Tracking Focus of Attention for Human-Robot Communication

Rainer Stiefelhagen[1], **Jie Yang**[2], **Alex Waibel**[1,2]

Interactive Systems Laboratories
[1]University of Karlsruhe, Germany
[2]Carnegie Mellon University, USA
stiefel@ira.uka.de, yang+@cs.cmu.edu, ahw@cs.cmu.edu

Abstract

In an intelligent working space, social robots should be capable of detecting and understanding human communicative cues. An important cue in human communication is focus of attention expressed by gaze direction. We have been developing technologies for gaze tracking and focus of attention modeling. In this paper we present our work on modeling focus of attention in meeting situations. We employ neural networks to estimate a persons head pose from camera images, and a probabilistic model to identify interesting targets in the scene based on the observed head pose. We are extending such technologies in building a gaze-aware human-friendly robot that is able to monitor a person's focus of attention.

1 Introduction

Recent development in humanoid robotics poses new challenges to both robotics and HCI communities. A major challenge is to develop robots that can behave like and interact with humans. Much research has been directed to advancing human-robot interaction [11, 2, 10, 1, 20]. In this research, we are interested in making a robot track human gaze and focus of attention.

Gaze plays an important role in human social interaction. During face-to-face communication people look at each other, monitor each other's lip-movements and facial expressions, and follow each other's gaze. In an intelligent working space, where humans and robots may interact with each other, gaze information could be used to detect what a person is looking at and paying attention to, to interpret what object or place a person is referring to when talking with a robot, or to determine whether a person is talking to the robot or not.

In the past few years, we have developed technologies for tracking and modeling human focus of attention. We are currently addressing 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 [16].

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 recorded evaluation meetings.

In the recently started humanoid robot project sponsored by the German Government, we have started to work on adapting and extending to build a gaze-aware human-friendly robot which is able to monitor a person's focus of attention. We describe the system and discuss potential extensions. The remainder of this paper is organized as follows: In Section 2 we introduce an approach to estimate a person's head pose from facial images using neural networks. In Section 3 we describe our system to simultaneously track participants in a meeting and estimate their head poses. In Section 4 we introduce a probabilistic approach to determine at which target a person is looking at based on his head pose and present experimental results on several meetings. In Section 5 we discuss how focus of attention tracking can be used to enhance human-robot communication and describe a demonstration system we built to illustrate the feasibility of our approach. We'll also discuss limitations of the current system and our future research directions.

2 Estimating Head Pose Using Neural Nets

In this section we describe how we have designed and trained a neural network to estimate a person's head pan and tilt from facial images.

The main advantage of using neural networks to estimate head pose as compared to using a model

Proceedings of the IEEE-RAS International Conference on Humanoid Robots

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

based approach is its robustness: With model based approaches to head pose estimation [8, 15, 9], head pose is computed by finding correspondences between facial landmarks points (such as eyes, nostrils, lip corners) in the image and their respective locations in a head model. Therefore these approaches rely on tracking a minimum number of facial landmark points in the image correctly, which is a difficult task and is likely to fail. On the other hand, the neural network-based approach does not require tracking detailed facial features. Instead, the whole facial region is used for estimating the user's head pose.

In our approach we are using neural networks to estimate pan and tilt of a person's head, given automatically extracted and preprocessed facial images as input to the neural net. This approach is similar to the approach described by Schiele et. al. [13]. However, Schiele et. al.'s system estimated only head rotation in pan direction. In this research we use neural network to estimate head rotation in both pan and tilt directions. In addition, we have studied two different image preprocessing approaches. Rae et. al. [12] describe a user dependent neural network based system to estimate head pan and tilt of a person. In their approach, color segmentation, ellipse fitting, and Gabor-filtering on a segmented face are used for preprocessing. They reported an average accuracy of 9 degrees for pan and 7 degrees for tilt for one user with a user dependent system.

