# Estimating Head Pose with Neural Networks - Results on the Pointing04 ICPR Workshop Evaluation Data

Rainer Stiefelhagen
Interactive Systems Laboratories
Universität Karlsruhe (TH)
Germany
E-mail: stiefel@ira.uka.de

## Abstract

*In this paper we report the results of a neural network based approach to head pose esimtation on the evaluation data set provided for the Pointing04 ICPR workshop. In the presented approach, we use neural networks to estimate a person's horizontal and vertical head orientation from facial images, which automatically were extracted from the provided data set.*

*With our approach, we achieved an average estimation error of 9.5 degrees for pan and 9.7 degrees for tilt estimation with a multi-user system that was trained on images from all 15 people in the database..*

## 1 Introduction

In recent years many researchers have addressed the problem of vision-based estimation of a person's head orientation (or head "pose"). Related work can be categorized in two approaches: model based approaches and appearance 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, 9, 3]. Detecting the facial features, however, is a challenging problem and tracking is likely to fail. Appearance based approaches either use some kind of function approximation technique such as neural networks [1, 6, 5], or a face database [4] to encode example images. With appearance based approaches no facial landmark detection is needed, instead the whole image of the face is used for classification.

In the Interactive Systems Lab, we have investigated both approaches. We employed purely neural network [6, 10] and model-based approaches to estimate a user's head pose [9].

In our work we found that robust head pose estimation results could be acchieved using an appearance based approach, where head pose is estimated from facial images using neural networks. This approach has proven to work well on high-resolution facial images as well as low resolution facial images captured with omnidirectional cameras[11, 7].

In this work we report the results of the neural network based approach to head pose esimtation on the evaluation data provided for the Pointing04 ICPR workshop.

## 2 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 [2, 9, 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 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. This also allow for head pose estimation on facial images of low resolution.

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 [6].

However, the system described in [6] 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 [5] 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 have reported head pose estimation results on good resolution facial images, captured with a pan-tilt zoom camera [8, 10] as well as on low resolution images captured with an omnidirectional camera [11, 7]. Figure 1 shows some sample images used in our previous work. On both good resolution images and low resolution images, head pose estimation results with average estimation errors of less than 4 degrees for pan and tilt for multi-user systems (12 users) were achieved. For new users, av-

a) b) c) d)

**Figure 1. Sample images used in our previous work on head pose estimation. Images c) and d) were captured with an omnidirectional camera [11].**

**Figure 2. Sample images from the Pointing'04 head pose data base.**

erage errors of less than 10 degrees for pan and tilt were achieved.

In the remainder of this section, we present our approach in detail and report the head pose estimation results on the data provided for the Pointing04 ICPR workshop.

### 2.1 The Pointing04 Workshop Head Pose Data Base

The database used for this evaluation consists of 15 sets of images. Each set contains of 2 series of 93 images of the same person at different poses. The first serie is used for learning, the second is for testing. There are 15 people in the database, wearing glasses or not and having various skin color. The pose, or head orientation is determined by 2 angles (h,v), which varies from -90 degrees to +90 degrees. A sample of a serie is depicted in Figure 2.

The images for this database have all been collected within the FAME project by the PRIMA Team in INRIA Rhone-Alpes. To obtain different poses, markers were put in the whole room, which correspond to certain poses (h,v) and at which the subjects had to look during data acquisition. Further details about the data set and the acquisition of the data can be found on the workshop website.

### 2.2 Preprocessing of Images

To locate and extract the faces from the collected images, we use a statistical skin color model [12]. 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 in our previous work we obtained 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 3 shows the preprocessed images of a user's faces.

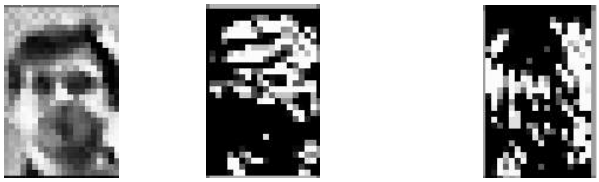

**Figure 3. Preprocessed images: normalized grayscale, horizontal edge and vertical edge image (from left to right)**

### 2.3 Neural Net Architecture, Training and Results

We have trained separate nets to estimate 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 80 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.

