Head Pose Estimation in Single- and Multi-view Environments – Results on the CLEAR'07 Benchmarks

Michael Voit, Kai Nickel, and Rainer Stiefelhagen

Interactive Systems Lab, Universitt 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, both on single- and multi-view head pose estimation. The benchmarks show a high contrast in the application domain: whereas the single-view task provides meeting recordings involving high-quality captures of the participants, the multi-view benchmark targets at low-quality, unobtrusive observations of people by means of multiple cameras in an unconstrained scenario. We show that our system performs with state-of-the-art results under both conditions.

1 Introduction

To obtain information about peoples' visual focus, targets they are referencing to during speeches, actions or interactions, tracking eye gaze is too difficult and obtrusive to capture when allowing natural behaviour patterns in uncontrolled environments. Instead, the estimation of peoples' head orientation easily allows to deduce knowledge about e.g. interaction dynamics without the need of wearing such special gear for detecting explicitely the participant's pupils. One of CLEAR's workshop task is to track head orientation within different domains. Therefore, CLEAR'07 introduced two different datasets, that both aim for separate scenarios: Head pose is to be estimated both for high-quality single-view meeting recordings provided by the AMI project [1], as well as for low-resolution, wideangle multi-view recordings that were captured by four upper-corner cameras during the CHIL project [2]. Whereas multi-view head pose estimation shows to be a rather young research field, head pose recognition on high quality video frames in general, already shows a lot of history both using model- [4,5,6] and appearance-based [3,7] approaches. In this work, we use one same approach for both domains: by training a neural network classifier, we are able to obtain hypotheses on a per-camera basis rather than estimate the overall posterior output immediately. In case of the multi-view scenario, a successive fusion scheme based on bayesian dynamics merges the single estimates into one final, joint system output. For both tasks we show that our technique produces state-of-the-art results.

R. Stiefelhagen et al. (Eds.): CLEAR 2007 and RT 2007, LNCS 4625, pp. 307–316, 2008.

2 Task Descriptions

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

The CHIL subtask in CLEAR'07's head pose estimation benchmark included the use of multiple cameras in order to gather and merge single-view hypotheses into one joint, robust estimate. The CHIL smartroom 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, for this task, four fixed and calibrated wideangle

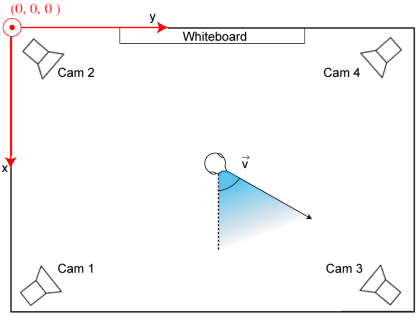

Fig. 1. Setup of the CHIL head pose task: four cameras were installed in a room's upper corners to capture the whole area underneath them. This surrounding setup allows people to move and behave without restrictions regarding a specific sensor. Using numerous cameras always guarantees to capture at least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

cameras were used, that were installed in the room's upper corners (Figure 1). The cameras do not obtain any zooming abilities and capture with a 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 can be observed, not allowing any detailed detection of nostrils, eye or mouth corners that might allow for detailed model-based approaches. The use of multiple cameras in a surrounding sensor setup allows people to move without restrictions but guarantees that at least half the sensors capture the back of the respective person's head only. However, always at least one frontal view of the head may be observed. During recording sessions, all people in the dataset were instructed to wear a magnetic motion sensor to initialize their groundtruth head orientation relative to a fixed transmitter, which was aligned with the room's coordinate system (hence, a horizontal head orientation of $0°$ would point straight along the room's positive x-axis). The tracker used

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 groundtruth head orientation.

allows to capture with 30Hz, thus providing angle annotations as fast as twice the cameras' rates. To avoid dedicated tracking and implicit head alignment, head bounding boxes were manually annotated and provided with the dataset both for training and evaluation. Overall, the final data corpus thus provided 15 recordings with one person each. Every recording was about 3 minutes long. For training, 10 of these 15 people were distributed. The successive evaluation step happened on the remaining 5 videos.

2.2 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 both in front of a table and a camera. Both persons are oriented towards the camera, hence their head orientation only varies within $-90°$ to $+90°$ for both pan and tilt rotations. The dataset included 8 meeting videos, hence 16 persons, to estimate head orientation 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 for tracking their groundtruth head orientation. The dataset itself was split into one trainingset, containing 5 videos (10 people) and a testing set, including 3 videos (6 people).

3 System Overview

We adopted and extended our system already presented in [8,9] to also cope with vertical pose estimation (tilt). The following subsections thus present a brief overview of the previous work.

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 captured 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.