We have trained neural networks to estimate a person's head rotation from two kinds of camera images: 1) images from a pan-tilt-zoom camera Canon VC-C1) and 2) an omnidirectional camera. The main difference between the two kind of images is the much poorer resolution of facial images obtained from the omnidirectional camera. However, while the pan-tilt-zoom camera is well suited to track the face of one user and therefore is suitable for human-robot interaction tasks, the omnidirectional camera has the advantage that all participants sitting around a table can be tracked simultaneously in one camera view and therefore is well suited for to simultaneous gaze tracking in meetings.

In the remainder of this section we will describe details how we trained neural nets to estimate head pan and tilt from good resolution images. Since preprocessing, training and network architecture are the same for pose estimation from the lower resolution images, we will present the results obtained with those images in section 3.

2.1 Training Networks with good resolution images

We collected training data from 14 persons in our lab. During data collection, users had to wear a head band with a sensor of a Polhemus pose tracker attached to it. Using the pose tracker, the head pose with respect to a magnetic transmitter could be collected in real-time. A camera was positioned approximately 1.5 meters in front of the users head. The user was asked to randomly look around in the room and the images together with the pose sensor readings were recorded. Figure 2 shows two sample images from that were taken during data collection with the Canon VC-C1 camera.

Figure 2: Two good resolution images taken with a pan-tilt-zoom camera during data collection.

2.2 Preprocessing of Images

To locate and extract the faces from the collected images, we use a statistical skin color model [19]. The largest skin colored region in the input image is selected as the face.

We have 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 towards normalizing against different lighting conditions. No additional feature extraction is performed. The normalized grayscale images are downsampled to a fixed size of 20x30 pixels and are then 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 downsampled to 20x30 pixels and are both used as input to the neural nets. Figure 5 shows the corresponding preprocessed facial images of a user. From left to right, the normalized grayscale image, the horizontal and vertical edge images of a user's face are depicted.

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

2.3 Neural Net Architecture, Training and Results

We have trained separate nets to estimate pan and tilt of the head. We have used a multilayer perceptron architecture with one output units, one hidden layer with 20 to 100 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 backpropagation.

To train a multi-user neural network, the data set of 12 users was divided into a training set consisting of 4.750 images, a cross-evaluation set of 600 images and a test set with a size of 600 images.

To determine how well the neural net based system can generalize to new users, we have also evaluated the performance of the neural network on the two remaining users whose images have not been in the training set.

Table 2 shows the results that we obtained on the the multi-user test set and on the new users using the different preprocessing approaches. Each cell of the table indicates the mean difference between the true pan (tilt) and the estimated pan (tilt) over the whole test set. Results are given in degrees.

preprocessing	$multi - user$	$newusers$
histogram	3.8 / 3.0	9.4 / 10.9
edges	4.6 / 3.6	10.1 / 9.9
histo + edges	**3.5 / 2.8**	**7.5 / 8.9**

Table 1: Head pose estimation accuracy from good resolution images on a multi-user test set and on two new users. Results for three different preprocessing methods are indicated: 1) using histogram-normalized images as input, 2) using edge images as input and 3) using both, histogram-normalized and edge images as input. The results indicate the mean error in degrees for pan/tilt.

It can be seen, that the best results were obtained when using both, the histogram normalized images and the edge images as input to the neural networks. On the multi-user test set a mean error of 3.5 degrees for pan an 2.8 degrees for tilt was obtained. On new users the mean error was 7.5 degrees for pan and 8.9 degrees for tilt.

3 Simultaneous Tracking of Head Poses in a Panoramic View

One focus of our research is to develop a multimedia meeting browser to automatically transcribe and summarize meetings [18, 3]. Within this project we also aim to monitor at whom or what participants are looking during the meeting. This information is useful to get a better understanding of the meetings and can later be used for image retrieval.

To simultaneously 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 [16], this has the advantage that only one video-stream has to be recorded, which eliminates the need for camera calibration, synchronization and camera control such as zooming on different participants.

From the view of the camera, a panoramic view of the whole scene can be computed, 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; see [17] for more details. To detect and track faces in the panoramic image, we use a statistical skin color detector and some movement information [19]. In addition, some heuristics are used to distinguish hands from faces; see [16] for details.

