A System for Probabilistic Joint 3D Head Tracking and Pose Estimation in Low-resolution, Multi-view Environments

Michael Voit[1] and Rainer Stiefelhagen[12]

[1] Interactive Analysis and Diagnosis, Fraunhofer IITB, Karlsruhe, Germany
[2] Institute for Anthropomatics, Universität Karlsruhe (TH), Karlsruhe, Germany

Abstract We present a new system for 3D head tracking and pose estimation in low-resolution, multi-view environments. Our approach consists of a joint particle filter scheme, that combines head shape evaluation with histograms of oriented gradients and pose estimation by means of artificial neural networks. The joint evaluation resolves previous problems of automatic alignment and multi-sensor fusion and gains an automatic system that is flexible against modifications in the available number of cameras. We evaluate on the CLEAR07 dataset for multi-view head pose estimation and achieve mean pose errors of 7.2° and 9.3° for pan and tilt respectively, which improves accuracy compared to our previous work by 14.9% and 25.8%.

1 Introduction

The automatic visual analysis of human social interaction and individual behavior strongly relies on the recognition of peoples' gaze and its derived visual focus of attention. No less than that, the observation of natural cooperation between people can only be ensured if their environment provides unconstrained work places and does not limit the scope of human actions to restraining sensors. Primary challenges hence originate from compromised sensor setups, as for example camera captures with bad lighting, low resolution or, as in the case of tracking gaze, the problem of rear and profile views of the people under study. Especially the problem of non-frontal observations can however be tackled with camera setups, that allow views from different angles. The low-resolution of far distant face captures however, and the still necessary selection between rear and possibly frontal appearances, still put limitations on every perception. A common approach to approximate gaze and the person's overall attention is therefore to use head orientation. Yet, most challenges also apply here besides the large variety of head appearances, that different poses and hair styles provide.

Official evaluations of multi-view head pose estimation, and with such, a common dataset to compare international systems, was conducted with the CLEAR evaluation workshop in 2006 and 2007. With them, several different approaches to integrate multi-sensor information and enhancements of single-view pose classifiers to fit this new scenario were presented and still provide current state of the art achievements.

Figure 1. Example scene from the CLEAR 2007 dataset for head pose estimation. The dataset provides four camera views from the room's upper corners. From this distance, facial details appear blurred and depending on the used camera view, the captured head region happens to be as small as 20×30 pixels and less.

Canton et al. [8] introduced color descriptors on 3D ellipsoidal shapes to derive likely pose distributions. Although the use of skin color is often used as a very strong feature in visual human-computer interaction tasks, the unrestricted environment challenges its robustness by introducing rear views of people with less or very short hair, whose hair color might even match the previously learned skin color cluster. Furthermore, our own experiments showed, that depending on lighting and background noise, the skin color segmentation is likely to happen on background objects that surround the detected head region, which strongly disbalances the color model of the person under study.

In [7], Ba et al. used a monocular pose classifier, consisting of texture and color analysis, and fused the individual single-view estimates upon the percentage of detected skin color in the respective view's head bounding box. Altough the underlying classification scheme achieves a high accuracy on monocular samples, its disadvantages are the strong dependence on a robust skin color segmentation in every view and the yet missing automatic 3D multi-view alignment.

Another work was presented by Lanz et al. in [12], in which a set of key views of the target is used to compute corresponding color histograms of the head and to generate new poses in between by interpolating from these pre-acquired key models. The training however has to happen on the same video sequence, as individual models have to be created that match the current color distribution of the corresponding person. Furthermore, the groundtruth pose angles during initialization are derived online, by applying frontal face template matching with symmetry analysis to find the best view and hence the observed rotation. With such, the model only allows to estimate in the scope of the initial setup and strongly depends on a robust color description with enough key views to interpolate over the complete set of possible pose angles.

