Tracking Head Pose and Focus of Attention with Multiple Far-field Cameras

Michael Voit, Rainer Stiefelhagen
Interactive Systems Labs, Universität Karlsruhe (TH)
Germany
voit@ira.uka.de, stiefel@ira.uka.de

ABSTRACT

In this work we present our recent approach on estimating head orientations and foci of attention of multiple people in a smart room, which is equipped with several cameras to monitor the room. In our approach, we estimate each person's head orientation with respect to the room coordinate system by using all camera views. We implemented a Neural Network to estimate head pose on every single camera view, a Bayes filter is then applied to integrate every estimate into one final, joint hypothesis. Using this scheme, we can track peoples' horizontal head orientations in a full 360° range at almost all positions within the room. The tracked head orientations are then used to determine who is looking at whom, i.e. people's focus of attention. We report experimental results on one meeting video, that was recorded in the smart room.

Categories and Subject Descriptors

I.4 [**IMAGE PROCESSING AND COMPUTER VISION**]: Scene Analysis

General Terms

Algorithms, Experimentation

Keywords

Focus of Attention, Gaze, Head Orientation, Head Pose, Neural Networks, Bayesian Filter

ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

1. INTRODUCTION

In order to build intelligent user interfaces, the perception of the user and his or her activities is a paramount goal. A number of research projects focus on developing technologies for the audio-visual perception of people, their locations, identities, gestures, activities etc. One important aspect for the analysis and understanding of human-human or human-computer interaction, is to somehow automatically gain knowledge about people's focus of attention, i.e. the knowledge about the targets, objects, or other people with whom they interact. This is for example important to build attentive interfaces [16, 15] and in particular to build attentive robots that should know, when they are addressed and when not [4]. Gaining information about people's focus of attention is also of particular interest for analyzing human-human interaction as in meetings. The analysis of focus of attention dynamics can for instance give relevant information about who is talking to whom [12], the roles of people, their dominance and possibly ranks, the structure of interaction [7], as well as about the type of interaction going on (for example discussion vs. presentation by one person).

In our own research we have been extensively working on analyzing focus of attention particularly in meetings. In the work presented in [10], we used an omnidirectional camera to track faces and head orientations for a fixed number of meeting participants around a table. A probabilistic framework was employed to find each person's most likely focus target, based on the participant's head orientation. The model could already automatically adapt to people's individual head orientation behaviors and to different scenarios (seating, number of people). We also investigated the use of speech activity cues and could show improvements by combining speech cues with head orientation for tracking focus targets.

One limitation of our previous system was, that the meeting participants' head orientations could only be estimated, if everybody was sitting around the table and not too far away from the omnidirectional camera. If a person wanted to stand up and give a presentation for example, his or her head and head orientation could not have been tracked anymore.

In the work we present here, we have overcome this limitation, since we now estimate people's head orientations with several remote cameras that can cover a whole room. In this paper, we present this new multi-view approach for head pose tracking in a smart room. We then also apply the probabilistic focus of attention model, which we already used in previous work, to detect the foci of attention of meeting

participants, in a meeting that we recorded with the remote cameras.

In our experiments we could determine focus of attention of four meeting participants at an average of 70% of the time with automatically learned model parameters (81% with hand-set parameters), based on the multi-view head pose estimates. Concerning head pose estimation, we observed a mean error of 11.2°.

1.1 Related Work

Head pose estimation is a popular task for recognizing peoples' looking direction. In general, approaches can be divided in either model based or appearance based techniques, whereas the first contributions mostly suffer from head movements that lead to the disappearing of important facial features such as nostrils, eye regions or mouth corners. The approaches described in [3, 2] are mostly affected by this constraint. Not only do they require facial images of quite high resolution to detect the relevant features, they also suffer from tracking problems due to fast head movements. In contrast, appearance-based approaches tend to achieve satisfactory results even with lower resolutions of extracted head images but typically lack precision. In [11] a neural-network-based approach was demonstrated for head pose estimation from very low resolution facial images which were captured by a panoramic camera. Here, however, the output only covered ranges from the left to the right profile. Also only one camera view was used, thereby limiting the application of the system to an area around a meeting table.

Another interesting work is described by Ba and Obodez in [1]. They classify facial images by modeling the responses of Gabor and Gaussian filters for a number of pose classes. One contribution of their work is the combination of head detection and pose estimation in one particle filter framework. Unfortunately the authors did not extend their system to more than one camera or evaluate on low resolution head captures.

