Head Pose Estimation in Single- and Multi-view Environments
Results on the CLEAR'07 Benchmarks
- 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 our system used and evaluated on the CLEAR'07 benchmarks on single- and multi-view head pose estimation. The benchmarks show a high contrast in the application domain: Whereas the single-view task provides simulated meeting recordings, involving high-quality recordings of the participants, the multi-view benchmark targets at low-quality, unobtrusive observations of people with multiple cameras in unconstrained scenarios. We show that our approach achieves satisfactory results in both domains.

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

Head pose estimation has got increasing attention due to its unobtrusive possibility to estimate peoples' looking direction. A lot of different approaches were presented which, in general, can be categorized into either model-based or appearance-based techniques. Model-based works such as [3, 2, 4] allow quite precise hypotheses about the orientation. Due to the necessary feature detection however, they are only applicable in areas where near frontal shots of peoples' faces are ensured. Further, high resolutions seem necessary since building on detailed face features (nostrils, eyes) mostly becomes impossible once the head's resolution decreases in its total dimensionality. Here, appearance-based approaches tend to achieve satisfactory results even with lower resolutions of extracted head images. In [5] a neural-network-based approach was demonstrated for head pose estimation from rather low resolution facial images which were captured by a panoramic camera. The output covered head poses from the left to the right profile.

Another interesting work is described in [1], where facial images are modeled by the response 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 joint particle filter framework. Integrating head detection and classification into one combined step allows to overcome alignment problems, from which most appearance-based techniques suffer the most.

2 Task Descriptions

2.1 The CHIL Data Corpus - Multi-view Head Pose Estimation

The CHIL task in CLEAR07 included the use of multiple cameras in order to gather singleview hypotheses into one joint, robust estimate. We used Neural Networks to gather monocular estimates and a particle filter without resampling for multi-view fusion. This section describes the task in detail and our system we are to present.

The CHIL smartroom is is equipped with several sensors to gather both audio and visual features about peoples' occupations and activities. Amongst numerous microphones and microphone arrays (both for speaker source localization and far field speech recognition), several cameras are installed to allow unobtrusive visual people tracking, person identification or head pose estimation. Overall, four fixed and calibrated cameras were used, that are placed in the room's upper corners (Figure 1). The cameras do not obtain any zooming abilities and capture with a

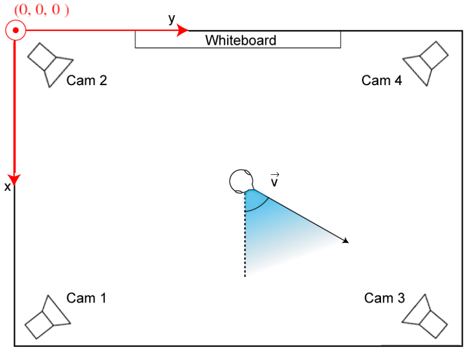

Fig. 1. Setup of the CHIL head pose task: Four cameras were installed in a room's upper corners to capture the whole area underneath them. This surrounding setup allows people to move and behave without restrictions regarding a specific sensor. Using numerous cameras always guarantees to capture at least one frontal view. However, it is inevitable that some cameras only capture the back of the head, depending on how the head is rotated.

resolution of 640×480 pixels at 15 frames per second, hence, concerning where a person is standing in the room, head captures tend to vary strongly in size. Overall, head captures as small as 20×30 pixels (see Figure **??** for some sample images) could be observed, not allowing any detailed detection of nostrils, eye or mouth corners that might allow any model-based approach. The use of multiple surrounding cameras allows people to move freely but guarantees that at least half of the sensors capture the back of the person's head only. However, always at least one frontal view of the head may be observed. In our previous work [6], we tried to overcome this problem by integrating a facial view classification step, where the likelihood of a head capture to actually depict a frontal view was estimated and used in the final fusion scheme of our multi-view head pose estimation. In this work, we overlap the single-view hypotheses and calculate a score for each possible head rotation in a particle-filter setup. As this scheme includes temporal smoothing by including a diffusion step, the results prove to be very robust. During recording sessions, all people in the dataset were instructed to wear a magnetic motion sensor to capture their groundtruth head orientation relative to a transmitter, which was aligned with the room's coordinate system. The tracker allowed a capture rate of $30Hz$, hence providing angle measurements that were as twice as fast as the cameras delivered pictures, thus providing enough information in real time. To avoid a tracking and alignment task, head bounding boxes were manually annotated and provided with the dataset both for training and evaluation. The final dataset contained 15 recordings with one

