From Gaze to Focus of Attention

Rainer Stiefelhagen, Michael Finke, Jie Yang, Alex Waibel
stiefel@ira.uka.de, finkem@cs.cmu.edu, yang+@cs.cmu.edu, ahw@cs.cmu.edu

Interactive Systems Laboratories
University of Karlsruhe — Germany, Carnegie Mellon University — USA

Abstract

Identifying human gaze or eye-movement ultimately serves the purpose of identifying an individual's focus of attention. The knowledge of a person's object of interest helps us effectively communicate with other humans by allowing us to identify our conversants' interests, state of mind, and/or intentions. In this paper we propose to track focus of attention of several participants in a meeting. Attention does not necessarily coincide with gaze, as it is a perceptual variable, as opposed to a physical one (eye or head positioning). Automatic tracking focus of attention is therefore achieved by modeling both, the persons head movements as well as the relative locations of probable targets of interest in a room. Over video sequences taken in a meeting situation, the focus of attention could be identified up to 98% of the time.

1. Introduction

During face-to-face communication such as discussions or meetings, humans not only use verbal means, but also a variety of visual cues for communication. For example, people use gestures; look at each other; and monitor each other's facial expressions during a conversation. In this research we are interested in tracking at whom or what a person is looking during a meeting.

The first step towards this goal is to find out at which direction a person is looking, i.e. his/her gaze. Whereas a person's gaze is determined by his head pose as well as his eye gaze, we only consider head pose as the indicator of the gaze in this paper. Related work on estimating human head pose can be categorized in two approaches: model based and example based approaches: In model-based approaches usually a number of facial features, such as eyes, nostrils, lip-corners, have to be located. Knowing the relative positions of these facial features, the head pose can be computed [2, 8, 3]. Detecting the facial features, however, is a challenging problem and tracking is likely to fail. Example based approaches either use some kind of function approximation technique such as neural networks [1, 7, 6], or a face database [4] to encode example images. Head pose of new images is then estimated using the function approximator, such as the neural networks, or by matching novel images to the examples in the database. With example based approaches usually no facial landmark detection is needed, instead the whole facial image is used for classification. In the Interactive Systems Lab, we have worked on both approaches. We employed purely neural network [7] and model-based approaches to estimate a user's head pose [8]. We also demonstrated that a hybrid approach could enhance robustness of a model based system [9]. In this paper, we extend the neural network approach to estimating the head pose in a more unrestricted situation.

A major contribution of this paper is to use hidden markov model (HMM) to detect a user's focus of attention from an observed sequence of gaze estimates. We are not only interested in which direction a user is looking at during the meeting, but also want to know at whom or what he is looking. This requires a way of incorporating knowledge about the world into the system to interpret the observed data. HMMs can provide an integrated framework for probabilistically interpreting observed signals over time. We have incorporated knowledge about the meeting situation, i.e. the approximate location of participants in the meeting into the HMMs by initializing the states of person dependent HMMs appropriately. We are applying these HMMs to tracking at whom the participants in a meeting are looking. The feasibility of the proposed approach have been evaluated by experimental results. Figure 1 shows an overview of our system: For each user, neural nets are used to produce a sequence of gaze observations ω given the preprocessed facial images $\mathbf{I}$. This sequence of gaze observations is used by the HMM to compute the sequence of foci of attention $\mathbf{F}$ of the user.

The remainder of the paper is organized as follows: Section 2 describes the neural network based head pose estimation approach. In section 3 we introduce the idea of interpreting an observed sequence of gaze directions to find a user's focus of attention in each frame; define the underlying probability model and give experimental results. We summarize the paper in section 4.

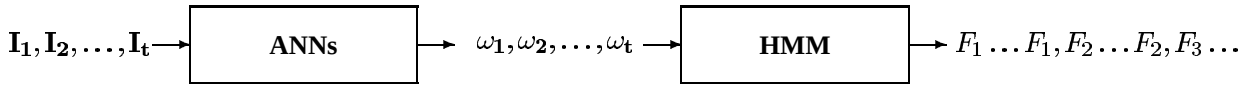


Figure 1. System overview: For each user, neural nets are used to produce a sequence of gaze observations ω given the preprocessed facial images $\mathbf{I}$. This sequence of gaze observations is used by the HMM to compute the sequence of foci of attention $\mathbf{F}$ of the user.