Pappu et al. present in [8] a textural approach, where synthetically created ellipsoidal texture models of a head are used to determine head pose by matching them with live images. And in [14], Tian et al. describe the use of wide baseline overhead stereo-cameras in a room to classify an observed head pose into one of a fixed set of discrete pose classes. There, also neural networks were implemented for estimating the head pose seen by each single camera. A maximum-likelihood search then results in the final pose hypothesis.

One of the few multi-view approaches, that also targets at estimating head orientation with multiple cameras in order to stabilize the final hypothesis is presented in [18]. Zhang et al. describe a system that uses weak-classifier cascades to detect rotated heads in every camera's capture over a sequence of frames. After three-dimensional hypotheses of the corresponding person's head are generated with the means of triangulation, dynamic programming is used to search for the optimal trajectory of the person's true head in three-dimensional space.

1.2 Paper Overview

The remainder of this paper is as follows: First, in Section 2, we introduce our smart room and the camera setup therein. In Section 3, we describe the main algorithm behind our head pose estimation. First, we introduce our general approach for estimating head pose on single-view head captures and later extend this method by adding a Bayesian filter in order to fuse the single pose estimates into one final joint hypothesis over all available camera views. A brief description of our technique for classifying focus of attention is given in Section 5. We show how we automatically detect and map head pose observations to actual focus targets. Section 6 will present our system's performance on our data collection. We first describe the collected training and testing data and discuss our evaluation. A conclusion and discussion about future work is given in Section 7.

2. SMART ROOM SETUP

We integrated a smart room that is equipped with several sensors in order to gather both audio and visual features about peoples' occupations and activities. Amongst numerous microphones and microphone arrays (both for speaker source localization and far field speech recognition), we installed several cameras to allow visual people tracking, person identification or head pose estimation.

For this work, we used four fixed and calibrated cameras that are placed in the room's upper corners (Figure 1). The cameras do not obtain any zooming abilities. The videos captured provide a resolution of 640×480 pixels at 15 frames per second, hence, concerning where a person is standing in the room, head captures tend to vary strongly in size. We observed head captures to be as small as 20×30 pixels (see Figure 2 for some sample images).

However, the use of four surrounding cameras allows people to move freely and makes sure to always obtain at least one frontal view of each respective face to estimate head pose on. This guarantees to always have at least one estimate with a generally higher confidence than the remaining views.

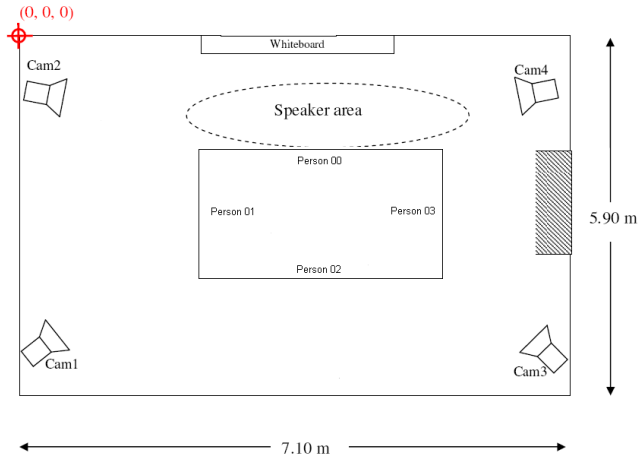

Figure 1. Top-down view into our smart room. We installed four fixed cameras to obtain views covering the whole room. In our meetings, a table was placed into the room, the four participants were to sit around it in front of the whiteboard's area.

One of the main challenges in using such a multi-view approach is to integrate views at the back of peoples' heads. Due to the surrounding setup, at least one or two camera views always depict peoples' hair instead of facial fea-

tures. An automatic single-view estimate of the corresponding head rotation fails due to the chaotic structure of hair. Although, the full integration of views with a rather unreliable, low confidence into a general fusing scheme shall be part of our statistic approach we are going to present here. In our previous work [17], we tried to overcome this problem by integrating a facial view classification step, where the likelihood of a head capture to actually depict a frontal view was estimated and used in the final fusion scheme of our multi-view head pose estimation.