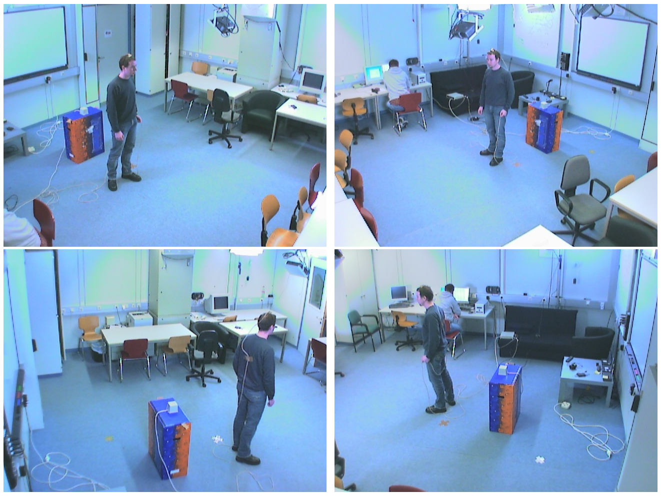

Fig. 2. Example captures of one frame from all four views in the CHIL corpus. The person recorded was to wear a magnetic motion sensor to capture his or her groundtruth head orientation.

person each. Every recording was about 3 min. long. For training, 10 of these 15 people were distributed. Evaluation took place on the remaining 5 videos.

3 The AMI Data Corpus - Single-view Head Pose Estimation

The AMI task provided single-view camera recordings of simulated meeting scenarios with two people sitting in front of a table in front of a camera. Both persons are oriented towards the camera, hence their head orientation only varies within $-90°$ to $+90°$ for both pan and tilt. No bounding boxes were provided, thus requiring an automatic alignment step. Since both persons are sitting at fixed positions, no tracking module is required for locating their coarse location. The dataset provided 8 meeting videos, hence 16 persons to estimate head pose in total. The overall length of one video is 1 min. As in the CHIL dataset, all persons involved were to wear a magnetic motion sensor to track their groundtruth head orientation. Finally, the dataset was split into a trainingset, containing 5 videos (10 people) and a testing set, including 3 videos (6 people).

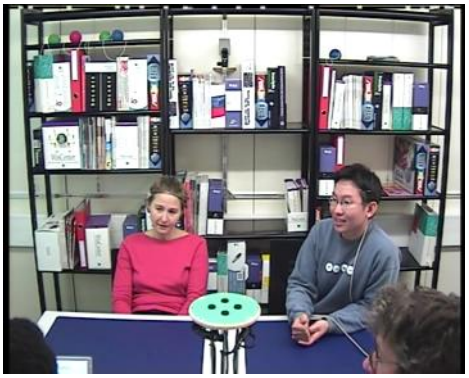

Fig. 3. Example capture of one frame from the AMI data corpus. Two meeting participants are sitting opposite to a camera. Their groundtruth head orientation is capture with a magnetic motion sensor. Due to the meeting scenario, the overall head pose range is limited to profile view reative to the capturing camera.

4 System Overview

We adopted our system already presented in [7, 8] and modified the approach to also fit on vertical pose estimation (tilt). The following subsections shall present a brief overview of the previous works.

4.1 Single-view Head Pose Estimation using Neural Networks

Neural Networks have proven, especially because of their generalisation, to be a robust classifier for the estimation of head orientation. We adopted this idea and applied this classifier for each camera view. Both horizontal and vertical head rotation were modeled with an individual classifier.

Either network follows a three-layered, feed-forward topology, receiving a preprocessed cropped head image, capturing the current observation at time t and outputting a hypothesis of the observed head rotation in either direction (horizontally or vertically).

The cropped head region is preprocessed by grayscaling, linearly stretching its grayvalued histogram to improve contrast and resizing the image to 32×32 pixel. A Sobel operator computes the normalized head region's derivation magnitude image which is concatenated to the normalized appearance, thus retrieving an overall feature vector consisting of 2048 components, derived from a merged head representation of 32×64 pixels.

The second layer was empirically chosen to contain 80 hidden units, all fully connected to both all input neurons as well as all output neurons.