Very good results on the CLEAR07 dataset were achieved by Yan et al., who presented in [10,11] a manifold learning algorithm, based on each individual's set of simplexes with propagated pose labels through the approximated submanifold. The submanifolds of all subjects in the training set are synchronized and new samples are classified by their median of the nearest neighbors in the manifolds' reduced feature space. A sample hereby consists of the concatenated intensities of cropped head image boxes over all camera views and, as above, does not yet include the automatic head alignment but relies on predefined head bounding boxes. Furthermore, the applied fusion on feature level, restricts the system to cope with modified camera setups where new views are added or existing cameras removed.

Considering above limitations and challenges, we now present a new fully-automatic 3D head tracking and pose estimation approach for low-resolution, multi-view environments. In a combined particle filter framework, we rate head appearances with local shape descriptors and estimate pose angles with artificial neural networks. Monocular observation likelihoods are computed for state projections into the respective camera views. The individual probability distributions are then merged by their joint product, which allows the overall system to be flexible for increasing or decreasing the number of available cameras without further retraining. Section 2 thereby describes our framework and both evaluations against pose and shape. Our results are presented in Section 3 where we discuss our testing on the CLEAR07 dataset for head pose estimation and put our experiments in contrast to our previous work. The conclusion in Section 4 then summarizes this paper and gives a short overview of yet unaddressed problems and future plans.

2 Head Tracking and Pose Estimation

We assume that the head's shape can be represented by the quadruple $s = (x, y, z, r_z)$, describing a 3D ellipsoid at position $(x, y, z) \in \mathbb{R}^3$ with radiuses of fixed proportions $r_x = r_y = k \cdot r_z$, and $k, r_z \in \mathbb{R}$. The ellipse's pose is represented by $\theta = (\theta_{pan}, \theta_{tilt})$, to describe rotations in horizontal (pan) and vertical (tilt) direction. In-plane rotation does not influence the overall viewing frustum, which is why we neglect it in this work. Hence, be $X_t \in \{s, \theta\}$ the head's state space at time t, then tracking the head's configuration can be defined as to follow the state evolution $X_t|X_{1:t-1}$, from gathered observations $Z_{1:t}$ up to now. A common Bayesian approach is to recursively calculate the state space density $p(X_t|Z_{1:t}, X_{1:t-1})$, for letting its expectation value $\mathbb{E}[X_t|Z_{1:t}, X_{1:t-1}]$ be the state hypothesis. If we assume, that state evolution is a first order Markovian process, i.e. letting the head's configuration X_t only depend on its predecessor and present measurement, such density is defined by:

$$p(X_t|Z_{1:t}) = \frac{p(Z_t|X_t)p(X_t|Z_{1:t-1})}{p(Z_t|Z_{1:t-1})} \tag{1}$$

with $p(X_t|Z_{1:t-1})$ as the Chapman-Kolmogorov prediction

$$p(X_t|Z_{1:t-1}) = \int p(X_t|X_{t-1})p(X_{t-1}|Z_{1:t-1})dX_{t-1} \tag{2}$$

2.1 Sequential Monte Carlo Sampling

To cope with the non-linear and non-Gaussian nature of the problem, we approximate equation 1, by applying sequential Monte Carlo sampling by means of a particle filter (PF) with sampling importance resampling (SIR) [9].

Following the law of large numbers, particle filters approximate a possibly multimodal target function with a set of support points $\{X_t^i, i = 1, \ldots, N_s\}$ and associated weights $\{\omega_t^i, i = 1, \ldots, N_s\}$. The sum of Dirac functions over this set of weighted samples then results in a discrete representation of the underlying signal, in our case the state space's probability distribution function (PDF):

$$p(X_t|Z_{1:t}) \approx \sum_{i=1}^{N_s} \omega_t^i \cdot \delta(X_t - X_t^i) \tag{3}$$

With this, a configuration hypothesis is easily obtained from the weighted mean of the support set:

$$\mathbb{E}[X_t|Z_{1:t}] = \sum_{i=1}^{N_s} \omega_t^i \cdot X_t^i \tag{4}$$

To propagate state evolution, the set of particles is updated as soon as a new measurement is available. New samples are drawn according to a predefined proposal distribution $X_t \sim q(X_t|X_{t-1}, Z_{1:t})$, which suggests samples in interesting regions of the state space. A corresponding weight update is then successively calculated with:

$$\omega_t^i \propto \omega_{t-1}^i \frac{p(Z_t|X_t^i)p(X_t^i|X_{t-1}^i)}{q(X_t^i|X_{t-1}^i, Z_t)} \tag{5}$$

