Towards Tracking Interaction Between People

Rainer Stiefelhagen, Jie Yang, Alex Waibel
Interactive Systems Laboratories
University of Karlsruhe — Germany, Carnegie Mellon University — USA
stiefel@ira.uka.de, yang+@cs.cmu.edu, ahw@cs.cmu.edu

Abstract

During face-to-face communication people make use of both verbal and visual behaviors. In this paper we address the problem of tracking visual cues between people. We propose a hybrid approach to tracking who is looking at whom during a discussion or meeting situation. A neural network and a model based gaze tracker are combined to track gaze directions of participants in a meeting. The neural network serves as two functions. First, the neural network coarsely detects gaze direction of a person, i.e., determines if the person is looking at front, or left, or right, or down at the table. Second, the neural network initializes a model based gaze tracker to find out more precise gaze information when a person is in a near front view. The feasibility of the proposed approach has been demonstrated by experiments. The trained neural network has achieved classification accuracy between 82% and 97% for different people. The experimental results have shown significant improvement of robustness for the model based gaze tracker initialized by the neural network.

Introduction

In face-to-face communication, people take advantage of various verbal and visual behaviors [Whittaker and O'Connail 1997]. For example, people use gestures, look at each other, and monitor each others facial expressions when they talk to each other. Therefore, both verbal and non-verbal cues play important roles in human communications. In order to automatically understand such communications, a system not only has to analyze the verbal content of a discussion, but also has to keep track of visual cues such as gestures, gaze and facial expressions of the participants.

Human gaze can serve several functions in a conversation [Whittaker and O'Connail 1997]:

- giving cues of people interest and attention;
- facilitating turn-taking during conversations;
- giving reference cues by looking at an object or person;
- indicating interpersonal cues such as friendliness or defensiveness.

Therefore, finding out at whom or where the speaker is looking at during a conversation is helpful to understand whom the speaker is talking to or what he/she is referring to.

In this paper we address the problem of automatically monitoring the gaze of participants in a meeting. We propose a hybrid gaze tracking approach that integrates a neural net based gaze tracker and a model based gaze tracker. The neural network serves as two functions. First, the neural network coarsely detects gaze direction of a person, i.e., determines if the person is looking at front, or left, or right, or down at the table. Second, the neural network initializes a model based gaze tracker to find out more precise gaze information when a person is in a near front view. We have trained several neural networks using the video data taken from a meeting. The trained networks have achieved user independent classification accuracy between 82% and 97%. The results have demonstrated the feasibility of reliably estimating the coarse gaze direction of a person using neural networks.

When a person is in a near front view, it is desirable to find out more precise gaze information. The model-based gaze tracker [Stiefelhagen et al. 1996] can play such a role. However, our experience has indicated that the model-based gaze tracker is sensitive to its initial configuration. We have integrated the neural net based estimation into the model-based gaze tracker. The output of the neural network is used to facilitate the search for the facial features in case of tracking failure by adjusting the search window locations of the facial features according to the estimated head pose. Using this hybrid gaze tracker, the average percent-

age of tracking failure on the recorded image sequences have been reduced from 40% to 25%.

The remainder of this paper is organized as follows. in Section 2 we describe the problem. In Section 3 we introduce the experimental setup that we used to collect the data. In Section 4 we describe the neural network based gaze estimation and gives detailed classification results. Section 5 explains how to combine the neural network based gaze tracker and model based gaze tracking system. Section 6 discusses some research issues and concludes the paper.

Problem Description

Consider a hypothetical scenario in which people are sitting around a conference table to have a meeting as shown in Figure 1. In the meeting, people are talking each other with gestures and facial expressions. The speakers gaze at listeners and visually monitor their environment. Likewise, listeners look at speakers. Listeners monitor speaker's facial expression and gestures, they nod their heads, change their facial expressions and physical postures depending on their interest in and attitude to the speaker's utterance. In this paper, we assume that the system knows the relative positions of participants. The system, therefore, can determine who is looking at whom by a person's gaze direction.

