Multi-view Head Pose Estimation using Neural Networks

Michael Voit Kai Nickel Rainer Stiefelhagen
Interactive Systems Labs
Universität Karlsruhe (TH)
Germany
{voit | nickel | stiefel}@ira.uka.de

Abstract

In the context of human-computer interaction, information about head pose is an important cue for building a statement about humans' focus of attention. In this paper, we present an approach to estimate horizontal head rotation of people inside a smart-room. This room is equipped with multiple cameras that aim to provide at least one facial view of the user at any location in the room. We use neural networks that were trained on samples of rotated heads in order to classify each camera view. Whenever there is more than one estimate of head rotation, we combine the different estimates into one joint hypothesis. We show experimentally, that by using the proposed combination scheme, the mean error for unknown users could be reduced by up to 50% when combining the estimates from multiple cameras.

Keywords: Neural Networks, Head Pose Estimation, Smart Rooms, Human Computer Interaction

1. Introduction

This work is embedded in the framework of the European Union research project CHIL (Computers in the Human Interaction Loop) [13]. Rather than requiring user attention to operate machines, CHIL services attempt to understand human activities and interactions to provide helpful services implicitly and unobtrusively. Realizing those kind of services requests to register all expressing modalities a human offers. One of the targeted scenarios is a so called smart-room which can be used for meetings and lectures. This room is equipped with a variety of sensors, such as cameras and microphones, providing data that is used to capture e.g. speech, gestures and head orientation in order to gather information about the users' current occupation, intention or focus of attention.

In particular, people's gaze gives a clue about their focus of attention [8] [1] [3] [5] [12]. Depending on the encountered focus, it is possible to decide whether humans are interruptible, how to present supportive information (acoustically or visually) and what location and device is most appropriate to display a certain type of information.

Ideally, the problem of tracking a person's line of sight would be addressed by measuring eye-gaze directly. However, in natural environments where people do not directly interact with computerized devices, tracking eye-gaze with distant cameras suffers from low resolution images and unrestricted human motion. In order to overcome these problems, we concentrate on estimating head pose instead of eye-gaze. We use multiple distant cameras that cover the entire room, and then combine their measurements in order to acquire a more robust and reliable head pose estimate. As argued by Stiefelhagen et al. [10], head orientation is a viable approximation for eye-gaze and can be used to implement many practical applications.

1.1. Related Work

In recent years, various approaches for visually estimating head pose were presented. Yet, the interacting person whose pose shall 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 to move is restricted anyway (like in a car or in front of a screen).

Especially model-based approaches as Horprasert et al. presented in [4] are affected by this constraint: By projecting facial features from their corresponding head model built in 3D to the observed image plane in 2D, the work describes how to hypothesize the observed head pose. Even though model-based hypotheses as such tend to provide very robust detection rates of a head's orientation, they require high-resolution captures for localizing the necessary features spread over the entire face. In contrast, appearance-based approaches tend to achieve satisfactory results even with lower resolutions of extracted head images. Stiefelhagen et al. [9] demonstrated the implementation of neural networks for detecting head orientations. Facial images

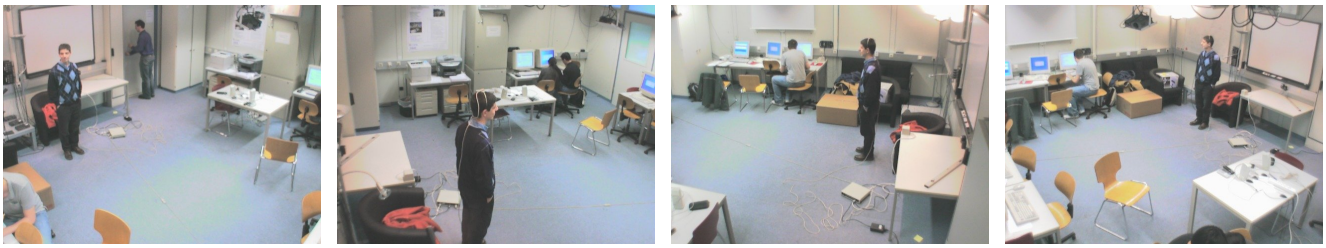

Figure 1. Views from the four cameras in our setup. The size of the head differs between the single views, two cameras suffer from back light. The resolution of the head region is very low. Thus, using multiple views of the head might help in reducing the error rate when estimating head pose.

were downscaled to a size of 20×30 pixels before being applied for pose estimation. However, the output only covered ranges from the left to the right profile and only one camera view was used for the estimation. Tian et al. [11] described the use of wide baseline overhead stereo-cameras 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 where more than one camera pair is used. Aside from this, we know no other work combining multiple views with low-resolution images.