Once a face is found in the panoramic view, a perspective view of the person is computed, and the face is again detected in the perspective view using the face detector. Perspective views of two participants are shown in Figure 4. The automatically detected faces are marked with boxes. Faces extracted from such perspective views are later used to estimate each participant's head pose with neural nets.

Figure 4: Perspective Views of two participants.

3.1 Neural Networks: Training and Results

Similar to the approach presented in section ??, we have trained networks to estimate pan and tilt from facial images that were taken with the omnidirectional camera.

Here, again data from 14 users was collected. During data collection, the user was automatically tracked in the panoramic view and a perspective view of the user was generated; see Figure 4. During data collection users had to wear a head band with a Polhemus pose tracker sensor attached to it to determine true head pose for training and evaluation of the networks.

Figure 5: Preprocessed images: normalized grayscale image, horizontal and vertical edge image (from left to right)

Neural networks were trained to estimate head pan and tilt from the preprocessed facial images, as described in section 2.1. As for the higher resolution images from the pan-tilt-zoom camera, the best result was obtained using histogram-normalized and edge images as input to the neural net. Figure 5 shows the preprocessed images for a face extracted from the panoramic view.

On a multi-user test set containing images from twelve users a mean error of 7.9 degrees for pan and 5.6 degrees for tilt was obtained. On a test set containing two new users, a mean error of 9.9 degrees for pan and 10.3 degrees for tilt was obtained.

In order to obtain additional training data, we furthermore 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. 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. This is only slightly worse than the accuracy on new users obtained with good resolution images – 7.5 degrees for pan and 8.9 degrees for tilt – reported in section 2.1.

Table 2 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 2: Head pose estimation accuracy from facial images taken with an omnidirectional camera. Results on a multi-user test set and on two new users for three different preprocessing methods are indicated: 1) using histogram-normalized images as input, 2) using edge images as input and 3) using both, histogram-normalized and edge images as input. The results indicate the mean error in degrees for pan/tilt.

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. Hence, the first step in determining a person's focus of attention is to track his/her gaze.

To map the person's gaze 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 gaze and 3) map each estimated gaze to its likely targets using a probabilistic framework.

4.1 A Probabilistic Model of 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.

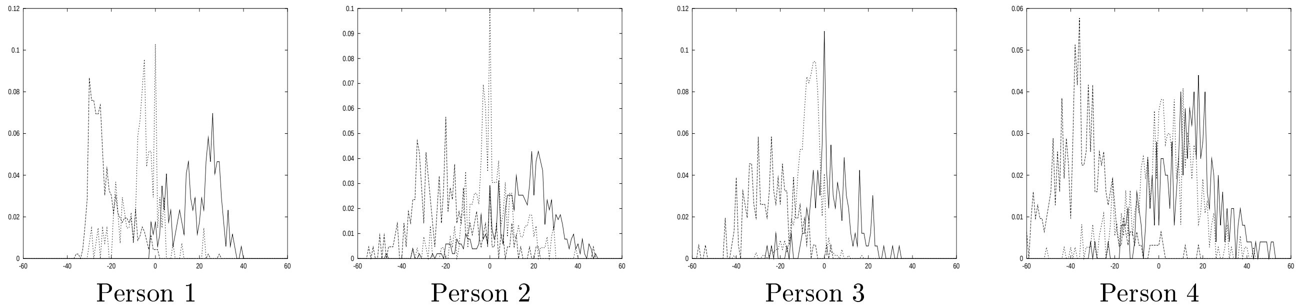


Figure 6: Head pan distributions of four persons in a meeting.

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 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 6 shows the head pan distributions of four participants in one of our recorded meetings. The head rotation of the user was estimated with the neural nets. It can be seen, for example, for Person 1, the three class-conditionals are well separated, whereas for Person 3 or Person 4, the peaks of some distributions are much closer to each other, and 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 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

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

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 Unsupervised adaptation of model parameters

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

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

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

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 [4]:

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

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

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

To initialize the means μ_j of the mixture model, k-means 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 classconditonals $p(x|\text{Focus} = T)$, 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 7 shows an example of the adaptation on pan observations from one user. In Figure 7a) the distribution of all head pan observations of the user is depicted together with the Gaussian mixture that was adapted as described above. Figure 7b) 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 7a). 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 7c) depicts the posterior probability distribution resulting from the adapted class-conditionals and class priors.