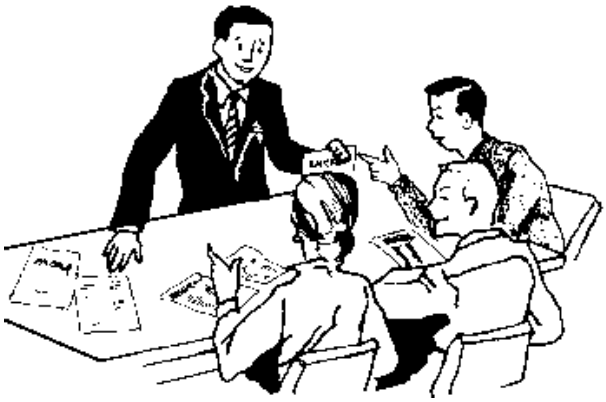

Figure 1: An example of interaction between people in a meeting

A person's gaze direction is determined by two factors: the orientation of the head, and the orientation of the eyes. While the orientation of the head determines the overall direction of the gaze, the orientation of the eyes is determining the exact gaze direction and is limited by the head orientation. Several approaches has been reported to estimate head pose. A neural network based system was used for estimating 2D head orientation [Schiele and Waibel 1995]. A model-based gaze tracker can be used to estimate a person's 3D head pose [Gee and Cipolla 1994, Stiefelhagen et al. 1996]. The system described in [Stiefelhagen et al. 1996] can estimate a person's gaze from images containing the users face with an average accuracy from 5 to 10 degrees. However, in the model based approach, it is essential to track a user's pupils, lip-corners and nostrils in the camera image to estimate his head pose. In order to robustly track these facial features, it requires high sampling rate and high resolution image. Furthermore, the model based system works only for a limited range. When occlusion of facial features happens, the system fails.

In fact, even for a human, we cannot estimate high accurate gaze for all the range. When a person is in a near front view, we can accurately identify his/her gaze. When a person is looking aside, we can no longer identify his/her gaze accurately. Based on this observation, we propose a coarse-to-fine approach to automatically identify person's gaze in a meeting. First, the system coarsely identify the person's gaze in 4 different directions: front, left, right and down. When the person is in a near frontal view, the system further estimates more precise gaze. This idea can be implemented by a hybrid system architecture. A neural network at the top of the system is used to coarsely detect the gaze and a model based gaze tracker is used to further determine more precise gaze if needed.

Experimental Setup

In order to automatically find out who is looking at whom in a meeting, it requires to obtain visual input. This can be achieved by placing cameras on the conference table so that all the participants appear in at least one of the camera views, or by using an omnidirectional camera on the table [Nayar 1997]. Once the participants' faces are in the view of field of the camera(s), the system can automatically locate their faces [Yang and Waibel 1996]. Once the faces are found the system then can try to detect where or at whom a person was looking to at a certain time, given the 3D locations of all the participants.

To develope such a "meeting tracking" system, we have collected data from a restricted meeting situation, where four people were sitting around a table and were asked to talk to each other. In the experiment, we put one camera in the middle of the table and videotaped each of the speakers for a period of time. The participants talked to each other and sometimes looked down at some papers lying in front of them on the table. In order to facilitate the labeling of the video data as well as to have a similiar number of frames of each gaze direction for each user, the person beeing recorded was

asked to look at the others following a script. We have recorded and digitized video sequences of six speakers. Each of the sequences is around two to three minutes long. We then tried to automatically estimate in which direction the persons were looking in the video data. Because the 3D locations of the participants are assumed to be known in the experiment, we can then map the estimated gaze direction directly to the person whom the speaker was looking at (if he wasn't looking down on the table). Figures 2 show some sample images from the recordings.

Figure 2: Sample frames from recorded video

Neural net based gaze estimation

In this section we describe a neural net based approach to estimate the coarse gaze direction of a person, i.e. if the person is looking to the left, the right, down, or straight into the camera. Our approach is similar to the one described in [Schiele and Waibel 1995]. However, in the approach desribed in [Schiele and Waibel 1995] face-color intensified images were used as net input, whereas in this approach the neural net input consists of preprocessed grayscale images of faces.

As neural net architecture we choose a regular multilayer perceptron with one layer of ten hidden units. The input of the neural net consists of preprocessed face images of size 20x30 pixel (see below). Output of the net consists of four units representing the four gaze directions we wanted to estimate (left, right, straight, down). Figure 3 shows the the used architecture of the neural net.

