Tracking Focus of Attention in Meetings

Rainer Stiefelhagen

Interactive Systems Laboratories
Universität Karlsruhe (TH)
Germany
E-mail: stiefel@ira.uka.de

Abstract

This paper presents an overview of our work on tracking focus of attention in meeting situations. We have developed a system capable of estimating participants' focus of attention from multiple cues. In our system we employ an omni-directional camera to simultaneously track the faces of participants sitting 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.

In addition this work reports recent experimental results: In order to determine how well we can predict a subject's focus of attention solely on the basis of his or her head orientation, we have conducted an experiment in which we recorded head and eye orientations of participants in a meeting using special tracking equipment. Our results demonstrate that head orientation was a sufficient indicator of the subjects' focus target in 89% of the time. Furthermore we discuss how the neural networks used to estimate head orientation can be adapted to work in new locations and under new illumination conditions.

1 Introduction

In recent years much research has been done in building computerized intelligent environments, which aim at supporting humans during various tasks and situations. Research projects include the “digital office” [5], ”intelligent house,” which adapts illumination and heating to a user's needs [12], ”intelligent classrooms,” 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 [7, 18].

In order to make such intelligent and interactive environments respond appropriately to their users' needs, it is necessary to equip them with perceptive capabilities to capture as much relevant information about its users and the context in which they act as possible. Obtaining knowledge about a person's focus of attention is a major step towards a better understanding of what users do, how and with what or whom they interact or to what they refer.

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

A body of research literature suggests that humans are generally interested in what they look at [19, 4, 9] and the close relationship between gaze and attention during social interaction has been emphasized [2, 3, 8]. In addition, recent user studies reported strong evidence that people naturally look at the objects or devices with which they interact [11, 6].

A first step to determine someone's focus of attention, therefore is, to find out in which direction the person looks. There are two contributing factors in the formation of where a person looks: head orientation and eye orientation. In this work head orientation is considered as a sufficient cue to detect a person's direction of attention. Relevant psychological literature offers a number of convincing arguments for this approach (e.g. [8, 3, 17]) and the feasibility of this approach is demonstrated experimentally in this paper.

A practical reason to use head orientation to estimate a person's focus of attention, is, that in scenarios such as addressed in this work, head orientation can be estimated with non-intrusive methods while eye orientation can not.

Our approach to tracking at whom participants look, i.e. their focus of attention, is the following:

1. Detect all participants in the scene,
2. estimate each participant's head orientation and
3. map each estimated head orientation to its likely targets using a probabilistic framework.

This approach is of course not perfect. Since eye gaze is neglected, a certain amount of errors is introduced. The noisy estimation of head orientations from camera images introduces additional errors.

To improve the robustness of focus of attention tracking, we therefore would like to combine various sources of information. We have found that focus of attention is correlated to who is speaking in a meeting and that it is possible to estimate a person's focus of attention based on the information of who is talking at or before a given moment. To estimate

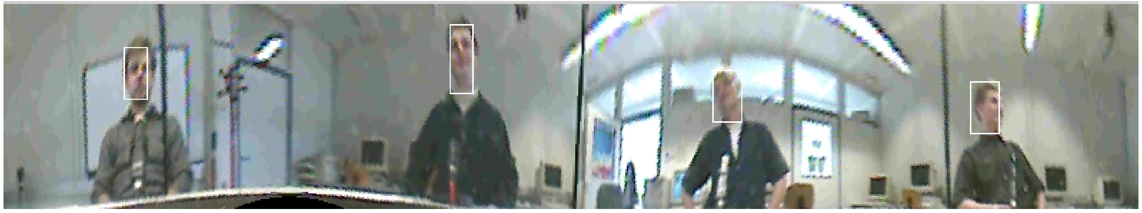


Figure 1. Panoramic view of the scene around the table. Faces are automatically detected.

where a person is looking, based on who is speaking, probability distributions of where participants are looking during certain “speaking constellations” are used.

The accuracy of sound-based prediction of focus of attention can furthermore significantly be improved by 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.

Finally, the head pose based and the sound-based estimations are combined to obtain a multimodal estimation of the participants’ focus of attention. This leaded to significant improvements compared to using just one modality for focus of attention tracking alone.