Aside from neural networks, many different classifiers have been presented for appearance-based head pose estimation. One interesting work is described by Park et al. in [6], though its output is limited to a small, fixed size of pose classes. By reducing extracted images enclosing the head to a 3×4 pixel grid, the image can be interpreted as a 12-dimensional vector that can be classified to its nearest neighbours. Although being focused on detecting head orientations ranging from $0°$ to $360°$, the support for multiple cameras is missing.

1.2. Sensor Setup

Figure 2 depicts our sensor setup: four calibrated colour cameras are mounted in the upper corners of the smart-room at a height of about 2.7 m. The size of the room is 5×7m. Because of this layout, the entire room is covered by the cameras' field of view, such that at least one facial view of the user's head can be obtained. However, the missing ability to zoom optically, results in very low-resolution images of the extracted head[1], depending on where the person is standing. Figure 1 shows the four views of the room as seen by the cameras.

Due to non-uniform lighting from different sources (sunlight, fluorescent tubes, halogene lamps), it is inevitable that the single head views suffer from different, non-optimal lighting conditions. Further, at least two cameras are confronted with back light.

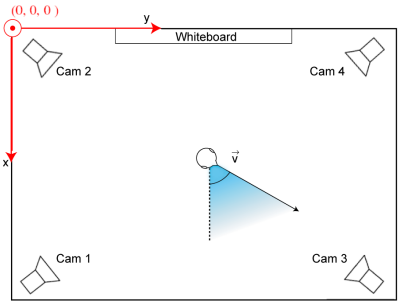

Figure 2. Four cameras are placed at the corners of the smart-room, such that at least one facial view of the head can be obtained. We estimate the horizontal rotation angle (pan) of a person's head by combining the estimates from multiple cameras.

2. Estimating Head Pose with Neural Networks

For each camera view, we estimate the head's orientation with a neural network that outputs the head's relative pan angle to the corresponding camera's line of sight. Hereby, the topology follows [9], which suggests using a three-layered feed-forward architecture.

1 Using the native camera resolution of 640×480 pixels, the typical size of a head is about 20×30 to 50×65 pixels.

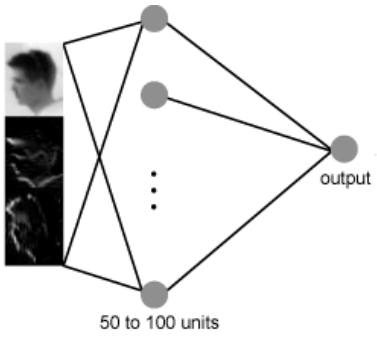

Figure 3. We use a three-layered, feed-forward neural network to estimate the pan angle relative to each camera. Estimating the relative angle allows for using the same network for all camera views and makes it easy to add cameras to the system without retraining the network. The network's input pattern consists of a normalized intensity image of the head as well as a vertical and horizontal edge image (Sobel). The single output unit provides the relative pan angle with respect to the camera, ranging from $-90°$ to $+90°$.

The net is being trained using standard error backpropagation with facial views of the head only, that range from $-90°$ to $+90°$. As depicted in figure 3, we decided to use a normalized intensity image of the head along with a vertical and horizontal edge image (produced by the Sobel operator) as the network's input pattern.

2.1. Extracting the Head

Head alignment is crucial in reducing noise whilst building input patterns for neural networks that estimate the head pose depicted in the extracted images. Because of this, we decided to keep head alignment and extraction independent from an actual head tracking technique that is implemented in a different module and outputs the head's 3D centroid.