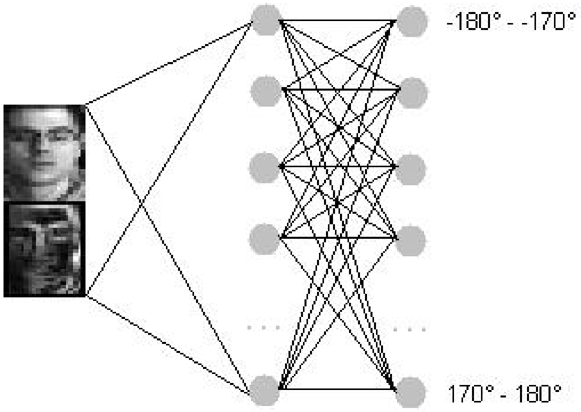

Fig. 4. In the multi-view setup, we trained one neural network with 36 output neurons. Each of them represents one discrete head pose class, relative to the camera's line of view (in 10° steps). The network was trained to estimate the class-conditional likelihood of the corresponding output class given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

Depending on the task, the network's output layer was trained to output either a likelihood distribution or a final, continuous estimation of the observed head orientation. The latter was used for the single-view task involving the AMI data corpus. Since no multiple cameras were used, no fusion scheme to merge numerous hypotheses was required, the networks' output could be used as the posterior system's output. Especially, since no uncertainty resulting from views at the back of peoples' heads is involved. Regarding our multi-view approach, the networks were trained to output a likelihood distribution of the possible head orientation over the whole range of observable rotation angles. To make the classifier sensor-independent, the networks were trained to output that likelihood over the range of relative poses to the camera's line of sight. That way, increasing the number of cameras in the setup is easily to achieve without retraining a new classifier for those cameras. Whereas, in the single-view task, we only used one output neuron both for the pan estimating as well as the tilt estimating network, the multi-view required numerous output units. We specified the range from $-180°$ to $+180°$ for pan and $-90°$ to $+90°$ for tilt, discretising the angle space to both 36 classes, 10° wide for pan, 5° respectively for tilt. Here, target outputs were modeled as gaussian densities as this uncertainty helps in matching the single views' hypotheses.

4.2 From Single-view to Multi-view Scenarios

To take advantage of having multiple views in the CHIL data corpus, single-view hypotheses are gathered from every available sensor and are to be merged into one joint, final estimation of the current observation. We apply the described network to retrieve these single-view hypotheses and merge and track with a bayesian filter. The bayesian filter resembles a general particle filter setup, omitting the resamling step, since, as described later, we only use a stationary, discrete set of states (thus particles) for pose tracking.

In our setup we compute a final estimate within a horizontal head rotation range of 360° (180° for tilt respectively). Hence, we use a fixed set of 360 (180) stationary filter states, each one representing their corresponding head rotation in horizontal (vertical) direction. The task is to compute a posterior likelihood distribution $p(x_i|Z_t)$ over this defined set of states $X = \{x_i\}$ at a given time t and single-view hypotheses Z_t. The posterior distribution can thus be described as

$$p(x_i|Z_t) = \frac{p(Z_t|x_i) \cdot P(x_i)}{p(x_i)} \tag{1}$$

The joint measurement $p(Z_t|x_i)$ is derived from the four single cameras' hypotheses with observations $Z_t = \{z_{j,t}\}$. The prior $P(x_i)$ denotes the probability to be in state x_i and includes the overall particle diffusion step, hence providing the temporal smoothing used for tracking. Each of these factors is going to be described in the following subsections.

4.3 Building a Joint Measurement

By mapping each possible head orientation x_i to an orientation $\phi_j(x_i)$, relative to camera j's line of view, we gather a combined measurement out of all single cameras' hypotheses by averaging the four class-conditional estimations, such that

$$p(Z_t|x_i) = \frac{1}{4}\sum_{j=1}^{4} p(Z_t|\phi_j(x_i)) \tag{2}$$

The intuition behind Equation 2 is that the hypothesis x_i is scored higher, the more cameras agree on it, i.e. the respective output neuron exhibits a high value. That means, if two or more hypotheses strongly agree on the very same head orientation, the final sum of these probabilities returns a much higher value than accumulating smaller likelihoods that describe rather uncertain, ambiguous estimations.