Our system for focus of attention detection in meetings has been successfully installed in both our labs at the Universität Karlsruhe, Germany and at Carnegie Mellon University in Pittsburgh, USA. A problem when porting the system to a new location is the need for appropriate training images for the neural network based approach to head orientation estimation. We therefore also investigated how much training/adaptation data is necessary to port the system to a new location.

The remainder of this paper is organized as follows: In section 2 we discuss how participants are tracked and how head pose is estimated in our system. In section 3 we introduce our probabilistic approach to model at whom subjects look at based on their head orientations. In section 4 we present a user study investigating how reliably focus of attention can be estimated based on head orientation alone in meetings. In section 5 we suggest that focus of attention tracking could benefit from also tracking other relevant cues and show that information about who is or has been speaking at a given moment can be used to improve focus of attention tracking accuracy. In section 6 we discuss portability issues of our system. We conclude the paper in section 7.

2 Simultaneous Head Pose Tracking in Meetings

We use an omni-directional camera to capture the scene around a meeting table. Compared to using several cameras to capture the scene, this simplifies the recording since no camera control, calibration or synchronization is necessary.

In the panoramic view of the meeting scene (see Figure 1 for an example) we then detect the participants faces by searching for skin-colored regions and use some heuristics to distinguish skin-colored hands from faces [13].

For each detected participant a rectified (perspective) view is computed (see Figure 2. Faces extracted from these views are then used to estimate each participant’s head pose.

2.1 Head Pose Estimation with Neural Networks

We use neural networks to estimate head pan and tilt from such facial images [13]. In our approach, preprocessed facial images are used as input to the neural networks, and the networks are trained so as to estimate the horizontal (pan) or vertical (tilt) head orientation of the input images. Separate networks were trained to estimate head pan and tilt. These networks contained one hidden layer and one output unit, which encodes the head orientation in degrees. By training multi-user networks on images from twelve users we achieved average estimation errors as low as three degrees for pan and tilt. On images from new users, head orientation could be estimated with an average error of 10 degrees for pan and tilt. More details can for be found in [13].

3 From Head Pose to Focus of Attention

In our approach we first estimate a persons head orientation and then detect at whom a person was looking based on his or her estimated head orientation.

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 multi-class 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 where participants sit at different locations than they were sitting during collection of the training data.

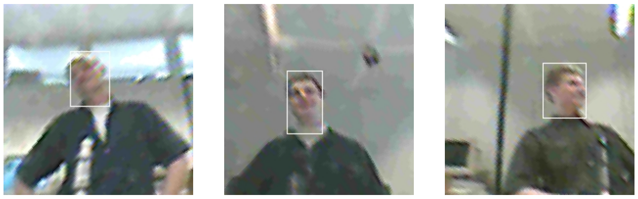
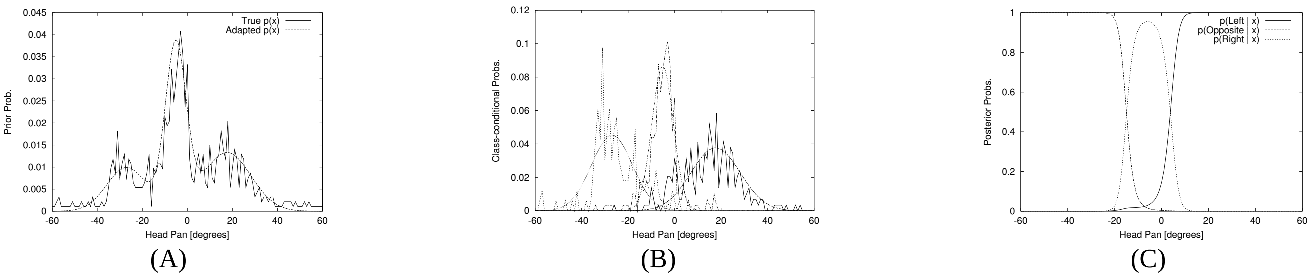


Figure 2. Perspective views of participants.



(A) (B) (C)

Figure 3. Unsupervised adaptation of model parameters to find $P(Focus|HeadPan)$ (see text).