Always being aware of the centroid's coordinates, we implemented a colour model for segmenting pixels that approximate the requested head's colour. Once a new head's centroid becomes available, a parametric colour model is initialized, using the pixels neighbouring that given source and is adapted with any following encountered head regions. Using histogram backprojection then results in a support map, depicting pixels that were declared as belonging to the head due to the colour model.

Furthermore, for stabilizing the extraction, an adaptive foreground segmentation as Stauffer et al. presented in [7] was used for building a secondary, binary image, showing pixels classified as belonging to the foreground.

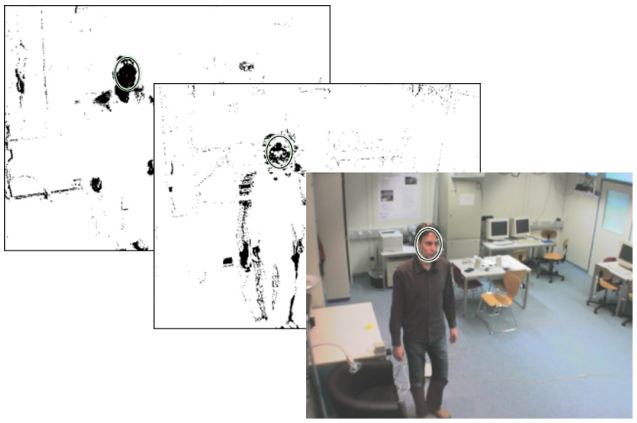

Figure 4. The head is aligned by using a colour model and adaptive foreground segmentation in order to build separate search spaces where an elliptical shape tracker approximates the best fitting ellipse.

On top of that, an elliptical shape tracker similar to [2], utilizes the intermediate results for finding a best fitting ellipse, enclosing as many supportive pixels as possible. Figure 4 depicts an example of our implemented head alignment, showing how an ellipse encloses a subset of pixels both in the support map of the colour-based tracking and in the binary result of the foreground segmentation. Let $s = (x, y, \sigma, \tau)$ be the state of an ellipse with its 2D position (x, y), horizontal radius σ and vertical radius τ. Further, be $\phi_C(s)$ the score of an ellipse with state s on a colour-based support map as denoted previously, $\phi_F(s)$ the respective score on the binary result image of the described foreground segmentation, then the best enclosing ellipse with state $\hat{s}$ can be described as

$$\hat{s} = argmax_s(\phi_C(s) + \phi_F(s)) \tag{1}$$

Both $\phi_C(s)$ and $\phi_F(s)$ rate the ellipse by scoring all of the pixels p that are enclosed in the underlying image and were defined as follows:

$$\phi_{C/F}(s) = \sum_p \zeta(p), \text{ with } \zeta(p) = \begin{cases} c & \text{if } f(p) = 0 \\ f(p) & \text{else} \end{cases} \tag{2}$$

$f(p)$ describes the support of a given pixel p, as defined by the respective support maps. The constant c is used as a penalty value for enclosing pixels that were negatively classified. A value of -1 showed sufficient results.

3. Combining Estimates from Multiple Views

We define $\Theta = \{\theta_i\}$, with $\theta_i \in \{1°, 2°, \ldots, 360°\}$ as the set of all possible head pose classes. Further, at each timestamp we have $H = \{h_1, h_2, \ldots, h_n\}$ a set of all single estimations made, where n resembles the number of cameras used.

In making a final decision about the true head pose, we score a pose hypothesis by summing up the a-posteriori probabilities of all available estimations regarding a hypothetical head pose as follows:

$$\pi(\theta_i) = \sum_{j=1}^{n} P(\theta_i|h_j) \tag{3}$$

Finding the best hypothesis then results in maximizing the score by searching for the best fitting hypothesis $\hat{\theta}$:

$$\hat{\theta} = \arg\max_{\theta_i \in \Theta} \pi(\theta_i) \tag{4}$$

This described procedure guarantees increasing hypothesis scores, the more camera views are being used and easily allows to extend an existing setup by adding more cameras in order to stabilise the estimation.

