Neural Network-based Head Pose Estimation and Multi-view Fusion

– Draft Version –

Michael Voit, Kai Nickel, and Rainer Stiefelhagen

Interactive Systems Lab, Universität Karlsruhe (TH), Germany,
{voit|nickel|stiefel}@ira.uka.de

Abstract. In this paper, we present two systems that were used for head pose estimation during the CLEAR06 Evaluation. We participated in two tasks: (1) estimating both pan and tilt orientation on synthetic, high resolution head captures, (2) estimating horizontal head orientation only on real seminar recordings that were captured with multiple cameras from different viewing angles. In both systems, we used a neural network to estimate the persons' head orientation. In case of seminar recordings, a Bayes filter framework is further used to provide a statistical fusion scheme, integrating every camera view into one joint hypothesis. We achieved a mean error of 12.3° on horizontal head orientation estimation, in the monocular, high resolution task. Vertical orientation performed with 12.77° mean error. In case of the multi-view seminar recordings, our system could correctly identify head orientation in 34.9% (one of eight classes). If neighbouring classes were allowed, even 72.9% of the frames were correctly classified.

1 Introduction

A lot of effort in today's research in human computer interfaces is put in analysing human activities and human-human interaction. An important aspect of human interaction is the looking behavior of people, which can give insight to their focus of attention, to whom they are listing, as well as about the general dynamics of interaction and the specific roles that people play. Since using special gear is prohibitive in real-life scenarios, visual analysis of people's head orientation has received more and more attention over the last years.

1.1 Related Work

Until now, various approaches for visually estimating head pose were presented. Yet, the interacting person whose pose was to be recognized often had to limit its movement and rotation to a fixed area around the camera. This prohibits natural behaviour and only allows to embed those systems in environments where the user's freedom of movement is restricted anyway (like in a car or in front of a screen).

Especially model-based approaches as presented in [4, 3, 5] are affected by this constraint. Since in these approaches, a number of facial features need to be detected to compute head pose, they require facial images of quite high resolution and also suffer of tracking problems due to fast head movements.

In contrast, appearance-based approaches tend to achieve satisfactory results even with lower resolutions of extracted head images. In [6] a neural-network-based approach was demonstrated for head pose estimation from very low resolution facial images which were captured by a panoramic camera. Here, however, the output only covered ranges from the left to the right profile. Also only one camera view was used, thereby limiting the application of the system to an area around a meeting table.

Another interesting work is described by Ba and Obodez in [2]. They classify facial images by modelling the responses of Gabor and Gaussian filters for a number of pose classes. An interesting contribution of their work is the combination of head detection and pose estimation in one particle filter framework. However, their work was limited to a monocular system.

Tian et al. [7] described the use of wide baseline overhead stereo-cameras in a room to classify an observed head pose into one of a fixed set of discrete pose classes. Neural networks were implemented for estimating the head pose seen by each camera. A maximum-likelihood search results in the final pose hypothesis. Though the architecture of the presented system seems to be usable for more than two cameras, the work lacks an example with more than one camera pair. To our knowledge, this is the only work combining multiple views for head pose estimation.

1.2 Paper Overview

This paper presents two independent systems that were used during CLEAR Evaluation 2006 [**?**]. The task of head orientation estimation comprises two distinct datasets: one is a monocular, synthetic setup with high resolution, frontal captures of different persons' rotated heads [1], the other are real seminar recordings that were recorded with four fixed, overhead cameras that were setup in the upper corners of a smartroom [**?**]. Due to the far distance of the cameras in the latter scenario, head regions mostly suffer from a rather poor resolution and the main task herein lies to take the advantage of using multiple views in order to stabilize the system's output.

Section 2 of this paper gives an overview of the neural network architecture we used for evaluating the monocular setup. Section 3 adapts that system's idea of using one neural network for single-view estimation, extends it to use multiple views and combine the single estimates to one joint hypothesis. Section 4 provides a short conclusion.