We have developed a Bayesian approach to estimate at which target a person is looking, based on his observed head orientation [14, 15]. More precisely, we wish to find $P(\text{Focus}_S = T_i|x_S)$, the probability that a subject S is looking towards a certain target person T_i, given the subject's observed horizontal head orientation x_S, which is the output of the neural network for head pan estimation. Using Bayes formula, this can of be decomposed into

$$P(\text{Foc.}_S = T_i|x_S) = \frac{p(x_S|\text{Foc.}_S = T_i)P(\text{Foc.}_S = T_i)}{p(x_S)},$$

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

In order to compute $P(\text{Focus}_S = T|x_S)$, 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 orientations of a person over time.

In order to adapt the parameters of our model to varying target locations and to the different head turning styles of the participants, we have developed an unsupervised learning approach to find the head pan distributions of each participant when looking at the others.

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

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

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

The number of Gaussians M is set to the number of other participants that are detected around the table. The parameters of the mixture model can be adapted so as to maximize the likelihood of the pan observations given the mixture model using the EM algorithm (for further details see [14]). To initialize the means μ_j of the mixture model, k-means clustering is performed on the pan observations. Parameters are iteratively updated as follows:

After adaptation of the mixture model, we use the individual Gaussian components as an approximation of the class-conditionals $p(x|\text{Focus} = T)$ of our focus of attention model described in equation (3). We furthermore use the priors of the mixture model, $P(j)$, as the focus priors $P(\text{Focus} = T)$. To assign the individual Gaussian components and the priors to their corresponding target persons, the relative position of the participants around the table are used.

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

Table 1. Correctly assigned focus targets based on head pan (in percent)

Figure 3 shows an example of the adaptation on pan observations from one user. The mixture of Gaussian distribution is adapted to the distribution of all head pan observations of the user (Fig. 3(A)). Figure 3(B) depicts components of the mixture model. For comparison, the real class-conditional head pan distributions are shown. Figure 3(C) depicts the resulting posterior distributions.

3.1 Experimental Results

We evaluated our approach on several meetings that we recorded. In each of the meetings four participants were sitting around a table and were discussing a freely chosen topic. Video was captured with the panoramic camera and audio was recorded using several microphones.

In each frame we manually labeled at whom each participant was looking. These labels could be one of *"Left"*, *"Right"* or *"Straight"*, meaning a person was looking to the person to his left, to his right, or to the person at the opposite. If the person wasn't looking at one of these targets; e.g., the person was looking down on the table or was staring up to the ceiling, the label *"Other"* was assigned. In addition, labels indicating whether a person was speaking or not, were manually assigned for each participant and each video frame.

Table 1 shows the evaluation results on the four recorded meetings. In the table, the average accuracy on the four participants in each meeting is indicated.

For the evaluation the faces of the participants were automatically tracked. Head pan was then computed using the neural network-based system to estimate head orientation.

For each of the meeting participants, the class-conditional head pan distribution $p(x|\text{Focus})$, the class-priors $P(\text{Focus})$ and the observation distributions $p(x)$ were adapted as described in the previous section, and the posterior probabilities $P(\text{Focus} = T_i|x)$ for each person were computed. During evaluation, the target with the highest posterior probability was then chosen as the focus of attention target of the person in each frame.

For the evaluation, we manually marked frames where a subject's face was occluded or where the face was not correctly tracked. These frames were not used for evaluation. Face occlusion occurred in 1.6% of the captured images. Occlusion sometimes happened, when a user covered his face with his arms or with a coffee mug for example; sometimes a face was occluded by one of the posts of the camera. In another 4.2% of the frames the face was not correctly tracked. We also did not use frames where a subject did not look at one of the other persons at the table. This happened in 3.8 % of the frames. Overall 8.2% of the frames were not used for evaluation since at least one of the above indications was given.

4 Head Pose versus Eye Gaze

In this work, head orientation is used to predict a person's focus of attention in meetings. This is done because head orientation is assumed to be a reliable indicator of the direction of someone's attention during social interaction and because eye gaze of several meeting participants cannot be easily tracked without the use of intrusive hardware.

Since we estimate where a person is looking at based on his head orientation, the following question suggests itself: how well can we predict at whom a person is looking at, merely on the basis of his or her head orientation?