The a-posteriori probabilities were derived from confusion matrices that were built per camera whilst evaluating the classification performance of the trained neural network on the cross evaluation set. Since confusion matrices transcribe the amount of observed estimations of facial views when the ground truth of head pose was known, they allow to compute the a-posteriori probabilities of pose classes when a specific single estimation is given. That way, the posterior probability of a class θ_i given the observation h_j can be computed as

$$P(\theta_i|h_j) = \frac{k_{ij}}{\sum_m k_{mj}} \tag{5}$$

where k_{ij} denotes the matrix element in row i and column j. While the columns define the different estimation classes being made, ground truth head pose classes are described by the rows of the matrix respectively.

3.1. Detecting Facial Views

Handling views of the back of the head is crucial for achieving reliable results when estimating head pose, since the back of the head only shows hair which differs from person to person, thus providing arbitrary information instead of reliable features. Therefore, we implemented another neural network which, using the same input patterns as described in section 2, estimates the a-posteriori probability of a given head image to actually depict a facial view of the head. Regarding the combination scheme for single estimates described previously, ignoring views of the back of the head was done by setting their respective probability $P(\theta_i|h_j)$ to zero in equation 3.

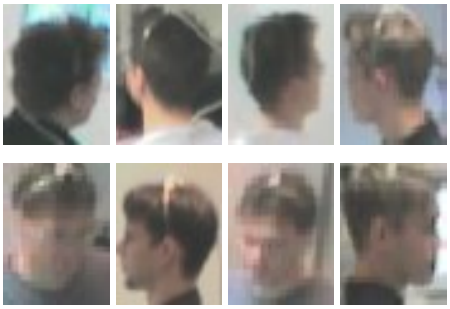

Figure 5. Recognizing facial views is crucial in achieving a robust estimation of the head's pose. We trained a neural network to output the a-posteriori probability of an extracted head image to actually depict a facial view of the head. Thereby, all images beyond $\pm 90°$ in regard to the corresponding camera's line of sight are declared as views at the back of the head.

4. Experimental Results

We evaluated the system with sequences from seven different persons. For each person, one minute of video at a frame rate of 15 fps was recorded. The test people wore a magnetic motion tracker in order to provide ground truth head pose. The persons were advised to move their head as free and natural as possible.

We implemented multiple network topologies and varied the size of the input patterns for both classifying facial views and estimating head pose. As it turned out, in case of a multiuser system, the best results of 87.9% in frontal-head classification were achieved using a size of 15×20 pixels for each of the three input images and 70 units in the hidden layer. Using this topology for recognizing views at the head's back, the minimum error of pose estimation was encountered using 60 hidden units and the same input image size. Hence, evaluating the system in an unknown user scenario shall be based upon these topologies.

The unknown user scenario was realised by implementing the leave-one-out method, where one person was removed from the training set and exclusively used for evaluation purposes. Table 1 shows the results for this scenario. As expected, the mean error of head pose estimation decreases, as more facial views could be recognized and used

Person	# facial views used 1	2	3	Ø error
1	45.8°	28.8°	23.5°	33.1°
2	39.9°	27.6°	n/a	36.9°
3	36.8°	22.1°	12.9°	25.3°
4	44.6°	28.6°	32.1°	38.0°
5	65.7°	38.2°	44.5°	53.7°
6	34.3°	19.7°	2.9°	28.2°
7	37.3°	30.3°	14.5°	31.8°

Table 1. Mean errors per person in the unknown user scenario. Obviously, the error decreases as more facial views are combined into a joint hypothesis. Person 2, 4 and 5 depict how crucial a robust classification of facial versus rear-views (Table 2) of a head is for the performance of the proposed combination scheme.

Person	Correct recognition of facial views
1	87.5%
2	77.9%
3	89.7%
4	80.2%
5	68.5%
6	87.2%
7	84.1%
Ø	82.4%

Table 2. Recognition rate of facial views in the unknown user scenario. Person 2, 4 and 5 show an unsatisfying high error which leads to an unstable head pose estimation as depicted in table 1.

for building the final hypothesis. This shows clearly how crucial a robust classification of facial views is, as disjoint views provide no possibility in building a stable hypothesis as two neighboured views at the head would. Table 2 displays the correct recognition rate of facial images for each of the seven test people, clearly showing the insufficient performance for test person 2, 4 and 5, which leads to the unsatisfying results regarding their head pose estimation.