4.3 Experimental Results

We have evaluated this approach on three meetings that we recorded in our lab. 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. To be able to evaluate our approach, we manually labeled for each of the participants at whom he was looking in each frame.

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 choosen 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 3 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 3: Percentage of correct assigned focus targets based on computing $P(\text{Focus}|\text{head pan})$.

5 Application to Human-Robot Communication

Many tasks require that a robot can track human's focus of attention when it interacts socially with humans. For example, a robot needs to identify message targets of a person in an intelligent space; i.e., a robot has to identify to whom a human is talking. We have applied focus of attention tracking technology to human-robot interaction. We have built a prototype system to demonstrate gaze-aware interaction with a household robot and other smart appliances such as a speech-controlled VCR in a room. The robot can differentiate if a human is talking to it or other devices by tracking the human's focus of attention. The system consists of the following main components:

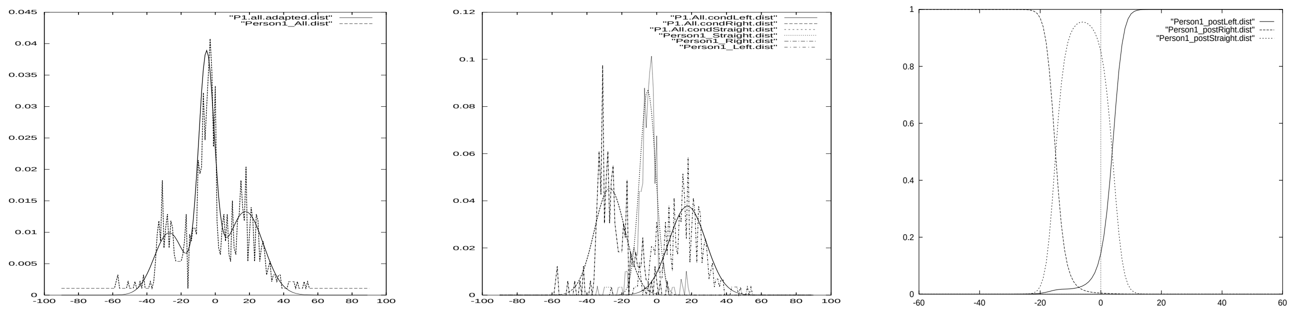
Robot Visualization For the demonstration we have simulated a robot using a 3D visualization toolkit and projected the robot onto one of the walls of our lab.

Speech Recognition A speaker independent large-vocabulary continuous speech recognizer was used for understanding the users' commands [14].

Parser A parser based on the system described in [7] was used to analyze the hypothesis received from the speech recognition module and to generate action commands that were sent to the robot visualization module.

Dialog Manager This module enabled the virtual robot to lead simple clarification dialogues, if necessary information is missing.

Speech Synthesis A Speech synthesis system [5] is used to provide spoken feedback to the user.



a) b) c)

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

Focus of Attention Tracker To observe the user's focus of attention, a pan-tilt-zoom camera was placed next to the simulated robot. The face of the user was tracked in the camera image and the user's head pose was estimated with a neural net as described in section 2.1.

Communication of all the components - recording, speech recognizer, parser, dialogue manager, visualization and focus-of-attention-tracker - was done using a client-server architecture that we adapted from [6].

For the demonstration three focus-targets were chosen: a) the robot, b) the VCR and c) none of them. Whenever the user was looking towards where the VCR was placed, the focus of attention module identified the VCR as target and the output of the speech recognizer was sent to the the VCR. Whenever the user was looking towards the simulated robot, the robot was chosen as the focus target, and therefore recognized speech was directed to the robot; i.e., the robot's parser, dialogue and visualization module, to generate appropriate actions of the robot simulation. Whenever the user was neither looking at the VCR nor to the robot, the user's speech was not recorded at all and neither the robot nor the VCR were responding.