To answer this question, we have analyzed the gaze of four people in meetings using special hardware equipment to measure their eye gaze and head orientation [16]. We have analyzed the gaze and head orientation data of the four people to answer the following questions:

1. How much does head orientation contribute to gaze?

2. How accurately can we predict at whom the person was looking at, based on his head orientation only?

4.1 Data Collection

The setting in this experiments is a round-table meeting. There are four participants in the meeting, and a session of data for about ten minutes with each participant is collected. In each session, one of the participants, the subject, wears a head-mounted gaze tracking system from ISCAN Inc. [10].This system can estimate and record the following data with a frame rate of 60 Hz: the subject's head position, head orientation, eye orientation, pupil diameter, and the overall gaze (line of sight) direction. All these estimations have a precision of better than one degree. head-mounted gaze-tracker.

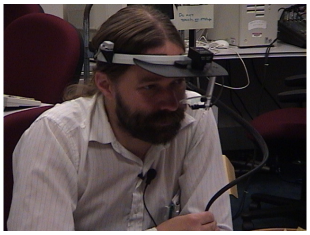


Figure 4. A participant wearing the head-mounted eye and head tracking system.

4.2 Contribution of Head Orientation to Gaze

First, we analyzed the contribution of head orientation and eye orientation to the overall gaze direction along the horizontal axis. On the data from the four participants we found that in 87% of the frames head orientation and eye gaze pointed in the same direction (left or right). For these frames we calculated the contribution of head orientation to the overall line of sight orientation. Since the horizontal component of the line of sight los_x is the sum of horizontal head orientation ho_x and horizontal eye orientation eo_x, the percentage of head orientation to the horizontal direction of gaze is computed as head contribution $= \frac{ho_x}{los_x}$.

Table 2 summarizes the results of four experiment sessions. From the results, we can see several interesting points: 1) Most of the time, the subjects rotate their heads and eyes in the same direction to look at their focus of attention target (87%). 2) The subjects vary much in their usage of head orientation to change gaze direction: from Subject 2's 53% to Subject 4's 96%, with an average of 68.9%. 3) Even for Subject 2, whose head contribution is the least among the four participants, head orientation still contributes more than half of the overall gaze direction. 4) Eye-blinks (or eye-tracking failures) take about 20% of the frames, which means even for commercial equipments as accurate as the ISCAN system we used, eye orientation, and thus the overall gaze direction cannot be obtained in about a fifth of the time.

Subject	eye blinks	same direct.	head contrib.
1	25.4%	83.0%	62.0%
2	22.6%	80.2%	53.0%
3	19.2%	91.9%	63.9%
4	19.5%	92.9%	96.7%
Average	21.7%	87.0%	68.9%

Table 2. Eye blinks and contribution of head orientation to the overall gaze.

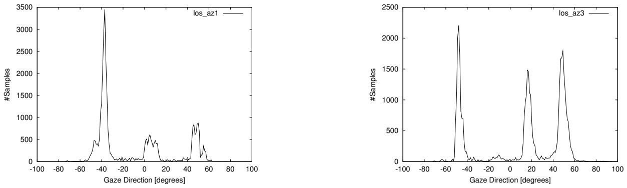


Figure 5. Histograms of horizontal gaze directions of two subjects.

4.3 Predicting the Gaze Target Based on Head Orientation

We approached the second question we proposed before in this particular meeting application: How accurately can we predict at whom the subject was looking at, on the basis of his head orientation? Answering this question gives us an idea of the upper limit of the accuracy that can be obtained when the focus of attention target is estimated based on head orientation alone.

Labeling Based on Gaze Direction

To automatically determine at which target person the subject was looking at (focus of attention), the gaze direction was used. Figure 5 shows the histograms of the horizontal gaze direction of two of the participants. In each of the histograms, it can be seen that there are three peaks. We assume that these belong to the direction where the other participants at the table were sitting. We have automatically determined the peaks in the horizontal line-of-sight datafiles using the k-means algorithm. The peaks found were then used as the directions where the other persons were sitting; and in each frame focus of attention labels were assigned based on the least distance of the actual horizontal line-of-sight to the three target directions.

Prediction Results

To see how accurate the focus target can be estimated based on observing head orientation alone, we used exactly the same method to find the focus targets as described in section 3. The only difference now is, that in the previous experiment, focus was determined based on noisy head pan *estimates* as given by the neural networks, whereas now, focus targets are found based on accurate head pan *measurements* as given from the gaze tracking equipment.