2 Monocular Head Pose Estimation

2.1 Task Overview

In the monocular head pose task, the Face Pointing04 database provided by the PRIMA Team in INRIA Rhone-Alpes [1] was being chosen. The database used for this evaluation consists of 15 sets of images. Each set contains 2 series of 93 images of the same person at different poses. The first series 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 vary from $-90°$ to $+90°$. To obtain different poses, markers were put in the whole room at which the subjects had to look during data acquisition. A sample of the dataset is depicted in Figure 1. Further details regarding this dataset can be found on the corresponding website [1].

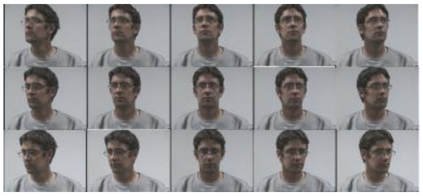

Fig. 1. Sample images from Face Pointing04 Database

2.2 System Overview

Since the task only contains the requirements to estimate head orientation on static, independent captures, no temporal filtering becomes necessary. Hence, we trained one single neural network similar to the one described in our previous work [8] with two output units for estimating horizontal and vertical head orientation continuously.

The network follows a three-layered, feed-forward topology, including 100 hidden neurons in the second layer. As input, the cropped head region is downsampled to an image size of 64×64 pixels, grayscaled and linearly stretched in its contrast to overcome small lighting changes. A Sobel operator is then applied to get the magnitude response in both horizontal and vertical derivation. Both images are then concatenated to obtain a feature vector of 8192 dimensions which is fed into the network's input layer (as depicted in Figure 2). Since the database does not provide head bounding boxes, a head segmentation step is necessary in order to align a bounding box around the region of interest. We implemented a linear boundary decision classifier in HSV color space to segment skin color cluster. The classifier was trained exclusively on the training images of the dataset. A connected component search over the segmented skin pixels results in the head

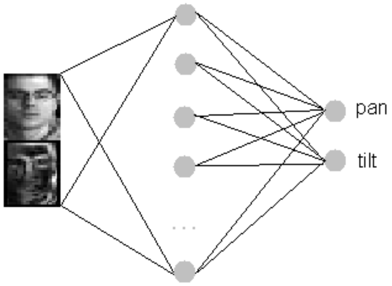

Fig. 2. We used one neural network for estimating both horizontal and vertical head pose in the monocular task. We trained two output neurons to estimate both orientations continuously.

bounding box. In order to double the training data, we mirrored the training images and added them to the training step.

The network was trained using standard error backpropagation and sigmoid activation functions. A cross evaluation set was used to obtain the best performing network among the 100 training cycles. As a final step the network's output is discretized into one of the defined classes.

2.3 Results

Table 1 shows our results on the described dataset. As it can be seen, our implementation performed with 12.3° mean error on horizontal orientation hypotheses and 12.77° mean error on vertical orientation estimations. We believe, the performance can well be increased by including a variance in cropping head bounding boxes such that inconcistent head alignment might be trained into the neural network and stabilise its performance. However, using a linear decision boundary in HSV space subjectively showed sufficient quality for segmenting the head region.

Pan Avg. Error	Tilt Avg. Error	Correct Pan Class	Correct Tilt Class
12.3°	12.8°	41.8%	52.1%

Table 1. Results of our monocular head pose estimation system on the Face Pointing04 Database.

3 Multi-view Head Pose Estimation

3.1 Task Overview

In the multi-view head pose estimation task, real seminar recordings provided by Universität Karlsruhe [**?**] were to be used in order to estimate the lecturer's head pose in horizontal direction only. The data consists of two datasets that are being used for training and evaluation respectively. The videos depict real seminar recordings from four fixed cameras that are placed in the upper corners of a seminar room. The lecturer's head bounding box and head orientation are annotated for each of the four camera views. Figure 3 depicts one sample video frame from the four cameras. Since the resolution of the captured cameras is $640x480$ pixels, the resolution of annotated head regions is poor, thus, the task in using multiple views targets at stabilising the system's output by using views from different angles. Since the lecturer's position varies, his or her head is being exposed to strong lighting changes such as the projector ray or whiteboard illumination. The background is cluttered, which is the reason why the task does not require automatic head alignment but provides manual annotations instead. The head orientation is classified into eight discrete classes: $0°, 45°, 90°, 135°, 180°, 225°, 270°$ and $315°$.