4.4 Integrating Temporal Filtering

Temporal information is implied by the prior distribution $P(x_i)$ within Equation 1. At each timestep t this factor implies the probability to observe state x_i. This factor is derived from the transition probability $p(x_i|x')$ to change into state x_i and the a-posteriori probability distribution $p(x'|Z_{t-1})$ which was computed at time $t-1$:

$$P(x_i) = \sum_{x' \in X} p(x_i|x')p(x'|Z_{t-1}) \tag{3}$$

We applied a Gaussian kernel function to provide state change propagation $p(x_i|x')$, hence updating the prior distribution can be defined as a convolution of the Gaussian kernel and the previous a-posteriori likelihoods:

$$P(x_i) = \sum_{x' \in X} N_{0;\sigma}(x_i - x')p(x'|Z_{t-1}) \tag{4}$$

In our evaluation we experimentally used a standard deviation $\sigma = 20°$.

By using a Gaussian kernel, short-term transitions between neighboring states are more likely than sudden jumps over a bigger range of states, hence the adaptation of the kernel's width directly influences, how strong temporal filtering and smoothing of the system's final output takes place.

5 Experimental Results

We evaluated our system on both the CHIL data corpus as well as the AMI data corpus. Since we only directly used the neural networks' outputs in the latter task, no temporal filtering was applied here. The CHIL corpus involved our bayesian filter scheme, which showed to improve the overall accuracy with approximately 2°.

5.1 Results on the CHIL Corpus

As described in 2.1, the dataset was split into one training set, containing recordings of 10 different people and one testset, providing videos of 5 different people. Each video was about 3 min. long, captured with a framerate of 15 frames per second. Overall, the training set provided 21636 head images which were mirrored to double the training amount. The training set was not filtered to include an evenly distributed amount of pose-specific examples. The testset was left unmirrored, hence included 13500 frames.

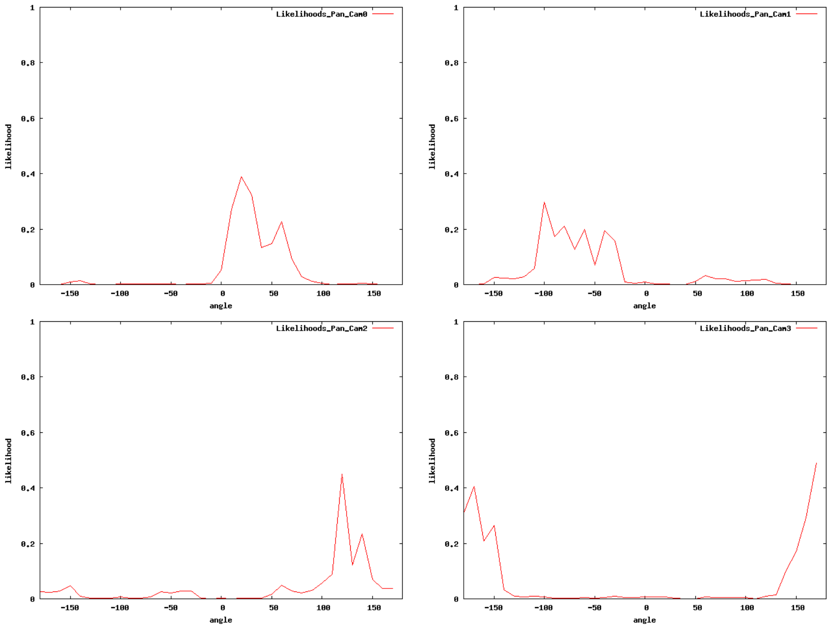

Fig. 5. Single-view likelihood distributions of all four used cameras for one single frame in the CHIL multi-view head pose task. Each distribution shows a significant cluster of high probability for a specific head orientation, relative to that cameras line of sight.

Either network's behaviour was learned in overall 100 training iterations. The training dataset was split into one training and one cross-evaluation subset (90% training, 10% cross-evaluation). Amongst 100 training iterations (in which the network's connectionist weights and activations were learned using standard error backpropagation algorithm), the one network minimizing the mean square error over the given cross-evaluation set was saved and extracted for later use, avoiding any overfitting of the network regarding the given training samples.

As can be seen in figure 5, the cameras' hypotheses generally seem follow the unimodal behaviour used during training. The uncertainty displayed in the wide variance of the distribution helps in tracking the head's orientation, since choosing the final head rotation is based on finding that specific system state, which maximizes the accumulation of the single-view hypotheses' corresponding likelihoods. Uncertainty in one view tend to be balanced with stronger confidences in the remaining views which leads to an unimodal posterior distribution as shown in figure 6.