Subject	1	2	3	4	Avg.
Accuracy	85.7	82.6	93.2	93.2	88.7

Table 3. Focus detection based on exact measurements of horizontal head orientation (in percent).

.

Table 3 summarizes the results on the four participants. The results show that the focus of attention target can be correctly estimated with only head orientation data in 82.6% (Subject 2) to 93.2% (Subject 3 and 4) of the frames, with an average of 88.7%. This can be seen as the upper limit of accuracy that we can get in head orientation based focus of attention estimation in such a scenario. These results also show that head orientation is indeed a reliable cue for detecting at whom participants look at in meetings.

5 Predicting Focus Based on Sound

Attention is clearly influenced by external stimuli, such as noises, movements or speech of other persons. Monitoring and using such cues might therefore help us to bias certain targets of interests against others.

We have found that focus of attention is correlated to who is speaking in a meeting and that it is possible to estimate a person's focus of attention based on the information of who is talking at or before a given moment [14, 15].

In our first experiment to predict focus from sound (speakers) we analyzed at whom the four participants in the recorded meetings were looking during certain "speaking" conditions. Here, "speaking" was treated as a binary vector; i.e., each of the four participants was either labeled as "speaking" or "not speaking" in each video frame. Now, using this binary "speaking" vector 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 [14].

By using only the speaker labels to make a sound-based focus prediction, we were able to predict the correct focus of each participant 56.3% of the time in the evaluation meetings.

5.1 Using Temporal Speaker Information to Predict Focus

We have also investigated, whether the prediction of the focus of attention could benefit from temporal speaker information.

Thus, we trained neural networks to estimate at which target person a subject is looking at, given a history of audio-observations as input. The neural net we use consists of an input layer of (N+1)*4 input units, corresponding to the (N+1) audio-observation vectors, one hidden layer and three output units, corresponding to the three target persons that a subject can look at. As audio-observations at each time step, the binary audio-observation vectors described in the previous section, were chosen.

As output representation a 1-of-N representation was used; i.e., during training the output corresponding to the correct target class was set to 1 and the other output units were set to zero. As error criterion, the commonly used mean square error criterion was used.

After training, such a network will approximate the a posteriori probabilities of the focus targets F_i given the sequence of observed audio-information:$P(\text{Focus}|A^t, A^{t-1}, ..., A^{t-N})$.

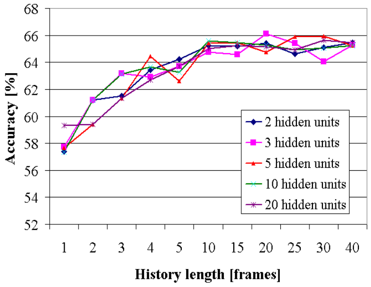


Figure 6. Sound-based focus prediction results with different audio-history lengths and different number of hidden units.

Figure 6 shows the average sound-based focus prediction results on the four evaluation meetings for different histories of audio-vectors used as input and for networks with different amounts of hidden units. The best accuracy is 66.1%. This was achieved using three hidden units and a history of 20 audio-vectors, corresponding to approximately eight seconds of audio-information. Please refer to [14] or [15] for more details.

5.2 Combining Head Pose and Sound to Predict Focus

The two independent predictions of a person's focus – $P(\text{Focus}|\text{Sound})$ and $P(\text{Focus}|HeadPose)$ – can be combined to obtain a prediction of a person's focus which is based on both the observation, who is speaking, and based on the person's head rotation.

We combined the predictions by computing the weighted sum of both modalities:

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

By setting α to 0.6, we achieved an average accuracy of 75.6% on the recorded meetings. Table 4(a) summarizes the results we obtained by using sound-only based focus prediction, head orientation-only based focus estimation and combined estimation.

	Head Pose only	Sound only	Combined
Meeting A	68.8	59.2	69.1
Meeting B	73.4	69.6	77.8
Meeting C	79.5	61.3	80.6
Meeting D	69.8	74.3	74.7
Average	72.9	66.1	75.6

Table 4. Focus-prediction results (in percent).

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.

6 Portability of the System

In this section we discuss how the presented system for focus of attention tracking can be installed in a new location.