Discussion

While this demonstration showed how gaze-awareness can enhance human-robot communication, there are obviously many limitations with our current approach. First, within the current system no 3D information of the scene is used. The system only works for a user standing within a certain region in front of an immobile robot and probabilistically maps the users head rotation to some fixed focus targets.

A less restrictive system should allow both a user and a robot to move freely in a room. In such a case, the position of the robot and the user has to be determined in the scene and a 3D model of the scene and the objects in it would be necessary to map the user's gaze direction onto one of the interesting objects.

In the presented demonstration, a user's gaze is only used to determine the current addressee of the user's speech. However, gaze could also be used during multimodal communication to determine to what object or place a person is referring to ("Put that there!"). We will address these issues in our future research.

6 Conclusion

In this paper we have described our work on estimating a person's focus of attention. We have presented a system to estimate visual 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 head poses using an unsupervised learning approach. These distributions are then used to predict focus of attention given a head pose. The focus of attention tracking technology can be used for a social robot. In an intelligent working space, a robot has to monitor shared attention of collaborators. We have developed a prototype system to demonstrate the concept of a gaze-aware robot. In a new humanoid robot project sponsored by the German Government, we are adapting and extending focus of attention tracking technology to allow a robot to monitor a person's focus of attention in an unconstrained environment.

Acknowledgments

Thanks to Thomas Schaaf for training the speech recognizer for the gaze-aware robot demo and to Christian Fügen for adapting the parser and dialogue component and hooking all components of the demo together.

This research is partially supported by the German Research Foundation (DFG) as part of the Sonderforschungsbereich 588 and by the US 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 NSF, or any other party.

References

[1] B. Adams, C. Breazeal, B. Brooks, and B. Scassellati. Humanoid robots: a new kind of tool. *IEEE Intelligent Systems*, 15(4):25–31, 2000.

[2] A. Agah. Human interactions with intelligent systems: research taxonomy. *Computers and Electrical Engineering*, 27(1):71–107, 2001.

[3] M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel. Multimodal meeting tracker. In *RIAO 2000 : Content-Based Multimedia Information Access*, Paris, France, April 2000.

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

[5] A. Black and P. Taylor. The festival speech synthesis system: system documentation. Technical report, Human Communication Research Center, University of Edinburgh, UK, 1997.

[6] C. Fügen, M. Westphal, M. Schneider, T. Schultz, and A. Waibel. Lingwear: A mobil tourist information system. In *Proceedings of the of the First International Conference on Human Language Technology (HLT 2001)*, San Diego, March 2001.

[7] M. Gavalda. Soup: A parser for real-world spontaneous speech. In *Proceedings of the 6th International Workshop on Parsing Technologies, IWPT-2000*, Trento, Italy, February 2000.

[8] 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.

[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] A. Koku, A. Sekmen, and A. Alford. Towards socially acceptable robots. In *Proceedings of 2000 IEEE International Conference on Systems, Man and Cybernetics*, pages 894–899, 2000.

[11] D. Perzanowski, A. Schultz, W. Adams, E. Marsh, and M. Bugajska. Building a multimodal human-robot interface. *IEEE Intelligent Systems*, 16(1):16–21, 2001.

[12] R. Rae and H. J. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Transactions on neural networks*, 9(2):257–265, March 1998.

[13] B. Schiele and A. Waibel. Gaze tracking based on face-color. In *International Workshop on Automatic Face- and Gesture-Recognition*, pages 344–348, 1995.

[14] H. Soltau, T. Schaaf, F. Metze, and A. Waibel. The isl evaluation system for verbmobil - ii. In *ICASSP 2001*, Salt Lake City, May 2001.

[15] 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.

[16] 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.

[17] 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.

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

[20] Y.Matsusaka, T.Tojo, S.Kubota, K.Furukawa, D.Tamiya, K.Hayata, Y.Nakano, and T.Kobayashi. Multi-person conversation via multi-modal interface -a robot who communicate with multi-user-. In *Proc. Eurospeech 99*, volume 4, pages 1723–1726, Sep. 1999.