\sigma_j^{new})^2 &= \frac{1}{d}\frac{\sum_n P^{old}(j|x^n)||x^n - \mu_j^{new}||^2}{\sum_n P^{old}(j|x^n)} && (4)\\
P(j)^{new} &= \frac{1}{N}\sum_n P^{old}(j|x^n). && (5)
\end{aligned}$$

To initialize the means μ_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 classconditionals $p(x|\text{Focus} = T)$, 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 equation (1). Furthermore, the individual Gaussian components can be assigned to corresponding target persons based on their relative position around the table.

Figure 5 shows an example of the adaptation on pan observations from one user. In Figure 5a) the distribution of all head pan observations of the user is depicted together with the Gaussian mixture that was adapted as described above. Figure 5b) depicts the real class-conditional head pan distributions of that person, together with the Gaussian components taken from the Gaussian mixture model depicted in Figure 5a). 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 depicted for comparison and are of course not necessary for the adaptation of the Gaussian components. Figure 5c) depicts the posterior probability distribution resulting from the adapted class-conditionals and class priors.

4.2.2 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 $p(x|\text{Focus})$, the class-priors $P(\text{Focus})$ and the observation distributions $p(x)$ were automatically adapted to compute the posterior probabilities $p(\text{Focus} = T_i|x)$ 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 twelve users in the three meetings, the correct focus target could be detect on average in 73.9% of the frames. Table 2 show the average results on the three meetings.

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

Table 2: Percentage of correct assigned focus targets based on computing $P(\text{Focus}|\text{head pan})$.

4.3 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.

$\vec{A} = (a_S a_L a_C a_R)$	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

Table 3: 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).

Table 3 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 ("**S**elf") 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 ("**C**enter")to S was speaking, and entry a_R indicating whether the person to its **r**ight 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 3 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 3 indicates in which direction subjects looked on average, regardless of speaking conditions. It can be seen that there is a bias towards looking straight; i.e., to the person opposite to the subject.

The entries of Table 3 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 $j \in \{$ "Left", "Straight", "Right" $\}$ denote the possible persons to look at, and where

$$\vec{A} = (a_{\mathbf{S}elf}, a_{\mathbf{L}eft}, a_{\mathbf{C}enter}, a_{\mathbf{R}ight})$$

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.

4.4 Combining Gaze and Sound to Predict Focus

In the previous section it was shown, how we can determine the probability $p(\text{Focus}|\text{Sound})$; i.e., the probability that a person is looking towards 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.

In section 4 we showed, how to compute $P(\text{Focus}_S = T_i|x_S)$, the posterior probability, that a person S is looking towards 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 $p(\text{Focus}|gaze)$ – 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 α, ranging from 0.0 to 1.0. On the three meetings, the optimal values of α ranged from 0.3 to 0.6 By setting α 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. Compared to the prediction accuracy of 73.9 % using gaze only, this corresponds to a relative error reduction of 3.4 %.

	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 %

Table 4: Focus-prediction using gaze only, sound only and prediction using both, gaze and sound.

4.5 Using the Sound History to Predict Focus

In the previous section, 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 $A^t, A^{t-1}, ..., A^{t-N}$ to predict the probability that a person S is looking towards one of the others; i.e., to estimate $P(\text{Focus}|A^t, A^{t-1}, ..., A^{t-N})$.

In this work, we have chosen to use a neural network to predict $P(\text{Focus}|A^t, A^{t-1}, ..., A^{t-N})$. 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 the previous section.

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 5.

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

Table 5: Focus-prediction using one frame and ten frames of speaker information. Neural networks were trained to predict $P(\textbf{Focus}|A^t, A^{t-1}, ..., A^{t-9})$.

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

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

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

Table 6 summarizes the results we obtained by using sound-only based focus prediction, gaze-only based focus estimation and combined estimation.

	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 %

Table 6: 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.

5. 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 towards 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.

Although moving participants can successfully be tracked in the camera, the current system cannot handle significant movements of the meeting participants. This limitation comes from both the software and the hardware. The current focus of attention model relies on probability distributions related to participants' locations. Significant movements will change those distributions. An adaptive procedure is needed to update the distributions when participants are moving. This will be our future work. In addition, due to the poor resolution of the omnidirectional camera that we are using, it is expected that estimating head orientation will not be possible when people are more than 2 meters away from the camera. A higher resolution camera will be helpful to solve this problem.

Acknowledgments

This Research is partially supported bt the Defense Advanced Research Projects Agency under contract number DAAD17-99-C-0061, and by the National Science Foundation under Grant No. IIS-9980013. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA and the NSF, or any other party.

6. REFERENCES

[1] G. D. Abowd, C. Atkeson, A. Feinstein, C. Hmelo, R. Kooper, S. Long, N. Sawhney, and M. Tani. Teaching and learning as multimedia authoring: The classroom 2000 project. In *Proceedings of the ACM Multimedia'96 Conference*, pages 187–198, November 1996.

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

[3] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, 1995.

[4] M. Black, F. Brard, A. Jepson, W. Newman, E. Saund, G. Socher, and M. Taylor. The digital office: Overview. In *Proceedings of the 1998 AAAI Spring Symposium on Intelligent Environments*, volume AAAI Technical Report SS-98-02. AAAI, AAAI Press, March 1998.

[5] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox. Room with a rear view: Meeting capture in a multimedia conference room. *IEEE Multimedia Magazine*, 7(4):48–54, Oct-Dec 2000.

[6] A. J. Diebold. *Animal Communication - Techniques of Study and Results of Research*, chapter Anthropology of the comparative psychology of communicative behavior. Bloomington: Indiana University Press, 1968.

[7] A. H. Gee and R. Cipolla. Non-intrusive gaze tracking for human-computer interaction. In *Proc. Mechatronics and Machine Vision in Practise*, pages 112–117, 1994.

[8] D. Gopher. *The Blackwell dictionary of Cognitive Psychology*, chapter Attention, pages 23–28. Basil Blackwell Inc., 1990.

[9] T. Jebara and A. Pentland. Parametrized structure from motion for 3d adaptive feedback tracking of faces. In *Proceedings of Computer Vision and Pattern Recognition*, 1997.

[10] S. R. Langton, R. J. Watt, and V. Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Neuroscience*, 4(2), 2000.

[11] M. Mozer. The neural network house: An environment that adapts to its inhibitants. In *Intelligent Environments, Papers from the 1998 AAAI Spring Symposium*, number Technical Report SS-98-92, pages 110–114. AAAI, AAAI Press, 1998.

[12] D. Perret and N. Emery. Understanding the intentions of others from visual signals: neurophysiological evidence. *Cahiers de Psychologie Cognitive*, 13:683–694, 1994.

[13] R. Stiefelhagen, J. Yang, and A. Waibel. A model-based gaze tracking system. In *Proceedings of IEEE International Joint Symposia on Intelligence and Systems*, pages 304 – 310, 1996.

[14] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing. In *Proceedings of ACM Multimedia '99*, pages 3–10. ACM, 1999.

[15] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In D. E. M. Penrose, editor, *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pages 281–286, Lansdowne, Virginia, February. 8-11 1998. DARPA, Morgan Kaufmann.

[16] J. Yang and A. Waibel. A real-time face tracker. In *Proceedings of WACV*, pages 142–147, 1996.