The main problem when installing the system in a new location is that the illumination conditions in the new location might be completely different from the conditions in which the training data for the neural networks for head orientation estimation was collected.

To investigate which steps are necessary to successfully move the focus of attention tracking system to a new location, we have installed the system in both our labs at the Universität Karlsruhe in Germany and at Carnegie Mellon University in Pittsburgh, USA.

In the remainder of this section we report about experiments on how the neural network for head pan estimation can be adapted to work under new conditions. We examine how much adaptation data is necessary to obtain reasonable focus of attention tracking performance and compare the results with adapted networks to the results obtained with neural networks that are trained from scratch with new data.

6.1 Data Collection at CMU

In order to train neural networks for head pan estimation in the new location, we have collected training images from twelve users in our lab at CMU (the new location). As during the data collection in Karlsruhe, subjects had to wear a head band with a Polhemus pose tracker sensor attached to it so that true head pose could be determined. Images of the person's head were captured with an omni-directional camera as described in section 2 and were recorded together with the person's head pose. From each person, we collected training images at several locations around the meeting table. The data collection took about fifteen minutes for each participant. Altogether we collected around 27.000 training images from twelve persons.

6.2 Training New Networks from Scratch

We first trained neural networks for head pan estimation using only the data that was collected at CMU. To see how much training data is necessary for reasonable generalization, we trained different networks using increasing subsets of the data. To evaluate the performance of the networks, data from four subjects was kept aside as a user-independent test set.

We trained networks on images from one up to all eight subjects in the training set. The neural network architecture and training was identical to those used with the networks trained with the data from Karlsruhe. The networks were trained on the training data set and a cross-evaluation set was used to determine the number of training iterations.

Figure 7 shows the results obtained on the user independent test set from CMU (top curve). It can be seen that the average pan estimation error on the test set is as high as twenty degrees when only images from one subject were used for training. The pan estimation error then gradually decreases, when training images from more subjects are

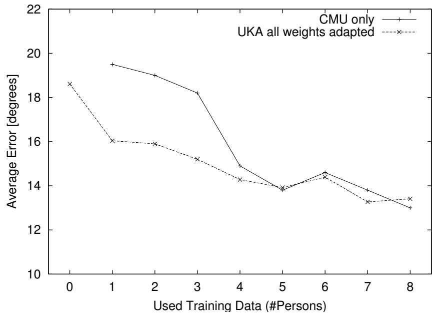


Figure 7. Pan estimation results in new location with a newly trained and with an adapted network (see text).

added. When all eight subjects were used for training, an average pan estimation error of 13 degrees was obtained.

We also trained one neural network on images on all the available twelve subjects. For training we used 80% of all the images. 10% of the images were used for cross-evaluation and the remaining 10% of the images were used as a test set. With this multi-user network for pan estimation, we achieved an average error of 7.6 degrees on the test-set.

6.3 Adapting a Trained Network

We then investigated whether and how well the network which was previously trained on data collected in Karlsruhe – the “UKA-network”' – could be adapted to the new CMU images, by using the different training data sets from CMU for adaptation.

We adapted the UKA-network by retraining all its weights on the different adaptation data sets from CMU. Training was done using standard back-propagation with a learning parameter of 0.1. To determine when the adaptation process should stop, a cross-evaluation set containing images from an additional subject was used. Typically, adaptation stopped after two to six iterations.

We adapted the UKA-net with images from one to all eight subjects of the CMU training set. The performance of the adapted networks was then evaluated on the the user-independent from CMU.

The results are also shown in Figure 7 (lower curve). With the unadapted UKA-network an average error of 19 degrees was obtained on the test set. By using images from one subject from CMU for adaptation, the average error decreases to 15.6 degrees. When all training data from eight subjects is used for adaptation, the average pan estimation error decreases to 13 degrees.

It can be seen that pan estimation works significantly better with the adapted networks when only little data is available for training or adaptation. In our experiments, the newly trained network only reached the performance of the adapted UKA-network, when training images from at least five subjects were available for training.