2. Estimating Head Pose with Neural Nets

The main advantage of using neural networks to estimate head pose as compared to using a model based approach is its robustness: With model based approaches to head pose estimation [2, 8, 3], 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 doesn't require tracking detailed facial features because 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. Our approach is similar to the approach as described by Schiele et. al. [7]. However, the system described in [7] 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. [6] describe a user dependent neural network based system to estimate 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 report an average accuracy of 9 degrees for pan and 7 degrees for tilt for one user with a user dependent system.

In the remainder of this section we describe our neural net based approach to estimate user's head pose (pan and tilt). First we desribe how we collected data to train and test the neural networks. Then the two different image preprocessing approaches that we investigated and the neural network architecture are described. Finally we present experimental results that we obtained using different types and combinations of input images for the neural nets.

2.1. Data Collection Setup

During data collection, the person that we collected data from had to sit on a chair on a specific location in the room, with his eyes at a height of approximately 130cm. In a distance of one meter and at a height of one meter a video camera to record the images was placed on a tripod. We placed marks on three walls and the floor on which the user had to look one after another. The marks where placed in such a way that the user had to look in specific well known directions. The marks ranged from -90 degrees to +90 degrees for pan, with one mark each ten degrees, and from +15 degrees to -60 degrees for tilt, with one mark each 15 degrees. This means that during data collection the user had to look at 19 x 6 specific points, from top to bottom and left to right. Once the user was looking at a mark, he could press a mouse-button, and 5 images were being recorded to hard-disk, together with the labels indicating the current head pose. This resulted in a set of 570 images per user. In order to collect slightly different facial images for each pose, the user was asked to speak with the person assisting the data collection. Figure 2 shows two example images recorded during data collection. In this way, we collected data of 14 male and 2 female subjects. Approximately half of the persons were wearing glasses.

Figure 2. Example images take during data collection as used for training and testing of the neural nets

2.2. Preprocessing of Images

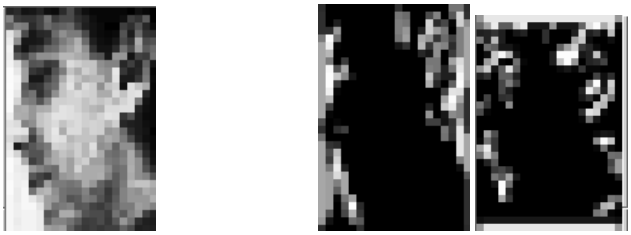
We investigated two different preprocessing approaches: Using normalized grayscale images of the user's face as the input to the neural nets and applying edge detection to the images before feeding them into the nets.To locate and extract the faces from the collected images, we have used a statistical skin color model [10]. The largest skin colored region in the input image was selected as the face.

In the first preprocessing approach, histogram normalization was applied to the grayscale face images as a means towards normalizing against different lighting conditions. No additional feature extraction was performed and the normalized grayscale images were downsampled to a fixed size of 20x30 images and then used as input to the nets.

In the second approach, we applied a horizontal and a vertical edge operator plus tresholding to the facial grayscale images. Then the resulting edge images were downsampled to 20x30 pixels and were both used as input to the neural nets. Figure 3 and 4 show the corresponding preprocessed facial images of the person depicted in Figure 2. From left to right, the normalized grayscale image, the horizontal and vertical edge images are depicted.

Person A

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



Person B

Figure 4. Preprocessed images: normalized grayscale, horizontal edge and vertical edge image (from left to right)

2.3. ANN Architecture

We trained separate nets to estimate pan and tilt of a person's head. Training was done using a multilayer perceptron architecture with one hidden layer and standard backpropagation with momentum term.