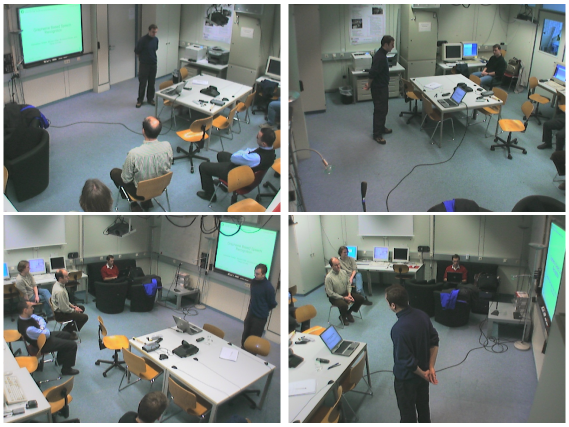

Fig. 3. Example video frame of UKA Seminar database. The lecturer of the seminar is observed by four fixed, overhead video cameras. In all views, the lecturer's head bounding box and horizontal head orientation is manually annotated.

3.2 System Overview

As in 2.2, we trained one neural network to estimate head orientation. Here, however, we trained the network to output the head orientation relative to one single camera: By using relative head pose angles, the very same neural network may be used for all camera views.

The network follows a three-layered, feed-forward topology, including 100 hidden neurons in the second layer. As input, the cropped head region is preprocessed in the same way as in section 2.2. Due to the low resolution of head captures, we only resampled to 32×32 pixels, thus the network only receives 2048 dimensions in total.

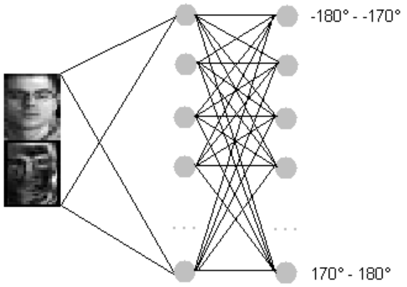

Fig. 4. In the multi-view setup, we trained one neural network with 36 output neurons. Each of them represents one discrete head pose class, relative to the camera's line of view (in 10° steps). The network was trained to estimate the class-conditional likelihood of the corresponding output class given the observation of that camera.

Further, the original network topology was modified to not outputting a continuous estimation of the horizontal head orientation but to output class-conditional probabilities $p(c_k|z_j)$ of a discretization c_k of possible head rotations, relative to camera j's line of view. The observation of camera j is denoted by z_j. Our experiments showed that a discretization into 36 classes, each 10° wide, performed best, thus allowing the network to give a hypothesis for the full range of observable head poses, from -180° to $+180^\circ$.

Concerning head orientation in room coordinates, we defined 360 states $X = x_i$, with $0 \leq i \leq 359$, where every state describes one possible head rotation. We implemented a Bayes filter for the transition between these states. Thus, given observations $Z = z_j$ of all cameras, our fusion can be written as:

$$p(X_t = x_i|Z_t) = p(Z_t|x_i) \cdot \sum_{x' \in X} p(X_t = x_i|X_{t-1} = x')p(X_{t-1} = x'|Z_{t-1}) \quad (1)$$

The observation model gathers the estimations of all n cameras, into one combined measurement, such that

$$p(Z_t|x_i) = \frac{1}{n} \sum_{j=1}^{n} p(Z_t|\phi_j(x_i)) \quad (2)$$

given the current observations Z_t. Here, $\phi_j(x_i)$ serves as a mapping from the absolute head pose angle x_i to one of the camera-relative rotation classes c_k of camera j.