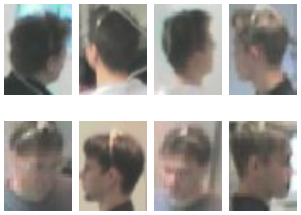

Figure 2. Head captures vary strongly in size, depending where the person is standing in the room and relative to the capturing camera. We observed head bounding boxes as small as 20×30 pixels. As it can be seen, facial details can hardly be detected. Eyes, nostrils and lips, which usually provide good features seem to be almost unrecognizable.

3. HEAD POSE ESTIMATION WITH NEURAL NETWORKS

Our head pose estimating system uses one neural network that is trained to estimate the horizontal head orientation of a person, relative to the observing camera's line of sight. With classifying head pose relative to the camera, our system is easily extensible to adding more cameras to the setup.

As described in Section 2, our head observations are rather small and poor in resolution, hence not allowing reliable detection of eye regions or nostrils. As for example reported in [11], neural networks have already proved good performance for head pose estimation even on rather low resolution images. We believe and show this strong performance is portable to our camera setup, too.

Our implemented neural network follows a three-layered, feed-forward topology, including 100 hidden neurons in the second layer. As input, the cropped head region is downsampled to an image size of 32×32 pixels, grayscaled and linearly stretched in its contrast to overcome small lighting changes. A Sobel operator is applied to get the magnitude response in both horizontal and vertical derivation. Both images are then concatenated to obtain a feature vector of 2048 dimensions which is fed into the network's input layer (as depicted in Figure 3). The feature vector is created by reading out the pixel intensities column by column, row by row.

The network was trained using standard error backpropagation and sigmoid activation functions. A cross evaluation set was used to obtain the best performing network among 100 training cycles.

The network outputs a class-conditional probability distribution $p(c_k|z_j)$ over 36 possible, discrete head pose classes c_k, each 10° wide. In total the network estimates over the whole pose range, from -180° to $+180^\circ$. Here, the observation of camera j is denoted by z_j. We used a Gaussian density function as corresponding target outputs during training, in order to imply a small fuzzification of the very correct class. We did this due to the fact, that with the latter use of estimations from multiple views, the integration of several network outputs into one joint measurement overcomes one single camera's uncertainty.

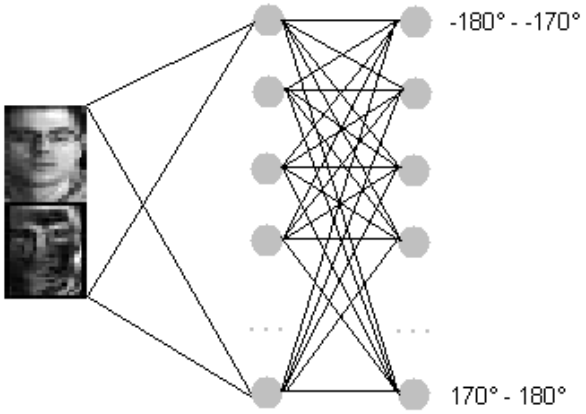

Figure 3. In the multi-view setup, we trained one neural network with 36 output neurons. Each of them represents one discrete head pose class, relative to the camera's line of view (in 10° steps). The network was trained to estimate the class-conditional likelihood of the corresponding output class given the observation of that camera. As input features, we use a grayscale image of the cropped head region along with its Sobel response.

4. A BAYESIAN FILTER FOR MULTI-VIEW FUSION

For integrating the several camera-dependent hypotheses, we use a Bayesian filter approach for fusing the single estimates into one joint estimation. Our state space consists of 360 states $X = \{x_i\}$, with $0 \leq x_i \leq 359$, where each one represents one individual, possible horizontal head rotation with respect to the room's coordinate system. At each timestep t, we compute a probability distribution over all 360 states by applying Bayes' rule, such as

$$p(x_i|Z_t) = \frac{p(Z_t|x_i) \cdot P(x_i)}{p(x_i)} \tag{1}$$

given a joint measurement $p(Z_t|x_i)$ that is derived from the four single cameras' hypotheses with observations $Z_t = \{z_{j,t}\}$. $P(x_i)$ denotes the prior probability to be in state x_i. Each of these factors is going to be described in the following subsections.