Extracting and preprocessing of faces

To find and extract the faces in these sequences we use a skin-color based approach, that we developped

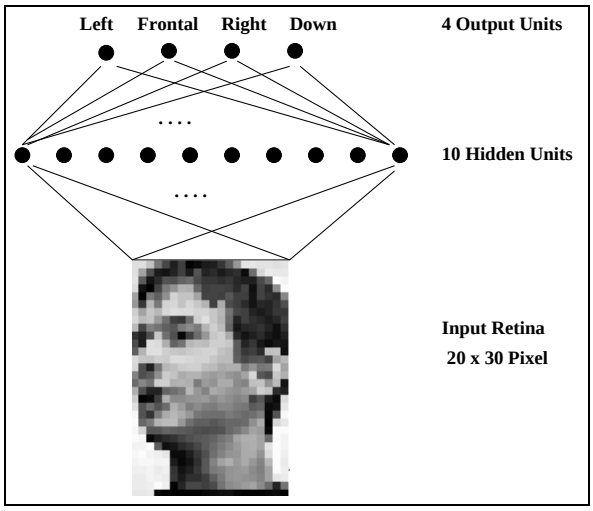

Figure 3: ANN to estimate coarse gaze

for our face tracking system [Yang and Waibel 1996]. In the approach, a two-dimensional Gaussian distribution of chromatic colors is used to model skin color. The largest connected region of skin-colored pixels is considered to be the face in the image.

The extracted faces are resized to a fixed size of 20x30 pixels. In order to compensate for different lighting conditions and to enhance contrast in the images, the resized face images are normalized by histogram normalization. These resized and histogram normalized face images are used as input to the neural net. Figures 4 show some sample input images.

Figure 4: Resized and histogram normalized input images

Training and Results

For training and testing of the net we used images of six different persons. We have trained user independent nets in a round robin way, i. e. we trained one net on images from five users and tested it on the sixth user. This was done for each of the six persons. Training was done using backpropagation with a momentum term. The average number of training samples in the training

(a) (b) (c) (d) (e)

Figure 5: Rapid head movement and occlusion of eyes

set were around 3300 images, average number of frames in the test sets was around 650 images.

We have achieved classification results between 82% and 97% on testset of the different persons. This shows clearly that it is possible to estimate a persons gaze direction using neural nets. Table 1 shows the results for all six persons.

Test Set	Accuracy
Subject 1	85.2 %
Subject 2	97.8 %
Subject 3	82.4 %
Subject 4	92.2 %
Subject 5	97.4 %
Subject 6	86.4 %
Average	90.2 %

Table 1: Neural net gaze classification results

Integrating the neural net with model based gaze estimation

In previous work we have demonstrated that a users gaze can be estimated by a model based gaze tracker. On test sequences we achieved accuracy around 5 to 10 degrees with this system [Stiefelhagen et al. 1996]. In the model based approach the 3D head pose (rotation and translation) is computed using the 2D locations of some facial features in the camera image and their corresponding 3D model locations. Therefore the users pupils, lip-corners and nostrils have to be searched and tracked in the camera image.

The model based gaze tracker sometimes fails to track the facial features. This is mainly due to rapid head movements and facial feature occlusion. Figure 5 is digitized from an image sequence of a person in the meeting. Note that the sampling rate is much lower than that of real-time tracking process because saving the image to a file takes a lot of time. It shows the process that the head of the subject turned from left side to a frontal view and then right side. In Figure 5(a) and (b), only the right eye is visible. In the Figure 5(e), the right eye is occluded. This example shows how easy the model based gaze tracker fails in an unrestricted case.

In case the gaze tracker fails to track the facial features, it starts searching these features in certain search windows inside the face again. To facilitate the recovery of the features, these search windows have to be positioned inside the face according to the actual head pose. For example if the person is looking to the right, the search windows for the eyes should be shifted to the right in the image. Figure 6 shows the search windows for three different head poses.