The output layer of the net estimating pan consisted of 19 units representing 19 different angles (-90, -80, ...,+80, +90 degrees). The output layer of the tilt estimating net consisted of 6 units representing the tilt angles +15, 0, -15, .. -60 degrees. For both nets we used gaussian output representation. With a gaussian output representation not only the single correct output unit is activated during training, but also its neighbours receive some training activation decreasing with the distance from the correct label. The input retina of the neural nets varied between 20x30 units and 3x20x30 units depending on the different number and types of input images that we used for training (see 2.4).

2.4. Training and Results

We trained separate user independent neural nets to estimate pan and tilt. The neural nets were trained on data from twelve subjects from our database and evaluated on the remaining four other subjects. The data for each user consisted of 570 images, which results in a training set size of 6840 images and a test set size of 2280 images.

As input to the neural nets, we have evaluated three different approaches:

1. Using histogram normalized grayscale images as input to the nets
2. Using horizontal and vertical edge images as input
3. Using both, normalized grayscale plus horizontal and vertical edge images as input.

Table 1 summarizes the results that we obtained using the different types of input images. When using normalized grayscale images as input we obtained a mean error of 12.0 degrees for pan and 13.5 degrees for tilt on our four user test set. With horizontal and vertical edge images as input, a slightly worse accuracy for estimating the pan was obtained. Using both, normalized grayscale image as well as the edge images as input to the neural net significantly increased the accuracy and led to accuracy of 9.0 degrees and 12.9 degrees mean error for pan and tilt respectively.

These results show, that it is indeed feasible to train a person independent neural net based system for head pose estimation. In fact, the obtained results are only slightly worse than results obtained with a user dependent neural net based system as described by Rae et. al.[6]. As compared to their results, we did not observe serious degradation on data from new users. To the contrary, our results indicate that the neural nets can generalize well to new users.

Net Input	Pan	Tilt
Grayscale	12.0	13.5
Edges	14.0	13.5
Edges + Grayscale	9.0	12.9

Table 1. Person independent results (Mean error in degrees) using different preprocessing of input images. Training was done on twelve users, testing on four other users.

However with the system that we have developed so far, we have observed a problem which still limits the use of the system significantly: when we tested the system on previously recorded data from a meeting that took place in another room, the accuracy of the estimation seriously degraded. We believe that this is mainly due to the very different lighting conditions between the room where data collection for training the nets took place (computer lab, no windows), and the room where the meeting took place (daylight + artificial illumination). Figure 5 shows two example images recorded during the meeting.

Possible solutions to this problem might be to investigate other preprocessing methods to reduce the influence of changing illumination and/or collecting more training data under different lighting conditions.

3. Modelling Focus of Attention Using Hidden Markov Models

The idea of this research is to map the observed variable over time namely the gaze direction to discrete states of what the person is looking at, i.e. his focus of attention. Hidden Markov Models (HMM) can provide an integrated framework for probabilistically interpreting observed signals over time. In this section we describe how we have designed the HMMs to estimate a user's focus of attention.

We have incorporated knowledge about the observed scene, i.e. the approximate location of likely foci of attention such as other people in the room, in the Hidden Markov Models. In our model, looking at a certain target is modelled as being in a certain state of the HMM and the observed gaze estimates are considered as being probabilistic functions of the different states. Given this model and an observation sequence of gaze directions, as provided by the neural nets, it is then possible to find the most likely sequence of HMM states that produced the observations. Interpreting being in a certain state as looking at a certain target, it is now possible to estimate a person's focus of attention in each frame. Furthermore, we can iteratively reestimate the parameters of the HMM so as to maximize the likelihood of the observed gaze directions, leading to more accurate estimates of foci of attention.

We have tested our models on image sequences recorded from a meeting. In the meeting, four people were sitting around a table, talking to and looking at each other and sometimes looking onto the table. During this meeting we had taped each of the speakers using a camera standing on top of the table and having one person in its field of view. Figure 5 shows two example images taken during data collection of the meeting. For two of the speakers we then estimated their gaze trajectory with the neural nets described in the previous section. For each user we have applied an HMM to detect his focus of attention given the observed gaze directions over time. We then applied the user-dependent HMMs to detect the foci of attention given the observed gaze directions over time.