The SIR filter scheme is applied to cope with these updates but provides two advantages: i) it sets the proposal function $q(\cdot)$ to the often applied and implicitly available prior PDF $p(X_t|X_{t-1}^i)$ and ii) instead of simply drawing new samples and propagating their old weights over time, it resamples and replaces the set of support points with every update step. This sets the particle priors to a uniform distribution with $\omega_{t-1}^i = N_s^{-1}$ and allows us to simplify equation 5 to a less complex reweighing scheme, which only depends on the observation likelihoods:

$$\omega_t^i \propto p(Z_t|X_t^i) \tag{6}$$

Resampling is usually done by drawing new samples from the current set of support points with probabilities according to their corresponding weights. These weights are then updated with respect to the given observation, as the

particles' respective state hypotheses are propagated along a known process motion model, including explicitly added noise, that is to cope with observation variances. The applied motion model filters state evolution for that diffused particles do not persist to local maxima of observation likelihoods and less particles become necessary for a matching approximation of the underlying signal. This is especially useful for complex motion patterns in high dimensional state spaces. However, since we only track the head's 3D position and rotation along two axes, we neglect state motion and only add Gaussian noise on top of the state vector.

2.2 Evaluating Pose Observations

To make use of the multi-view environment, our goal is to merge individual estimates from every camera in a combined framework, instead of applying a best view selection beforehand that introduces a further error source. This can easily be achieved with training the classifier to estimate pose angles with respect to the particular camera's line of sight. However, applying the same classifier on different views, introduces rear captures of the tracked head, as well as large distances to the observing camera, depending on the trajectory of the person under study. The trained system needs to cope with possibly low-resolution head captures and noise induced from different hair styles. It therefore has to show strong generalization capabilities and should allow to measure its confidence along with its given hypothesis. In [6], we showed that classifying artificial neural networks (ANN) sufficed all these conditions. We trained two feed-forward networks with three layers, to respectively estimate a PDF over a set of possible pose angles. One network was applied for pan, another for tilt angle classification. The class width was chosen to span over $10°$, hence leading to 36 output neurons for an estimate over $-180° - +180°$ horizontally, and 18 output neurons for $-90° - +90°$ in vertical direction. By using a Gaussian PDF as target output during training, we implied uncertainty concerning neighbor angle classes. In our experiments this has shown to enhance robustness when applying the same network on different camera views: unimodal estimates, that peaked over wrong angles, could still provide enough support, if the actual class lied in the near neighborhood. A joint PDF over all obtained network outputs hence still managed to average its peak near the correct target angle.

We therefore applied this classifier in our current framework as follows: With Z_t^c the observation of camera c, be $Z_t^c(s_t^i)$ the cropped image patch, which is obtained from projecting the 3D bounding box around state estimate X_t^i's ellipsoid into camera view c. This 2D region is preprocessed, for that it is resampled to a fixed size of 32×32 pixels in order to provide invariance to different head sizes and observation distances. After resampling, the head region is grayscaled and equalized in intensity to deliver a better contrast. The result is concatenated with its 3×3 Sobel magnitude response, to provide a vectorized feature vector $\Upsilon(Z_t^c(s_t^i))$, consisting of the head intensity and edge appearances. The ANN then estimates the posterior PDF over the set of relative angles θ_{pan}^c or θ_{tilt}^c, for horizontal or vertical rotations. Since we consider both angles to be conditionally independent, we can build a joint posterior with:

$$p^{pose}(\theta^{i,c}|\Upsilon(Z_t^c(s_t^i))) = p^{ann}(\theta_{pan}^{i,c}|\Upsilon(Z_t^c(s_t^i)))p^{ann}(\theta_{tilt}^{i,c}|\Upsilon(Z_t^c(s_t^i))) \quad (7)$$