3.1 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 one network respectively. Either network follows a three-layered, feed-forward topology, receiving a cropped and preprocessed head image (according to a tightest fitting head bounding box), capturing the current observation at time t and stating a hypothesis of the observed head rotation in either direction (horizontally or vertically).

The cropped head region is preprocessed by grayscaling, equalizing its histogram and resampling to 32×32 pixels. A Sobel operator computes the normalized head region's edge magnitude image which is concatenated to the normalized appearance, thus retrieving an overall feature vector of 2048 dimensions, 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.

Depending on the task, the network's output layer was trained to represent 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 ($-180°$ to $+180°$ for pan and $-90°$ to $+90°$ for tilt). To achieve sensor-independent classification,

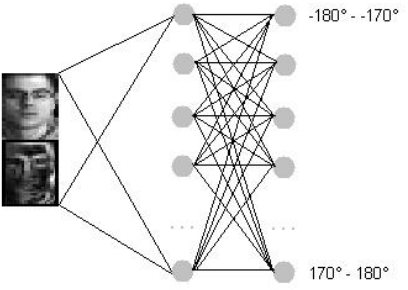

Fig. 4. In the multi-view setup, we trained one neural network with 36 output neurons. Each of them represents one discrete head pose class, relative to the camera's line of view (in 10° steps). The network was trained to estimate the likelihood of a possible rotation, given the observation of that camera.

all networks were trained to estimate those likelihoods over the range of relative poses to the respective camera's line of sight. That way, extending the setup by adding further cameras allows for no need of retraining a new classifier. Whereas, in the single-view task, we only used one single output neuron both for a pan as well as a tilt estimating network, the multi-view task thus required numerous output units to approximate the corresponding distribution. We therefore discretised the relative angle space into 36 classes each. Target outputs were modeled as gaussian densities as this uncertainty helps in correlating the single views' hypotheses as described later on.

3.2 From Single-view to Multi-view Scenarios

Taking advantage of having multiple views as in the CHIL data corpus, single-view hypotheses are gathered from every available sensor and merged into one joint, final estimation of the current observation. We apply the previously described networks to retrieve single-view distributions 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, which only need for reweighing. 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) filter states, each one representing a 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 $X = x_i$ for a given time t and single-view hypotheses Z_t . The posterior distribution can thus be described as

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

As defined, the joint measurement $p(Z_t|x_i)$ is derived from all single cameras' hy- potheses with observations $Z_t = z_{j,t}$. The prior $P(x_i)$ denotes the probability to be in state x_i, modelling diffusion and providing temporal smoothing used for tracking. Each of these factors is going to be described in the following subsections.

3.3 Building a Joint Measurement

After 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 over all cameras by averaging the four class-conditional likelihoods, 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 estimates.

3.4 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 from state x' at time $t-1$ into the current state x_i and the a-posteriori 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:

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

We used the empirically evaluated 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.

4 Experimental Results

We evaluated our system on both the CHIL [2] data corpus as well as the AMI [1] 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 by approximately 2°.

4.1 Results on the CHIL Corpus

As described in 2.1, the dataset was split into one training set, containing recordings of 10, and one testset, containing videos of 5 individual, different people. Each video was about 3 minutes long, captured with a framerate of 15 frames per second. For every 5th frame, a manually annotated head bounding box was provided. During training stage, all cropped head boxes were mirrored to double the amount of training data. 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), that network minimizing the mean square error over the given cross-evaluation set was saved and extracted for later use, thus avoiding overfitting to the given training samples. As can be seen in Figure 5, the cameras' hypotheses generally seem to 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 that maximizes the accumulation of the single-view hypotheses' corresponding likelihoods. Uncertainty in one view tends to be balanced with stronger confidences in the remaining views which leads to an unimodal posterior distribution as shown in Figure 6. The final results are depicted in Table 1: our system showed to perform with an accuracy of 8.5° for horizontal orientation estimation and 12.5° for its vertical counterpart. Omitting temporal smoothing during bayesian filtering resulted in an overall performance loss of 2°.