In the remainder of this section we describe the design of the HMM, how we have adapted HMM parameters and give evaluation results on the two video sequences.

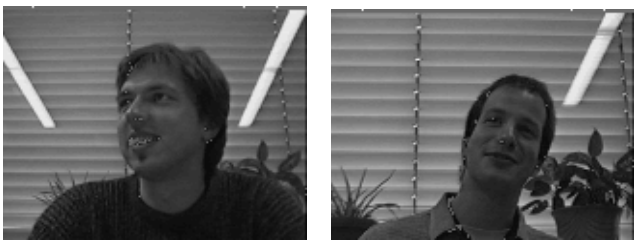


Figure 5. Example images from "meeting" data as used for HMM evaluation

3.1. HMM Design

Knowing that there were four people sitting around a table, we modelled the targets for each person P as the following four states: P is looking to the person sitting to his right, P is looking to the person to his left, P is looking to the person in front of him, P is looking down on the table.

In our model the observable symbols of each state are the pose estimation results as given by the neural nets, that is the angles for pan and tilt ω_{pan} and ω_{tilt}. We have parameterized the state dependent observation probabilities $B = b_i(\omega)$ for each state i, where $i \in left, right, center, table$, as two-dimensional gaussian distributions with diagonal covariance matrices:

$$b_i(\omega) = \frac{1}{2\,\pi\,\sigma_{pan}\,\sigma_{tilt}} e^{-\frac{1}{2}\left[\frac{(\omega_{pan}-\mu_{pan})^2}{\sigma_{pan}^2}+\frac{(\omega_{tilt}-\mu_{tilt})^2}{\sigma_{tilt}^2}\right]}$$

Assuming that we know the approximate positions of the participants of the meeting relative to each other, we initialized the observation probability distributions of the different states with the means of the gaussians set to the expected viewing angle, when looking at the corresponding target. Table 2 shows the initial values that we have chosen for the respective means. All variances were set to the same value initially. The transition matrix $\mathbf{A} = (a_{ij})$ was initialized to

State	μ_{pan}	μ_{tilt}
left	-45	0
center	0	0
right	+45	0
table	0	-45

Table 2. Initializaton of HMM states

have high transition probabilities for remaining in the same state ($a_{ii} = 0.6$) and uniformly distributed state transition probabilities for all other transitions. The initial state distribution was chosen to be uniform.

3.2. Probabilistic Model

Let $O = \omega_1\omega_2\cdots\omega_T$ be the sequence of gaze direction observations $\omega_t = (\omega_{pan,t}, \omega_{tilt,t})$ as predicted by the neural nets. The probability of the observation sequence given the HMM is given by the sum over all possible state sequences q:

$$\begin{aligned} p(\mathbf{O}) &= \sum_q p(O,q) \\ &= \sum_q p(O|q)\, p(q) \\ &= \sum_q \prod_t p(\omega_t|q_t)\, p(q_t|q_{t-1}) \\ &= \sum_q \prod_t b_{q_t}(\omega)\, a_{q_t,q_{t-1}}. \end{aligned}$$

To find the single best state sequence of foci of attention, $q = q_1 \ldots q_n$ for a given observation sequence, we need to find

$$max_q(p(O,q)).$$

This can be efficiently computed by the Viterbi algorithm [5]. Thus, given the HMM and the observation sequence of gaze directions, we can efficiently find the sequence of foci of attention using the Viterbi algorithm.