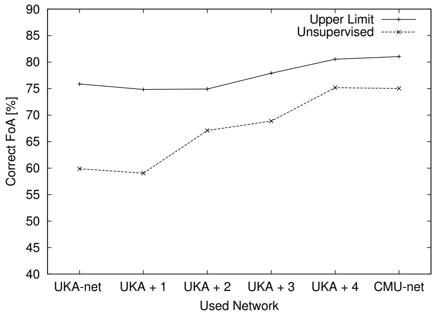


Figure 8. Accuracy of focus of attention detection on a meeting recorded at CMU (see text).

6.4 Focus of Attention Detection Results

To measure how well focus of attention can be estimated using the different neural networks, we have collected a meeting with four participants in our lab at CMU.

The focus of attention tracking system was run on the recorded meeting with different networks for pan estimation. For the evaluation we used the unadapted UKA-network, the adapted UKA-networks and the neural network that was trained on images from all twelve subjects in our data set from CMU.

For each network we evaluated the focus of attention detection accuracy using the mixture of Gaussian approach presented in chapter 3. All parameters of the Gaussian mixture model were adapted completely unsupervised.

Figure 8 shows focus of attention detection accuracy on the meeting that was recorded at CMU for the different networks used for head pan estimation.

Using the UKA-network for head pan estimation, focus of attention could be detected in only 60% of the time on the meeting, with a possible upper limit of 76%. By adapting the UKA-network with data collected at CMU the performance increases to 75% focus of attention detection accuracy when images from four subjects were in the adaptation set (“UKA + 4”). This performance is already as good as the performance obtained with the CMU-network, which was trained on images from twelve subjects collected at CMU.

6.5 Discussion

Our experiments suggest that a network which has already been trained to estimate head pan from images taken in one location can be adapted to work in a new location and under different illumination conditions by collecting a limited number of images in the new location and adapting the networks' weights with the new images. In our experiments we achieved good focus of attention tracking results in the new location by using adaptation images from only four subjects. These images could be collected in approximately one hour. Our experiments also showed that adapting an existing network for pan estimation, which has been trained on images taken in different lighting and camera conditions, leads to better pan estimation results than training networks from scratch with images from the new location when only a small amount of training images are available.

7 Conclusions

In this paper we presented a system to track the focus of attention of participants in a meeting. 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 their head orientations 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 73 % 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.

In addition, 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.

To answer how precisely focus of attention can be predicted in a meeting just based on the participants' head orientations we have recorded eye gaze and head orientations of four subjects in a meeting. The user study clearly demonstrated that head orientation is a reliable cue to detect at whom someone is attending to. In the meetings which we recorded for this study, we were able to correctly determine at whom the subject was looking in 89% of the time just based on the subject's head orientation.

Finally, we have investigated how a neural network for head pan estimation can be adapted to work in a new location. Our experiments showed that adaptation images from only four subjects were sufficient to achieve good focus of attention detection accuracy in a new location with completely different illumination conditions.

Acknowledgments

I would like to thank everybody at the Interactive Systems Labs at Carnegie Mellon University and in Karlsruhe for their great support over the years. Special thanks to those who helped organizing the meetings and data collection sessions at CMU and to everybody who participated in these sessions.

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. *Social Interaction*. Methuen, London, 1969.

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

[4] P. Barber and D. Legge. *Perception and Information*, chapter 4: Information Acquisition. Methuen, London, 1976.

[5] 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.

[6] B. Brumitt, J. Krumm, B. Meyers, and S. Shafer. Let there be light: Comparing interfaces for homes of the future. *IEEE Personal Communications*, August 2000.

[7] 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.

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

[9] A. J. Glenstrup and T. Engell-Nielsen. Eye controlled media: Present and future state. Technical report, University of Copenhagen, http://www.diku.dk/users/panic/eyegaze/, 1995.

[10] ISC. Iscan inc. http://www.iscaninc.com/.

[11] P. P. Maglio, T. Matlock, C. S. Campbell, S. Zhai, and B. A. Smith. Gaze and speech in attentive user interfaces. In *Proceedings of the International Conference on Multimodal Interfaces*, volume 1948 of *LNCS*. Springer, 2000.

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

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

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

[15] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks. Special Issue on Intelligent Multimedia Processing*, July 2002.

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

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

[18] 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.

[19] A. L. Yarbus. Eye movements during perception of complex objects. In L. Riggs, editor, *Eye Movements and Vision*, pages 171–196. Plenum Press, New York, 1967.