4.1 Building a Joint Measurement

By mapping each possible head orientation x_i to an orientation $\phi_j(x_i)$, relative to camera j's line of view, we gather a combined measurement out of all single cameras' hypotheses by averaging the four class-conditional estimations, such that

$$p(Z_t|x_i) = \frac{1}{4} \sum_{j=1}^{4} p(Z_t|\phi_j(x_i)) \tag{2}$$

The intuition behind Equation 2 is that the hypothesis x_i is scored higher, the more cameras agree on it, i.e. the respective output neuron exhibits a high value. That means, if two or more hypotheses strongly agree on the very same head orientation, the final sum of these probabilities returns a much higher value than accumulating smaller likelihoods that describe rather uncertain, ambiguous estimations.

4.2 Integrating Temporal Filtering

Temporal information is implied by the prior probability distribution $P(x_i)$ within Equation 1. At each timestep t this factor implies the probability to observe state x_i. This factor is derived from the transition probability $p(x_i|x')$ to change into state x_i and the a-posteriori probability distribution $p(x'|Z_{t-1})$ which was computed at time $t-1$:

$$P(x_i) = \sum_{x' \in X} p(x_i|x')p(x'|Z_{t-1}) \tag{3}$$

We applied a Gaussian kernel function to provide state change propagation $p(x_i|x')$, hence updating the prior distribution can be defined as a convolution of the Gaussian kernel and the previous a-posteriori likelihoods:

$$P(x_i) = \sum_{x' \in X} N_{0;\sigma}(x_i - x')p(x'|Z_{t-1}) \tag{4}$$

In our evaluation we experimentally used a standard deviation $\sigma = 20°$.

By using a Gaussian kernel, short-term transitions between neighboring states are more likely than sudden jumps over a bigger range of states, hence the adaptation of the kernel's width directly influences, how strong temporal filtering and smoothing of the system's final output takes place.

5. USING HEAD POSE FOR FOCUS OF ATTENTION TRACKING

It is well known that a person's head orientation – in addition to eye gaze and body posture, for example – is an important cue to derive a person's focus of attention [9, 5]. In particular during social interaction, head orientations seem to play an important role, since *most lookers in effect cooperate by making head movements, or other special expressive movements accompanying shifts of gaze*, as Argyle constitutes in [6].

Our approach therefore is to derive likely focus targets based on an individual's observed head orientation. Instead of directly classifying specific focus targets from head images (without the need of head orientation or gaze in general), we aim for an approach that allows for a flexible setup and number of focus targets. In addition, we aim for modeling the subjects' different head orientation behaviors, since there is evidence that people use individual head and eye orientation behavior when interacting with surrounding humans or objects (see for example [13]).

Thus, given an estimated head pose observation x, we want to compute the probability of a possible focus target w_i. Following Bayes' rule, this can be described by the following equation:

$$p(w_i|x) = \frac{p(x|w_i)P(w_i)}{\sum_{j=1}^{n} P(w_j)p(x|w_j)} \tag{5}$$

The most likely focus can then be derived by choosing the target with highest probability, i.e.

$$w_i = \arg\max_i p(w_i|x) \tag{6}$$

The difficulty resides in approximating each participant's individual head orientation style, that is the class-conditional distribution $p(x|w_i)$. For every person, an individual model has to be adapted. Our solution uses Gaussian mixture models to approximate each person's observed head orientations. Not only does this allow for an automatic clustering of head orientations, but also implies both class-conditional distributions and a-priori likelihoods for every automatically detected target. Every component's mean and variance is initialized with k-Means algorithm. The mixture model's final parameters (mean vector, covariance matrix and prior likelihoods) of its components are obtained using EM algorithm. We use as many Gaussian components as there are relevant targets in the scene. As a final step, we compute the a-posteriori distribution of every target, given the possible range of allowed head rotations.

Once the Gaussian mixture models have been adapted to an individual's head orientation observations, they can be used to obtain the class-conditional probabilities $p(x|w_i)$ and priors $P(w_i)$ for our Bayesian model (Eq. 5). Here, the individual Gaussian components are taken as the class-conditional distributions and their corresponding weights are taken as the priors in the Bayesian model. The denominator in Equation 5 is of course constant for one observation x and can thus be ignored for Equation 6.

Figure 6.1 depicts one of our trained Gaussian mixture models and the corresponding class-conditionals for all detected targets of one person. It can be seen that the class-conditional distributions, which correspond to looking towards the people sitting left, opposite and right to our subject, are well defined and therefore allow for a quite robust mapping to either one of these focus targets.