#### 2.3.1 Results with Multi-User System

To train a multi-user neural network, we divided the data set of the 15 users into a training set consisting of 2232 images (80% of the data), a cross-evaluation set of size 279 images (10%) and a test set with a size of 279 images (10%). After training, we achieved a mean error of 10.6 degrees for pan and 10.4 degrees for tilt on the multi-user test set.

| | pan | tilt |
|---|---|---|
| basic set of training images | 10.6 | 10.4 |
| + additional mirrored images used | 9.5 | 9.7 |

**Table 1. Average error in degrees for pan and tilt estimation on the Pointing04 ICPR workshop 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, of course). As a result, the available amount of data could be doubled without having the effort of additional data collection.

After training with the additional data, we achieved an average error of 9.5 degrees for pan and 9.7 degrees for tilt on the multi-user test set.

Table 2 summarizes the pan and tilt estimation results on the Pointing04 ICPR workshop data.

It has to be noted that the face orientation data used here consists of faces collected at orientations of 15 degree steps for horizontal rotation, and 30 degree steps for vertical rotation. The task of head estimation head orientation could therefore be treated as a classification problem, where the correct orientation class for pan and tilt orientation has to be detected from an input image.

Instead of treating the face orientation task as a classification problem, however, we have decided to estimate the absolute head orientation directly, as in our previous work.

To obtain a classification measurement of our approach, we have mapped the obtained head pose estimations to the head orientation classes provided in the data. By doing this we achieved a classification accuracy of 52% for horizontal orientation (13 classes) and 66.3% for vertical orientation (7 classes). The corresponding confusion matrices are depicted below.

## 3 Conclusion

In this paper we have reported head pose estimation results on the Pointing04 workshop evaluation set for facial orientation. We have used a neural network based approach to estimate head pose from facial input images. On a randomly chosen test set containing 10% of the images from the provided evaluation data, we achieved head pose estimation with an average error of 9.5 degrees for horizontal head rotation (pan) and 9.7 degrees for vertical head rotation (tilt). This corresponds to correct classification rates of 52% for classifying the correct horizontal head orientation class (1 out of 13 possible classes) and 66.3% for classifying the correct vertical head rotation class (1 out of 7 possible classes).

| | -90 | -60 | -30 | 0 | 30 | 60 | 90 | sum |
|---|---|---|---|---|---|---|---|---|
| -90 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 3 |
| -60 | 0 | 24 | 9 | 0 | 0 | 0 | 0 | 33 |
| -30 | 0 | 5 | 26 | 6 | 0 | 0 | 0 | 37 |
| 0 | 0 | 1 | 26 | 80 | 30 | 0 | 0 | 137 |
| 30 | 0 | 0 | 0 | 5 | 24 | 4 | 0 | 33 |
| 60 | 0 | 0 | 0 | 1 | 4 | 29 | 0 | 34 |
| 90 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 |
| | 1 | 32 | 61 | 92 | 58 | 34 | 1 | 279 |

**Table 3. Confusion matrix for classifying vertical head orientation. Correct classification was achieved 66.3% of the time.**

## Acknowledgments

This work has partially been funded by the European Commission under contract nr. 506909 within the project CHIL (http://chil.server.de).

## 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] 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.

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

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

[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.

| | -90 | -75 | -60 | -45 | -30 | -15 | 0 | 15 | 30 | 45 | 60 | 75 | 90 | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -90 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -60 | 0 | 1 | 15 | 9 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 |
| -45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -30 | 0 | 0 | 0 | 9 | 21 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 37 |
| -15 | 0 | 0 | 0 | 2 | 6 | 24 | 12 | 1 | 0 | 0 | 0 | 0 | 0 | 45 |
| 0 | 0 | 0 | 0 | 0 | 1 | 11 | 23 | 15 | 0 | 0 | 0 | 0 | 0 | 50 |
| 15 | 0 | 0 | 0 | 0 | 0 | 1 | 7 | 23 | 11 | 0 | 0 | 0 | 0 | 42 |
| 30 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 18 | 8 | 0 | 0 | 0 | 33 |
| 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 60 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 7 | 21 | 3 | 0 | 34 |
| 75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 90 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| | 0 | 2 | 17 | 20 | 36 | 41 | 46 | 45 | 31 | 15 | 22 | 4 | 0 | 279 |

**Table 2. Confusion matrix for classifying horizontal head orientation. Absolute counts are given. Correct classification was achieved 52.0% of the time.**


[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] R. Stiefelhagen, J. Yang, and A. Waibel. Simultaneous tracking of head poses in a panoramic view. In *International Conference on Pattern Recognition*, volume 3, pages 726–729, September 2000.

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