One particular problem facing multi-view head pose estimation as described in our setup is characterized by the large distance at which different views of the corresponding head are captured from. Especially estimates built on only one recognized facial view suffer from the low resolution of the extracted head images. The lack of not being able to zoom optically even worsens this issue. Moreover, because of different head sizes and shapes, provided from different camera angles, input patterns suffer from a lot of noise. Also, the mounting of the cameras at a height of 2.7m produces views at the head seen from above, which of course implies a lot of hair being captured inescapably, especially if the head is tilted further.

Nevertheless, it gets obvious that using more cameras stabilizes the estimation of a final hypothesis. Especially regarding the final results as shown in Table 3, a definite decrease of the mean error can be observed. Inevitably, this leads to the conclusion that using even more cameras than we did in our setup should further be investigated. In fact, depending on the room's dimensions and the location of the cameras, using four cameras seems to be the very minimum in covering the complete room as it was presented here.

	# facial views used 1	2	3	overall error
multiuser	26.7°	18.4°	15.5°	21.8°
unkn. persons	44.1°	27.8°	21.9°	35.1°

Table 3. Overall mean error for both the known user scenario and the unknown user scenario. Both results show clearly how the error decreases the more camera views could be combined to form the final hypothesis.

5. Conclusions

In this work, we present a neural-network based system for estimating horizontal head rotation in a smart-room. Two neural networks are used in this system: one to estimate head orientation from facial head images in a range between $\pm 90°$, and the other to recognize facial views of the head. We describe a statistical combination scheme that integrates the rotation estimations from different camera views in order to form a joint hypothesis. Our experiments show, that the mean error decreases as more facial views are available: accomplished by two facial views of the head, the overall mean error averaged $27.8°$. When using even three facial views it dropped to $21.9°$. The performance of the frontal-head classifier was 82.4% in average, thus indicating that stabilizing the elimination of views of the back of the head might lead to even better results in building a final statement about the head's pose.

Due to the ability of the system to integrate estimates from multiple views (thus covering the entire room), tracked

persons are allowed to move freely instead of being limited to a fixed area as often premised by previous work. Moreover, the described setup depicts an unobtrusive system which is working in the background without the need of wearing head mounted devices in order to achieve complete freedom in movement.

Due to the lack of optical zoom, the cameras currently used in our system only provide low-resolution images of the head that further suffer from the varying illuminations at different locations. However, the experiments show, that the appearance-based classifier described in this paper is able to produce viable results for many applications. Concerning that the error might still be too high for fine granulated pose estimation, a subdivision in coarser pose classes would allow to classify people's focus to surrounding objects such as the whiteboard or audience. However, the approach we present in this paper is not limited to the described low resolution, overhead camera setup. In particular, the additional use of high resolution sensors with the ability to zoom optically should improve the results dramatically when estimating people's head orientation. But also being limited to a similar setup as described here, we would expect to obtain even lower error rates as more cameras become available.

It should also be noted, that the system currently does not consider temporal information that could well be used to smooth the final hypotheses.

Acknowledgements

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

References

[1] M. Argyle. *Social Interaction*. Methuen, London, 1969.

[2] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–237, 1998.

[3] N. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24(6):581–604, 2000.

[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] P. P. Maglio, T. Matlock, C. S. Campbell, S. Zhai, and B. A. Smith. Gaze and speech in attentive user interfaces. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 1–7, 2000.

[6] S. Park and J. K. Aggarwal. Head segmentation and head orientation in 3d space for pose estimation of multiple people. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, 2000.

[7] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

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

[9] R. Stiefelhagen, J. Yang, and A. Waibel. Simultaneous tracking of head poses in a panoramic view. In *International Conference on Pattern Recognition*, 2000.

[10] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems*, 2002.

[11] 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.

[12] M. von Cranach. The role of orienting behaviour in human interaction. In A. H. Esser, editor, *Environmental Space and Behaviour*. Plenum Press, New York, 1971.

[13] A. Waibel, H. Steusloff, R. Stiefelhagen, and the CHIL Project Consortium. Chil: Computers in the human interaction loop. In *5th International Workshop on Image Analysis for Multimedia Interactive Services*, 2004.