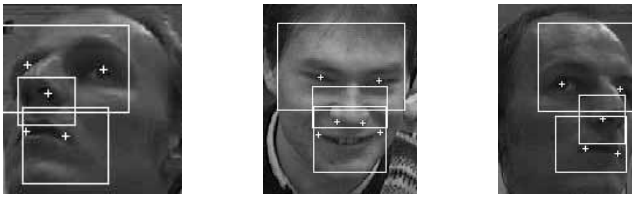

Figure 6: Initialization of search windows according to neural net based pose estimation

In our previous gaze tracker these search windows were adjusted according to the previously found pose. We have now integrated the neural net based gaze estimation into the model based gaze tracker. We now use the gaze estimate of the neural net to initialize the search windows for the facial features. Integrating this approach led to much faster recovery of the gaze tracker and therefore to a reduction of frames with lost features by almost a factor of two. Table 2 shows the percentage of frames were failure of the facial feature tracking occurred with and without the integrated neural net based estimation. In average the percentage of tracking failure on these sequences could be reduced from 40% to 25%.

Conclusion and Future Work

We have proposed a hybrid approach to track gaze directions of participants in a meeting. A neural network based gaze tracker has been developed for detecting coarse gaze directions. It has been demonstrated that

Person	# frames	without NN	with NN
Subject 1	801	41.8 %	29.1 %
Subject 2	791	47.1 %	42.7 %
Subject 3	997	36.8 %	17.8 %
Subject 4	728	33.6 %	20.8 %
Subject 5	646	47.2 %	28.9 %
Subject 6	617	36.9 %	14.3 %
Average		40.6 %	25.6 %

Table 2: percentage of frames, where facial features are lost

a neural network can reliably detect coarse gaze directions of participants in a meeting. The experimental data were recorded from a meeting where four people were sitting around a table and talk to each other. By assuming known relative positions of participants, the system can automatically track who was looking at whom. The measured average accuracy user independent classification is about 90%. We have further combined the neural network based gaze tracker and the model based gaze tracker [Stiefelhagen et al. 1996]. It has been demonstrated that the robustness of the model based gaze tracker has been greatly improved using the neural network based gaze tracker for initialization. The failure rate of the new gaze tracker has been reduced almost 50%.

We are currently working on improving the robustness of our model based gaze tracker to obtain more precise gaze estimation whenever the person's face is in a near frontal view. In order to develop a fully functional system for tracking who is looking at whom in a meeting, several further problems have to be addressed. First, identifications and locations of the participants have to be identified automatically. This requires the system integrates face recognition and 3D tracking modules. Second, sensors and sensor fusion techniques have to be studied carefully. In order to track all the participants simultaneously, multiple cameras have to be mounted on the wall or ceiling in the room and an omnidirectional camera can be placed on the table. These cameras have to coordinated appropriately.

Acknowledgements

We would like to thank some colleagues in Interactive Systems Lab for participating in experiments. The neural networks used in this research were trained using the Stuttgart Neural Net Simulator tool [SNNS].

Support for this work has come from the NSF, under contract CDA-9726363, and the DARPA, under contracts N00014-93-1-0806 and N6601-97-C8553.

References

[SNNS] SNNS: The Stuttgart Neural Net Simulator. University of Stuttgart, Institute of Parallel and Distributed High-Performance Systems. http://www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html.

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

[Nayar 1997] Nayar, S. K. . Catadioptric omnidirectional camera. In Proceedings of Computer Vision and Pattern Recognition, 482-488.

[Schiele and Waibel 1995] Schiele, B. and Waibel, A. Gaze tracking based on face-color. In Proceedings of International Workshop on Automatic Face- and Gesture-Recognition, 344-348. University of Zurich, Department of Computer Science.

[Stiefelhagen et al. 1996] Stiefelhagen, R.; Yang, J. and Waibel, A. A model-based gaze tracking system. In Proceedings of IEEE International Joint Symposia on Intelligence and Systems, 304 - 310. Los Alamitos, CA: IEEE Computer Society Press.

[Whittaker and O'Connail 1997] Whittaker, S. and and O'Connail, B. The role of vision in face-to-face and mediated communication. In Kathleen E. Finn, Abigail J. Sellen and Sylvia B. Wilbur eds.: Video-mediated communication, 23-49. Mahwah, NJ: Lawrence Erlbaum Associates.

[Yang and Waibel 1996] Yang, J. and Waibel, A. A real-time face tracker. In Proceedings of IEEE Workshop on Applications of Computer Vision, 142-147. Princeton, NJ.