A Bayesian Approach for Multi-view Head Pose Estimation

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

***Abstract*— In this paper, we present a system for estimating human head pose with the use of multiple camera views. We apply a neural network to each of the views, and fuse the output using a Bayesian filter framework. Thus, we achieve a more robust estimation compared to pure monocular approaches. The system is evaluated on low resolution seminar video recordings with rather bad lighting, on which the captured head size varies around 20×25 pixels. In total we achieved a correct classification in 39.4% of all frames (one of eight classes). If neighbouring classes were allowed, even 73.4% of the frames were correctly classified.**

I. Introduction

The looking behavior of people gives an important insight on their focus of attention and to whom they are listening. An automatic understanding of observed foci allows machines to perceive their surroundings, analyze scene events and human behavior and detect whether they are being looked at.

This work is embedded in the framework of the European Union research project CHIL (Computers in the Human Interaction Loop). In this project, services are to be developed that provide useful help implicitly and unobtrusively by analyzing and understanding human behavior. We implement these services in a so called smart-room that is equipped with a variety of sensors, such as cameras and microphones. All sensors provide data that is used to capture and detect e.g. speech, gestures and head orientation in order to gather information about the users' current occupation, intention or focus of attention. In order to allow people to move as freely as naturally, it is necessary to obtain data without restricting the users behavior. In case of visual focus of attention tracking, this can be achieved by using multiple cameras that provide captures from different viewpoints, thus making sure that no matter how the user moves, behaves or acts, at least one optimal view can be obtained. Although focus of attention tracking has become an important factor in human computer interaction research, the combination of multiple viewpoints is still a rather unexplored field.

In order to detect peoples' focus of attention, it is necessary to detect where people look at. This can either be achieved by tracking their eye gaze or by an approximation of gaze with detecting their head pose. Head pose estimation is a popular task and many different approaches co-exist to strive for the best possible performance. A big focus thereby resides on appearance-based approaches, that seem to perform satisfactory even in low-resolution camera setups as shown in [1] where neural networks were used to estimate head pose from participants in a meeting, which was captured with a panoramic camera that was placed on the table. In [2], Obodez et al. describe the integration of Gabor and Gaussian filters into a joint particle filter tracking framework. Pappu et al. present in [3] a textural approach, where synthetically created ellipsoidal texture models of a head are used to determine head pose by matching them with live images. And in [4], Tian et al. describe 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. There, also neural networks were implemented for estimating the head pose seen by each single camera. A maximum-likelihood search then results in the final pose hypothesis.

One of the main problems head pose estimating systems such as [2] and [1] mostly suffer from is the restriction of head orientation to a fixed range of angles relative to the capturing camera. In monocular setups, often views depicting head orientations within $-90°$ to $90°$ only are allowed to be classified, hence prohibiting any natural and unrestricted behavior and only allowing 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). Further, allowing people to move around freely, often implies a strong variance in lighting changes as the observation of a person walking around in a room tends to include multiple light sources that affect the brightness of the person's head region more or less. Especially texture-based approaches such as [3] suffer from this circumstance. Therefore, our main goal is to both overcome the problem of poor resolution and strong lighting changes besides allowing to capture head images that depict orientations in the whole pose range of $360°$.

A. Paper Overview

In this paper, we present a system to estimate head pose on low resolution seminar recordings under poor lighting conditions. The video sequences were captured with multiple cameras that were placed in the upper corners of our seminar room. Due to the far distance of the cameras, extracted head regions mostly vary around 20×25 pixels in size. Therefore, the main goal of this work is to integrate the information coming from multiple views in order to stabilize the system's final output and overcome the particular problem of low-resolution head appearances.

Section 2 will give a brief description of the data that was being used. Since we are going to estimate head pose on every single view and combine the single hypotheses into one joint decision about the observed head pose, section 3 gives an overview of the neural network architecture we used for this monocular step. Section 4 then extends the system to use multiple views and combine the single estimates to one joint hypothesis. Section 5 provides a short conclusion.

II. Data Description

The database we used, consists of actual seminars that were also provided for the CLEAR06 evaluation workshop[1]. The videos were recorded with four fixed cameras that are placed in the upper corners of our seminar room (see Fig. 1). The lecturer's head bounding box and head orientation are annotated for every 10th frame in each of the four camera views. In order to gather more data, we interpolated the remaining 9 frames linearly. In total there are 42330 frames to estimate head pose on (for each of the four cameras). Figure 2 depicts one sample video frame from the four cameras. Since the native resolution is $640x480$ pixels, the resolution of annotated head regions is poor (20×25 pixels). Thus, the task is to use multiple views to stabilize the system's output. Since the lecturer's position changes continuously, his or her head is further being exposed to strong lighting changes such as the projector ray, whiteboard illumination or general room illumination. The head orientation is classified into eight discrete classes $\Theta = \{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\}$.

