Simultaneous Tracking of Head Poses in a Panoramic View

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

Abstract

In this paper we present an approach to simultaneously estimate gaze directions of multiple people in the view of a panoramic camera. Human faces are located and tracked using a probabilistic skin-color model and motion detection. Neural networks are used to estimate head poses of the detected faces. With this approach, it is possible to simultaneously track the locations of multiple people around a meeting table and estimate their gaze directions using only a panoramic camera. We have achieved an accuracy of 9 degrees for head pan estimation and 6 degrees for tilt estimation for a multi-user system.

1 Introduction

In a multi-party communication situation such as a meeting, knowing the head poses and relative positions of the participants would be useful to determine whom is looking at whom during the meeting. Much previous research has been focused on tracking head pose of one person. Simultaneously tracking head poses and positions of many persons can be achieved using multiple cameras. However, using multiple cameras causes synchronization and calibration problems. Furthermore, the number of cameras might have to be adjusted to the number of persons apparent in the scene. In this paper, we present an approach to track people's head poses and relative position using one panoramic camera.

Related work on estimating human head pose from images 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 [3, 9, 4]. 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 [2, 7, 6], or a face database [5] to encode example images. With example based approaches no facial landmark detection is needed, instead the whole image of the face is used for classification.

In this work we employ neural networks to estimate head pan and tilt from facial images. The work presented in this paper extends our work published in [10], where we used the panoramic camera only to detect the locations of people around the table, but used additional cameras to zoom on each participant's face to get the facial images needed for head pose estimation. The results presented here demonstrate, that it is possible to detect head pan and tilt with low resolution images obtained from a panoramic camera, with only marginal decrease of accuracy.

The remainder of this paper is organized as follows: In section 2 we describe the use of a panoramic camera to detect and track faces around a table. To track each person's gaze we have trained neural nets to estimate head pan and tilt from facial images. Details about architecture, training and results of these nets are given in section 3. We summarize the paper in section 4.

2 Tracking Faces in a Panoramic View

In our system, we use a panoramic camera put on top of the conference table to capture the scene.

Figure 2: The panoramic camera used to capture the scene[1]

[1] Image courtesy of CycloVision Technologies, Inc.

Figure 1: Panoramic view of the scene around the conference table. Faces are automatically detected and tracked (marked with boxes).

Figure 2 shows a picture of the panoramic camera system. The camera is located in the top cylinder and is focusing on a parabolic mirror on the bottom plate. Through this mirror almost a whole hemisphere of the surrounding scene is visible. Figure 3 shows the view of a meeting scene as it is seen in the parabolic mirror and as it is captured with this camera. As the topology of the mirror and the optical system are known, it is possible to compute panoramic views of the scene as well as perspective views at different angles of the panoramic view [1]. Figure 1 shows the rectified panoramic image (with faces marked) of the camera view depicted in Figure 3.

Figure 3: Meeting scene as captured with the panoramic camera

2.1 Using Color and Motion for Face Detection

To detect and track faces in the panoramic camera view, a statistical skin color model consisting of a two-dimensional Gaussian distribution of normalized skin colors is used. The color distribution is initialized so as to find a variety of face colors and is gradually adapted to the faces actually found. The interested reader is referred to [11]. To detect faces, the input image is searched for pixels with skin colors. Connected regions of skin-colored pixels in the camera image are considered as possible faces. Since humans rarely sit perfectly still for a long time, motion detection is used to reject outliers that might be caused due to noise in the image or skin-like objects in the background of the scene that are not faces or hands. Only regions with a response from the color-classifier and some motion during a period of time are considered as faces. In addition some heuristics are applied in order to distinguish (skin-colored) hands from faces.

3 Estimating Head Pose Using Neural Nets

A major 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 [3, 9, 4], 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 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 preprocessed facial images as input to the neural net. This approach is similar to the approach described by Schiele and Waibel [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.