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

Mean Error Pan	Mean Error Tilt	Mean Angular Error
8.5°	12.5°	16.4°

4.2 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

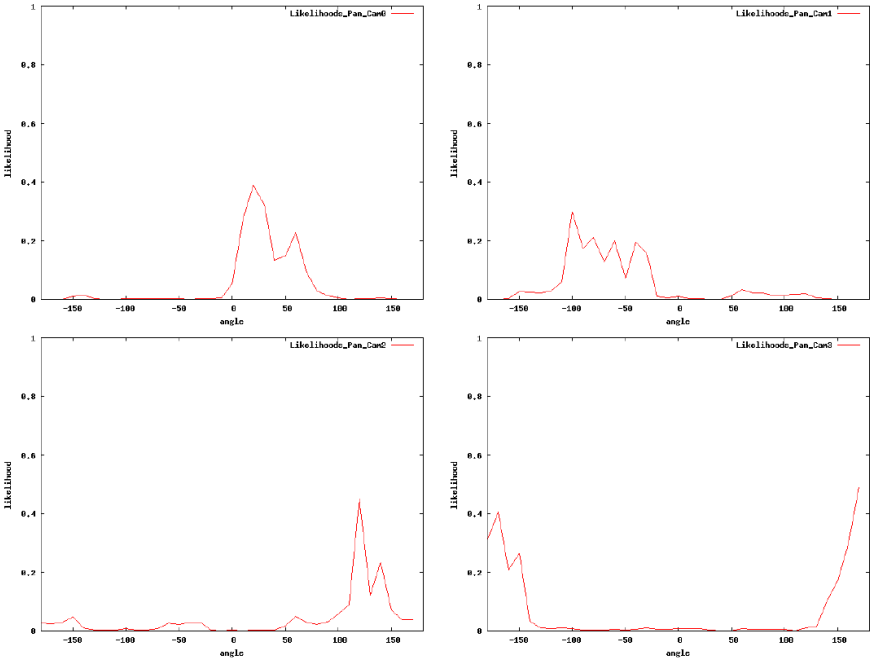

Fig. 5. Single-view pan likelihood distributions of the four used cameras for one single frame in the CHIL multi-view head pose task. Each distribution shows a significant cluster of high probability for a specific head orientation, relative to that cameras line of sight.

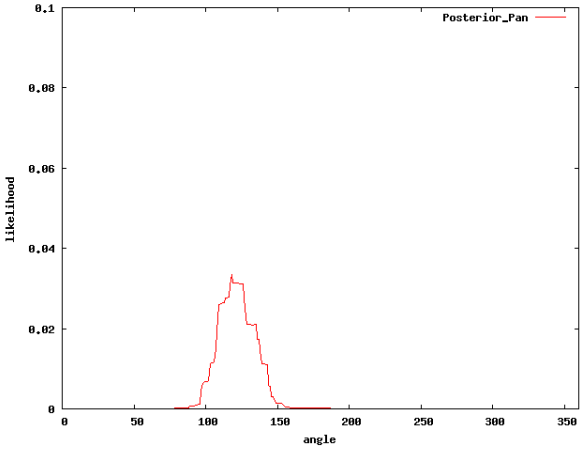

Fig. 6. The posterior distribution resulting after applying the bayesian filter on the given single-view likelihoods shown in Figure 5. The distribution is unimodal and unambiguous.

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

Mean Error Pan	Mean Error Tilt	Mean Angular Error
14.0°	9.2°	17.5°

the translation of the magnetic sensor to the recording camera, we trained individual classifiers for both the left 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 training set was split into one training and one cross-evaluation subset. Here, too, the cropped head regions during training stage were mirrored to double the amount of samples. Since no bounding boxes were provided, a skin-color classifier helped to detect the corresponding person's head bounding box. Due to the static seating locations of the participants, no tracking became necessary.

The networks were trained with standard error backpropagation algorithm, using 100 iterations to extract that network minimizing the mean square error on the cross-evaluation set. The latter was set to include 10% of the overall training samples.

5 Conclusion

In this paper we presented the evaluation of our head pose estimation approach on the CLEAR'07 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 and state-of-the-art 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 or frontal views. Since the single-view meeting scenarios only provided fixed locations of the participants, only head rotations within profile range 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. For the fusion of multiple views', 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 over all included sensors.

Acknowledgement

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

References

1. http://www.amiproject.org
2. http://chil.server.de
3. Ba, S.O., Obodez, J.-M.: A probabilistic framework for joint head tracking and pose estimation. In: Proceedings of the 17th International Conference on Pattern Recognition (2004)
4. Gee, A.H., Cipolla, R.: Non-intrusive gaze tracking for human-computer interaction. In: Proceedings of Mechatronics and Machine Vision in Practise, pp. 112–117 (1994)
5. Horprasert, T., Yacoob, Y., Davis, L.S.: Computing 3d head orientation from a monocular image sequence. In: Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition (1996)
6. Stiefelhagen, R., Yang, J., Waibel, A.: A model-based gaze tracking system. In: Proceedings of the IEEE International Joint Symposia on Intelligence and Systems, pp. 304–310 (1996)
7. Stiefelhagen, R., Yang, J., Waibel, A.: Simultaneous tracking of head poses in a panoramic view. In: Proceedings of the International Conference on Pattern Recognition (2000)
8. Voit, M., Nickel, K., Stiefelhagen, R.: A bayesian approach for multi-view head pose estimation. In: Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI) (2006)
9. Voit, M., Nickel, K., Stiefelhagen, R.: Neural network-based head pose estimation and multi-view fusion. In: Proceedings of the CLEAR Workshop (2006)