6. EXPERIMENTAL EVALUATION

6.1 Data Description

For training the head pose estimating neural network, we collected video recordings of 15 different people in our smart room. Each person was to wear a magnetic motion sensor that allowed us to gather information about the relative position and orientation of the sensor to a stationary transmitter. We recorded all four described camera views at 15 frames per second. The video recordings are all about three minutes in length, each person was to rotate and move his or her head in all possible directions. For every participant, about 10810 frames from all four camera views were collected; in total, the network has been trained with over 162150 head images from 15 different persons.

We further recorded one meeting video with four participants sitting around one table for evaluating our focus of attention estimation system. The video is about six minutes long, also captured with 15 frames per second. Here, person 0 was also to wear the magnetic motion sensor, in order to gather annotations about this participant's true head rotations. The video starts with all people arriving at the table, meeting and greeting each other. Then all participants sit down and begin a discussion for about five minutes. We manually annotated the video, so that at each time frame

Figure 4. Example scene from our captured meeting video.

we obtain knowledge about who was looking at whom. Further, the head centroid was marked in each frame too, so we overcome the problem of tracking and recognizing people during the meeting for the here reported experiments. No further persons were allowed to enter or leave the room or the scene in general.

6.2 Evaluating Head Pose Estimation

Due to the use of the magnetic motion sensor, we were able to evaluate our head pose estimation for person0 who was wearing the device. Table 1 depicts the results for this person.

Video	$error \leq 10^\circ$	$error \leq 20^\circ$	mean error
Correct	54.5%	87.3%	11.2°

Table 1. Analysis of head pose estimation quality on person0, for whom the ground truth was available (a magnetic pose sensor was used).

As it can be seen, the mean average error for this person was 11.2°, which seems to be quite well, given the difficulty of the task. In about 87.3% of all frames, we estimated pose correctly with a mean error $\leq 20^\circ$.

6.3 Evaluation of Focus of Attention Mapping

We compared our results of focus of attention recognition with our annotated focus targets for each person. Table 2 shows our results. As it can be seen, our system achieved a correct classification of 70% in average for the four participants. The best result (87% correct) was obtained for person 1. This result does not surprise at all; watching the video, this person seems to include head movement a lot for focusing discussion partners. It seems eye movement for focusing discussion partners or a rather passive participation during the meeting were consciously avoided - being contrary to the remaining participants. For person 3 our approach only achieved a correct classification of 53.3% of all frames. Here, head rotation was often ambiguous. Subjectively drawing conclusions from watching the video, this person often stopped his head rotation in between two opposite sitting participants and focused either one with eye gaze. Further, the analysis showed that this participant rather showed a passive discussion style, mostly sitting calmly and following the discussion without much movement and active contribution.

	P0	P1	P2	P3	Avg.
Correct	75.8%	87%	62.1%	53.3%	70.3%

Table 2. Classification results with automatically learned model parameters on visual head pose estimations.

6.4 Manually Setting the Parameters

In addition to only evaluating the fully-automatic mapping using Gaussian mixture models, we also manually adjusted the model parameters of our focus of attention model (see Equation 6). This was done by supervised analysis of each individual's head pose distributions for the different targets. By manually choosing these optimal parameters we can obtain an upper limit of the possible accuracy of the approach, given the noisy head orientation estimates.

As it can be seen in Table 3, at an average of 81% of all frames, head pose could be well used to find the corresponding focus target for all participants.

	P0	P1	P2	P3	Avg.
Correct	83.3%	88.4%	77.8%	72.8%	80.9%

Table 3. Classification results with manually set model parameters on visual head pose estimations.

7. CONCLUSION

In this work we presented a system for estimating horizontal head orientations from multiple camera views. This approach allows to estimate people's head pose in a 360° range and is able to track head pose in the whole smart room.

Based on head pose observations, we implemented a focus of attention classifier to recognize who was looking at whom in a recorded meeting. We evaluated our system experimentally on one recorded meeting video. The results indicate, that we can detect the participants' focus of attention at an average of up to 81% of the time.

This work extends our previous work on tracking head pose and focus of attention in meetings, in that the here presented multi-view approach now allows the participants to move freely in a smart room, and does not restrict them to stay close to a camera on the meeting table.

7.1 Future Work

While the presented multi-view pose tracking approach now allows us to track head orientations of people moving freely in a smart room, our current model to map head orientations to likely focus targets has not yet been extended to handle dynamic scenes. This is one of the most important issues that we want to address in the future. In addition, we hope to get further improvements, by also incorporating information about people's body orientation, as well as information about the current speakers and their position in a room.