So far we have considered the HMM to be initialized by knowledge about the setup of the meeting. It is furthermore possible to adapt the model parameters $\lambda = (\mathbf{A}, \mathbf{B})$ of the HMM so as to maximize $p(O|\lambda)$. This can be done in the EM (Expectation-Maximizaton) framework by iteratively computing the most likely state sequence and adapting the model parameters as follows:

- means:

$$\hat{\mu}_{pan}(i) = E_i(\omega_{pan}) = \frac{\sum \phi_{i,t}\, \omega_{pan,t}}{\sum \phi_{i,t}}$$

$$\hat{\mu}_{tilt}(i) = E_i(\omega_{tilt}) = \frac{\sum \phi_{i,t}\, \omega_{tilt,t}}{\sum \phi_{i,t}}$$

, where $\phi_{i,t} = \begin{cases} 1 & : \quad q_t = i \\ 0 & : \quad otherwise \end{cases}$

- variances:

$$\begin{aligned} \sigma^2_{pan}(i) &= E_i(\omega^2_{pan}) - (E_i(\omega_{pan}))^2 \\ \sigma^2_{tilt}(i) &= E_i(\omega^2_{tilt}) - (E_i(\omega_{tilt}))^2 \end{aligned}$$

- transition probabilities:

$$a_{i,j} = \frac{\text{number of transition from state } i \text{ to } j}{\sum_t \phi_{i,t}}$$

On our two evaluation sequences, parameter reestimation converged after three and five iterations respectively.

3.3. Results

To evaluate the performance of the proposed model, we compared the state-sequence given by the Viterbi-decoding to hand-made labels of where the person was looking to. Both of the evaluated sequences contained 500 frames and lasted about one and a half minute each. We evaluated the performance of the HMM without model parameter adaption and with automatic parameter adaption. Furthermore we evaluated the results obtained by directly mapping the output of the neural nets to the different viewing targets. This mapping was obtained by assigning the network output directly to a specific target as described in table 3. Table 4 reports the obtained results. It can be seen that compared to directly using the output of the neural nets, a significant error reduction can already be obtained by using an HMM without parameter adaption on top of the ANN output. Using parameter reestimation however, the error can be furthermore reduced by a factor of two to three on our evaluation sequences.

While performing parameter reestimation on the two sequences, a significant improvement of the accuracy of the adapted HMMs over the HMMs that were initialized using knowledge could be observed. The means of the gaussians, which represent the viewing angles of the different targets, shifted from their initial estimates to values that better matched the observations over time.

ω_{pan}	ω_{tilt}	assigned state
[-90,-30]	any	"left"
[-20,+20]	[-15, +15]	"center"
[+30,+90]	any	"right"
[-20,+20]	[-30, -60]	"table"

Table 3. Direct mapping of ANN output to viewing targets

Seq.	no HMM	HMM, no reest.	HMM, reest.
A	9.4 %	5.4 %	1.8 %
B	11.6 %	8.8 %	3.8 %

Table 4. Percentage of falsely labelled frames without using the HMM and with using HMM before and after parameter reestimation

4. Conclusion

In this paper we have addressed the problem of tracking a person's focus of attention during a meeting situation. We have proposed the use of a HMM framework to detect focus of attention from a trajectory of gaze observations and have evaluated the proposed approach on two video sequences that were taken during a meeting. The obtained results show the feasability of our approach. Compared to hand-made labels, accuracy of 96% and 98% was obtained with the HMM-based estimation of focus of attention.

To estimate a person's gaze we have trained neural networks to estimate head pose from facial images. Using a combination of normalized grayscale images, horizontal and vertical edge images of faces as input to the neural nets, we have obtained accuracy of 9.0 degrees and 12.9 degrees for pan and tilt respectively on a test set of four users which have not been in the training set of the neural nets.

However we observed that under changed lighting conditions the neural network based pose estimaton seriously degraded. Possible solutions to this problem could be using other preproceesing methods to reduce the influence of changing illumination and/or collecting more data under different lighting conditions to train the neural nets.

References

[1] D. Beymer, A. Shashua, and T. Poggio. Example-based image analysis and synthesis. In *Proceedings of Siggraph'94*, 1994.

[2] 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.

[3] 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.

[4] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[5] L. R. Rabiner. *Readings in Speech Recognition*, chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267–295. Morgan Kaufmann, 1989.

[6] 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.

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

[8] 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.

[9] R. Stiefelhagen, J. Yang, and A. Waibel. Towards tracking interaction between people. In *Proceedings of the AAAI Spring Symposium on Intelligent Environments*, pages 123–127. AAAI Press, March 1998.

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