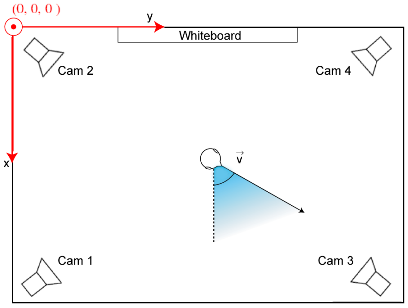

Fig. 1. Setup of our smart room. We installed four fixed, overhead cameras in the upper corners to allow an unintrusive setup of surveillance. The lecturer is allowed to move freely and rotate his or her head without any limitations.

III. Monocular Head Pose Estimation with Neural Networks

Because of the low resolution of the lecturer's head captures, we decided to classify head pose using an appearance based approach. As we described in our previous work [5], neural networks showed to be good classifiers for this task, hence we adopted the idea and trained a network to estimate head orientation relative to the camera's line of sight. Using

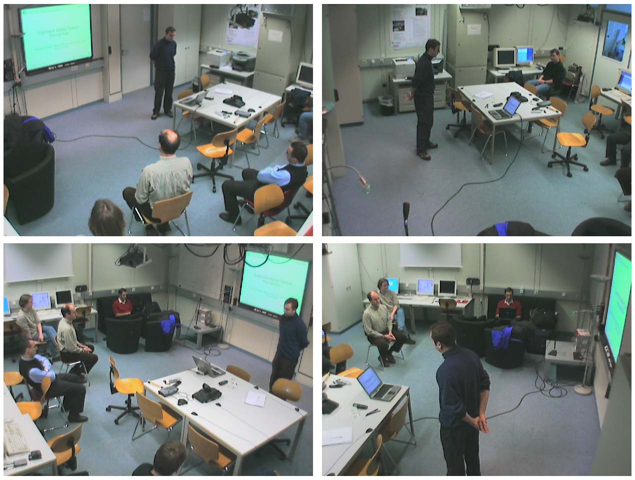

Fig. 2. Example video frame from the CLEAR06 database. The lecturer of the seminar is observed by four fixed, overhead video cameras. In all views, the lecturer's head bounding box and horizontal head orientation is manually annotated.

these relative angles, the very same classifier may be used on each camera views, only a mapping becomes necessary, transforming the relative estimation into an angle value relative to the room's coordinate system.

The network follows a three-layered, feed-forward topology, including 100 hidden neurons in the second layer. As input, the cropped head region is rescaled to an image size of 32×32 pixels, grayscaled and linearly stretched in its contrast to overcome small lighting changes. A Sobel operator is applied to get the magnitude response in both horizontal and vertical derivation. Both images are then concatenated to obtain a feature vector of 2048 dimensions which is fed into the network's input layer (as depicted in Figure 3).

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 100 training cycles.

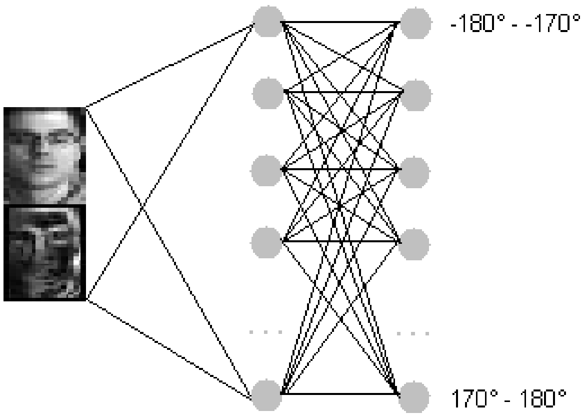

Fig. 3. In the multi-view setup, we trained one neural network with 36 output neurons. Each of them represents one discrete head pose class, relative to the camera's line of view (in $10°$ steps). The network was trained to estimate the class-conditional likelihood of the corresponding output class given the observation of that camera.

[1] http://www.clear-evaluation.org

Instead of using one output neuron that classifies head pose continuously in between $0°$ and $360°$, we used 36 output neurons 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 $+180°$. We used gaussian density functions as target outputs in order to imply a small fuzzification of the very correct class. We did this due to the fact, that with the latter use of estimations from multiple views, the integration of several network outputs into one joint measurement overcomes one single camera's uncertainty.