The sum in equation 1 is made up of two factors: $p(X_t = x_i|X_{t-1} = x')$ describes the transition probability to go from state x' to x_i. The second factor $p(X_{t-1} = x'|Z_{t-1})$ represents the posterior probability distribution at time $t-1$.

Having computed the distribution of all states and transitions, we accumulate the probabilities of all states, which fall into the very same output orientation class θ_l of those defined by the task ($\Theta = 0°, 45°, 90°, \ldots$). The final output can then be given as the highest scored orientation $\hat{\theta}$ such that:

$$\hat{\theta} = \arg\max_{\theta_l \in \Theta} \sum_{x_i \in \theta_l} p(X_t = x_i | Z_t) \tag{3}$$

3.3 Experimental Results

The system has been trained on the training dataset only, evaluation took place on the evaluation set exclusively. No further head alignment was done, the annotated head bounding boxes were used directly to extract the head region.

Avg. Error	Correct Class	Correct + neighbouring class
49.2°	34.9%	72.9%

Table 2. Results of our multi-view head pose estimation system on the UKA Seminar Database.

Our system performed with 34.9% correct classification, when allowing the system's output to lie within the correct or neighbouring classes the performance increased two 72.9%. We believe that an additional alignment step would further increase the system's performance, since the manual labelling still varies in position and size.

4 Conclusion

In this work, we have presented two system for estimating head pose under different conditions that were used during CLEAR Evaluation 2006. One task was to estimate head orientation both in horizontal and vertical direction on monocular, synthetic head captures (Face Pointing04 Database). The second task was to hypothesise horizontal head orientation on multi-view, real seminar recordings (UKA Seminar Database). In both systems, head orientation is estimated per camera using a neural network. In case of the multi-view seminar scenario, an attached Bayes filter both fuses the single cameras' estimations as well as provides temporal filtering to smooth the system's output on each video recording. Using one single neural network that is applied on every camera, our approach is flexible and allows for easy change of camera positions and additional sensors without the necessity of retraining the whole system. The Bayes filter framework is independent of the amount of cameras and can easily be extended by further information coming from even than the four views that were provided in the dataset.

In case of the monocular setup, our system estimated horizontal head orientation with a mean error of 12.3°, vertical orientation estimation performed with a mean error of 12.77°. Since the dataset did contain static face captures only, no temporal filtering was applied.

Our multi-view head pose estimation system, we used on UKA Seminar Database, performed with a correct classification of 34.9%. When allowing for neighbouring classes, even 72.9% are correctly classified.

5 Acknowledgement

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

References

1. Pointing'04 icpr workshop, http://www-prima.inrialpes.fr/pointing04/.
2. S. O. Ba and J.-M. Obodez. A probabilistic framework for joint head tracking and pose estimation. In *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.
3. A. H. Gee and R. Cipolla. Non-intrusive gaze tracking for human-computer interaction. In *Proceedings of Mechatronics and Machine Vision in Practise*, pages 112–117, 1994.
4. T. Horprasert, Y. Yacoob, and L. S. Davis. Computing 3-d head orientation from a monocular image sequence. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, 1996.
5. R. Stiefelhagen, J. Yang, and A. Waibel. A modelbased gaze tracking system. In *Proceedings of the IEEE International Joint Symposia on Intelligence and Systems*, pages 304–310, 1996.
6. R. Stiefelhagen, J. Yang, and A. Waibel. Simultaneous tracking of head poses in a panoramic view. In *International Conference on Pattern Recognition*, 2000.
7. Y.-L. Tian, L. Brown, J. Connell, S. Pankanti, A. Hampapur, A. Senior, and R. Bolle. Absolute head pose estimation from overhead wide-angle cameras. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003.
8. M. Voit, K. Nickel, and R. Stiefelhagen. Multi-view head pose estimation using neural networks. In *Second Workshop on Face Processing in Video (FPiV'05), in Proceedings of Second Canadian Conference on Computer and Robot Vision. (CRV'05), 9-11 May 2005, Victoria, BC, Canada*, 2005.