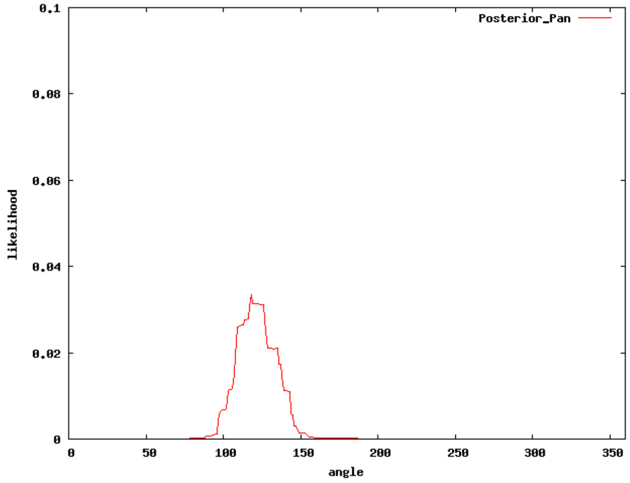

Fig. 6. The posterior distribution resulting after applying our bayes filter on the given single-view likelihoods shown in figure 5. The distribution shows to be unimodal and unambiguous.

The final results are depicted in table 5.1. Our system showed to perform with an accuracy of $8,5°$ for horizontal orientation estimation and $12,5°$ for the vertical counterpart. Omitting any temporal filtering during our bayesian filtering, resulted in an overall performance loss of $2°$.

Mean Error Pan	Mean Error Tilt	Mean Angular Error
$8,5°$	$12,5°$	$16,4°$

Table 1. Results on the CHIL data corpus. The CHIL corpus provided multi-view recordings.

5.2 Results on the AMI Corpus

The AMI training corpus included 10 recordings of two persons sitting either to the left or right side of the camera. Because of missing 3D information regarding the translation of the magnetic sensor to the recording camera, we trained individual classifiers for both the left person and the right person in order to avoid including ambiguous head pose appearances from shifted locations. Overall we evaluated with four neural networks, two for pan (left person, right person) and two networks for tilt estimation (left person, right person). All networks were trained in a similar way to our scheme in the multi-view task: The trainingset was split into one training and one cross-evaluation subset. For either side, originally 1501 frames per person were provided, hence 15010 cropped head regions which were mirrored to double the amount of training samples. The networks were trained with standard error backpropagation algorithm, using 100 iterations to extract that network, which minimized the mean square error on the cross-evaluation set. The latter was set to include 10% of the overall training samples.

Mean Error Pan	Mean Error Tilt	Mean Angular Error
$14,0°$	$9,2°$	$17,5°$

Table 2. Results of the UKA head pose system on the AMI data corpus. The corpus provided single-view recordings of meeting scenarios.

6 Conclusion

In this paper we presented the evaluation of our head pose estimation approach on the CLEAR07 head pose benchmarks. We adopted our previously presented work for horizontal head pose estimation to hypothesise the vertical rotation, too and evaluated our approach on different multi-view (CHIL data corpus) and single-view (AMI data corpus) recordings. Under both circumstances, our system proved to produce reliable results of up to $8,5°$ mean pan error and $12,5°$ mean tilt error on the multi-view dataset and $14,0°$ and $9,2°$ on the single-view dataset respectively. In the multi-view setup, people were to move their head without any restrictions, views at the head's back were as often observable as profile views or frontal views. Since the meeting setup only provided stationary sitting locations of the participants, only head rotations within profile view relative to the camera used were involved. Whereas the latter benchmark focused on interaction scenarios with multiple people involved, the multi-view recordings were oriented towards unobtrusive head pose estimation in environments where people need to move their head freely without restrictions. Both goals were successively achieved. Our system hereby uses neural networks on each camera view for estimating head orientation in either direction. Fusing multiple views' hypotheses, a bayesian filter was applied to both diffuse prior estimates (temporal propagation) as well as search for the most coherent match of overlapping single-view hypotheses of each included sensor.

7 Acknowledgement

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

References

1. 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.
2. 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.
3. 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.
4. 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.
5. R. Stiefelhagen, J. Yang, and A. Waibel. Simultaneous tracking of head poses in a panoramic view. In *International Conference on Pattern Recognition*, 2000.
6. 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.
7. M. Voit, K. Nickel, and R. Stiefelhagen. A bayesian approach for multi-view head pose estimation. In *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 4-6 September 2006, Heidelberg, Germany*, 2006.
8. M. Voit, K. Nickel, and R. Stiefelhagen. Neural network-based head pose estimation and multi-view fusion. In *Proceedings of the CLEAR Workshop 2006, Southampton, UK*, 2006.