IV. From Single-view to Multi-view

Although the final decision has to be given within eight output classes of possible head orientations, our state space consists of 360 states $X = \{x_i\}$, with $0° \leq x_i \leq 359°$, where each one represents one individual horizontal head rotation. At each frame t, we can therefore compute a probability distribution over the 360 states by applying the Bayes rule such as

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

given a joint measurement $p(Z_t|x_i)$ that is derived from the four single cameras' hypotheses with observations $Z_t = \{z_{j,t}\}$. The prior probability $P(x_i)$ defines the likelihood to be in state x_i, hence temporal information is implied within this factor.

A. Measurement

In order to gather a combined measurement from the single cameras' hypotheses, we averaged the four class-conditional estimations to one estimate with respect to a given state x_i such that

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

Here, $\phi_j(x_i)$ denotes a mapping from state space X to a description relative to camera j's line of view. 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.

B. Temporal Propagation

Temporal information is implied by the prior probability distribution $P(x_i)$ within the state space. At each frame t this factor implies the probability to observe state x_i, and is derived by the transition probability $p(x_i|x')$ to change into state x_i and the posterior 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 implemented a gaussian kernel function to provide state change propagation $p(x_i|x')$, hence updating the prior distribution can be defined as the convolution of the gaussian kernel and the previous posterior 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 $\sigma = 20°$.

C. Deriving the final estimation

Having computed a probability distribution over 360 possible head rotations, the final step involves classifying head pose within the eight defined classes Θ. We implemented the scoring of each output class θ_l by accumulating the likelihood of all those states that correspond to the same output θ_l. Hence, the final head pose estimate $\hat{\theta}$ can be given as:

$$\hat{\theta} = \arg\max_{\theta_l \in \Theta} \sum_{x_i \in \theta_l} p(x_i|Z_t) \tag{5}$$

V. Experiments & Results

No further head alignment was done, the annotated head bounding boxes were used directly to extract the relevant head region. We evaluated our system under three different conditions:

1) Monocular Estimation: The network was applied to all cameras' frames independently. The output was chosen to be the highest scoring neuron. All outputs were evaluated, no decision fusion took place.
2) Bayes Filter Approach: Head pose is estimated on all four camera views and subsequently combined with our described Bayes filter approach.
3) Maximum Likelihood Approach: Similar to (2), but without temporal filtering, that is the prior distribution is constant for all frames.

As it can be seen in table I, performance increased over twice as much when decision fusion was used in the Bayes framework.

Condition	Correct Class	Correct + neighboring class
(1)	15.8%	38.8%
(2)	39.4%	73.4%
(3)	37.5%	73.4%

TABLE I. The correct classification of head pose increased over twice as much when multiple views were combined in our Bayes filter framework.

We achieved a correct classification in 39.4% of the time. However, even during manual annotations, humans seemed to have problems with choosing either the very correct or neighbouring class for single frames, such that allowing the automatic classification to lie either within the annotated or neighbouring classes the system's performance increased to up to 73.4%. Including temporal smoothing - condition (2) - gained 2% better classification into the correct pose class compared to our fusion scheme that disregards temporal information, as described in condition (3). However, the maximum likelihood integration of multiple views alone clearly shows how information from different positions overcomes the uncertainty of one single view. The accumulation of the network's class-conditional likelihoods over all cameras (see equation 2) scores those states higher where > 2 cameras tend to output similar observations, either within the very same pose estimation or at least neighboring ones. Compared to our previous work [5], this clearly shows that views at the heads' back do not need to be dealt with separately by including an extra facial view classification to highlight best matching camera pairs.

VI. Conclusion

In this work, we presented a system for estimating horizontal head pose on low resolution seminar video recordings from multiple cameras. For each frame, head orientation is hypothesized on every single camera. A subsequently attached Bayes filter framework integrates all hypotheses into one joint measurement and outputs the most likely head orientation. 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 number of cameras and their geometry.

We evaluated our system under three different conditions: (1) we only used the network's output on every camera frame and did no integration into one joint measurement. No Bayes filter was attached, the system was evaluated on every camera frame independently. Given this scenario, the network estimated head pose with a correct classification rate of 15.8%. When allowing the estimation to lie either within the very correct or the neighbouring classes, in 38.8% of the time the correct pose was detected. (2) When both our described Bayes filter is attached on a combined measurement gathered from every camera's estimation and temporal information is used to smooth the system's output, the performance doubled to 39.4% and 73.4% respectively. Ignoring temporal information, as given in condition (3), the correct-class performance dropped to 37.5%.

VII. Acknowledgment

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

References

[1] R. Stiefelhagen, J. Yang, and Alex, "Simultaneous tracking of head poses in a panoramic view," in *International Conference on Pattern Recognition*, vol. 3, September 2000, pp. 726–729.

[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] R. Pappu and P. Beardsley, "A qualitative approach to classifying gaze direction," in *Proceedings of FG98*, 1998, pp. 160–165.

[4] 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.

[5] 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.