Rae and Ritter [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 our previous work we used one camera to zoom on each of the participants to obtain the input images for pose estimation. Using these high-resolution images, we achieved an accuracy of 8.8 degrees for pan and 5.7 degrees for tilt with a multi-user system. On new users, an accuracy of 10.6 / 9.9 degrees for pan/tilt could be achieved. In the work presented here, we are using only the panoramic camera to generate the input for the neural nets. While the facial images extracted from the panoramic view are of considerably lower resolution than images taken with close up views of each of the participants, we could still obtain accuracy that is only slightly lower as the accuracy achieved with additional cameras to zoom on each person.

3.1 Data Collection

We collected training data from fourteen users. During data collection, the user was automatically tracked in the panoramic view and a perspective view of the user was generated. To determine head pose the 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. The user was asked to randomly look around in the room and the perspective images of the user were recorded together with the pose sensor readings. Figure 4 shows some sample images taken during data collection.

Figure 4: Training Samples: The perspective images are generated from the panoramic view. Head pose labels are collected with a magnetic field pose tracker.

3.2 Preprocessing of Images

To locate and extract the faces from the collected images, we use a statistical skin color model [11]. 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 [8]: 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 down-sampled to a fixed size of 20x30 pixels and then are 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 down-sampled to 20x30 pixels and are both used as input to the neural nets. Since our previous experiments (as described in [8]) showed, that we obtain the best results when combining the histogram normalized and the edge images as input to the neural nets , we are only presenting results using this combination of preprocessed images as input to the neural net here. Figure 5 shows the preprocessed images of a user's faces.

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

3.3 Neural Net Architecture, Training and Results

We have trained separate nets to estimate head pan and tilt. For each net, a multi-layer perceptron architecture with one output unit (for pan or tilt), one hidden layer with 20 to 60 hidden 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 back-propagation.

3.3.1 Results with Multi-User System

To train a multi-user neural network, we divided the data set of 12 users into a training set consisting of 6080 images, a cross-evaluation set of size 760 images and a test set with a size of 760 images. After training, we achieved a mean error of 9.1 degrees for pan and 6.2 degrees for tilt on the test set.

3.3.2 User Independent Results

To determine how well the neural nets can generalize to new users, we have also evaluated the multi-user system on two new users, that have not been in the training set. On the two new users we we obtained an average error of 12.1 degrees for pan and 11.4 degrees for tilt. These results demonstrate, that the neural networks can generalize also to faces of new users.

3.3.3 Adding Artificial Training Data

In order to obtain additional training data, we 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 having the effort of additional data collection. Having more training data should especially be helpful in order to get better generalization on images from new, unseen users. Indeed, after training with the additional data, we achieved an average error of only 10.1 degrees for pan and 10.6 degrees for tilt on the two new users. This corresponds to an error reduction of 17 % for pan and 7% for tilt. Table 1 summarizes the results.

	multi-user	user-independent
basic data	9.1 / 6.2	12.1 / 11.4
+ artificial data	9.3 / 6.4	10.1 / 10.6

Table 1: Average error in degrees (pan/tilt) for multi-user and user-independent systems.

4 Conclusion

We have described a system to automatically detect and track participants around a meeting table and to estimate their head gaze. To capture the images, only one panoramic camera is used. Faces are detected in the view of this camera using a skin-color model and motion detection. We have demonstrated, that is it possible to estimate each participant's gaze using neural networks from images that are generated from the panoramic view. We have achieved an accuracy of 9.1 degrees for pan and 6.2 degrees for tilt for a multi-user system and achieved accuracy of 10.1 degrees (pan) and 10.6 degrees (tilt) on two new users.

Acknowledgments

This Research is partially supported bt the Defense Advanced Research Projects Agency under contract number DAAD17-99-C-0061, and by the 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 DARPA and the NSF, or any other party.

References

[1] S. Baker and S. K. Nayar. A theory of catadioptric image formation. In *Proceedings of the 6th International Conference on Computer Vision , ICCV 98*, pages 35–42, Bombay, India, January 1998.

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

[3] 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.

[4] 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.

[5] 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.

[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, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. In M. Turk, editor, *Proceedings of Workshop on Perceptual User Interfaces: PUI 98*, pages 25–30, San Francisco, CA, November, 4-6th 1998.

[9] 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.

[10] 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.

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