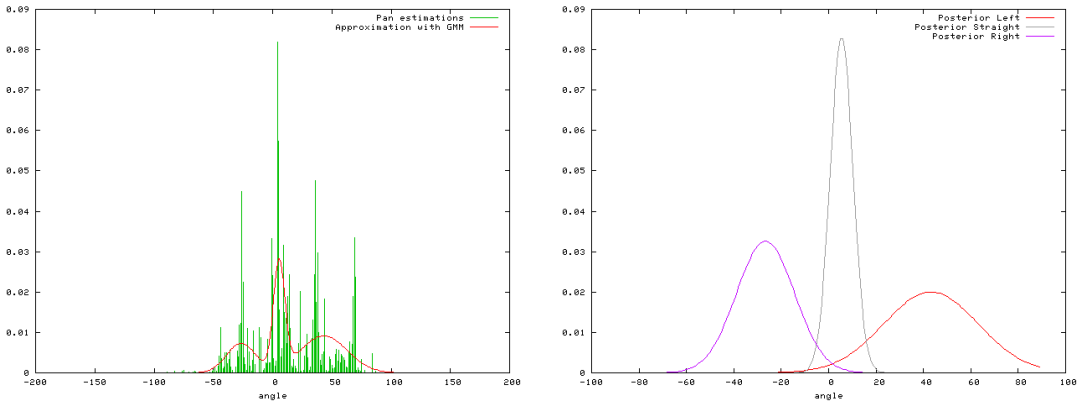

Figure 5. Example of a Gaussian mixture model approximating the observed head pose observations. There are three Gaussian components describing the three possible focus targets relative to the tracked person. The right Figure depicts the approximated class-conditional distributions for every possible target. It is obvious how every target resides within its individual head pose range.

8. ACKNOWLEDGEMENT

This research was supported by the German Research Foundation (DFG) within SFB 588 Humanoid Robots and the European Commission under contract no. 506909 within the project CHIL (Computers in the Human Interaction Loop; http://chil.server.de).

9. REFERENCES

[1] S. O. Ba and J.-M. Obodez. A probabilistic framework for joint head tracking and pose estimation. In *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.

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

[3] T. Horprasert, Y. Yacoob, and L. S. Davis. Computing 3-d head orientation from a monocular image sequence. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, 1996.

[4] M. Katzenmaier, R. Stiefelhagen, T. Schultz, I. Rogina, and A. Waibel. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *International Conference on Multimodal Interfaces ICMI*, 2004.

[5] S. R. Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. In *The Quaterly Journal of Experimental Psychology, 53A(3):825845*, 2000.

[6] M. C. Michael Argyle. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[7] K. Otsuka, Y. Takemae, J. Yamamoto, and H. Murase. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions and utterances. In *Proceedings of the International Conference on Multimodal Interfaces - ICMI*, 2005.

[8] R. Pappu and P. Beardsley. A qualitative approach to classifying gaze direction. In *Proceedings of FG98*, pages 160–165, 1998.

[9] V. B. Stephen R.H. Langton, Roger J. Watt. Do the eyes have it? cues to the direction of social attention. In *Trends in Cognitive Neuroscience, 4(2):5058*, 2000.

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

[11] 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.

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

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

[14] Y.-L. Tian, L. Brown, J. Connell, S. Pankanti, A. Hampapur, A. Senior, and R. Bolle. Absolute head pose estimation from overhead wide-angle cameras. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

[15] K. van Turhnout, J. Terken, I. Bakx, and B. Eggen. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of the International Conference on Multimodal Interfaces - ICMI*, 2005.

[16] R. Vertegaal. Attentive user interfaces. *Communications of the ACM*, 46(3), 2003.

[17] M. Voit, K. Nickel, and R. Stiefelhagen. Multi-view head pose estimation using neural networks. In *Second Workshop on Face Processing in Video (FPiV'05), in Proceedings of Second Canadian Conference on Computer and Robot Vision. (CRV'05), 9-11 May 2005, Victoria, BC, Canada*, 2005.

[18] M. L. Z. Zhang, G. Potamianos and T. Huang. Robust multi-view multi-camera face detection inside smart rooms using spatio-temporal dynamic programming. In *Proceedings of Automatic Face and Gesture Recognition (FG), Southampton, United Kingdom*, 2006.