If we assume a uniform distribution of training samples per pose class, then following Bayes' rule gains an observation evaluation, proportional to the ANNs' estimated posterior:

$$p^{pose}(Z_t^c|X_t^i) = p^{pose}(Z_t^c|s_t^i, \theta_t^c) \propto p^{pose}(\theta^{i,c}|\Upsilon(Z_t^c(s_t^i))) \quad (8)$$

2.3 Evaluating Head Alignment with Local Shape Descriptors

The nature of the trained ANNs is to output pose likelihoods for any given image patch along with strong generalization capabilities. This makes the estimates only as reliable as the implicit head alignment, that is to crop 2D image head regions consistent with used training samples. To gain a measurement for the fitness of an aligned head region and with such, confidence in the ANN's estimates, we use local shape descriptors by means of histograms of oriented gradients (HOG) as a second evaluation of a state hypothesis.

HOGs were presented by Dalal and Triggs in [1] and describe an appearance-based approach to represent an object's shape by merging local histograms of its binned edge orientations into a joint feature vector. The histograms are computed for non-overlapping subwindows of fixed size, to cover the object's image patch in total. The concatenation of histograms of neighboring subwindows then gives the final description. In the original work, a support vector machine then discriminatively detects learned objects.

We adopt the descriptor and directly rate possible head appearances against a given head model by means of the l^2 norm. An initial mean HOG representation of heads is computed over a set of training samples. This model is then gradually adapted online with sparse head detector hits from a custom-trained Haar-like feature cascade using OpenCV's implementation [2,3]. A general preprocessing thereby happens similar to 2.2: a given image patch $Z_t^c(s_t^i)$ is resampled to a fixed width and height, grayscaled and equalized for a better contrast. Its HOG descriptor is then computed on its 3×3 Sobel filter response. We obtained satisfactory results, by scaling image patches to 24×24 pixels, using 8 bins for discretizing orientations into 45°-wide segments and concatenating the final descriptor over 3×3 neighboring histograms.

With defining $\Gamma(Z_t^c(s_t^i))$ to be the respective HOG descriptor of the image patch corresponding to hypothesis s_i, and $\widehat{\Gamma}$ a mean shape model that we computed upon training samples, a corresponding observation likelihood equates to the similarity of the two vectors:

$$p^{shape}(Z_t^c|X_t^i) = p^{shape}(Z_t^c|s_t^i) = p^{shape}(\Gamma(Z_t^c(s_t^i))|\widehat{\Gamma}) \propto \lambda exp - \lambda|\Gamma(Z_t^c(s_t^i)) - \widehat{\Gamma}| \quad (9)$$

The parameter λ defines how strong the fitness converges against zero and was empirically set to 0.25.

Figure 2. Histograms of oriented gradients for an exemplary head image patch. The head region was grayscaled and resampled to fit a width and height of 24×24 pixels (left). A division into 8×8 pixel sized subregions (middle) then defines the areas for which individual histograms over the edge orientations are computed (right). Each line in the right image, resembles an orientation bin in that local histogram. The brighter the line, the stronger that edge orientation is supported.

2.4 Joining Multi-Sensor Observation Likelihoods

Using two or more cameras implies a set of multi-view sensor stream observations denoted by $Z_t = \{Z_t^c, c = 1, \ldots, N_C\}$, with pose and shape observations $Z_t^c = \{\Upsilon(\cdot), \Gamma(\cdot)\}$. Both evaluations estimate likelihoods invariant to the used camera view. To cope with multiple streams, a common acceptance in Bayesian tracking is therefore to build the product density over the set of all single-view likelihoods [12]. Considering that observations for pose and shape evaluations are conditionally independent in every camera view, we can therefore build a final joint observation PDF with:

$$p(Z_t|X_t^i) = \prod_{c=1}^{N_C} p^{pose}(Z_t^c|X_t^i) p^{shape}(Z_t^c|X_t^i) \tag{10}$$

3 Experimental Validation

To allow a comparison with current state of the art systems, we evaluated our approach on the CLEAR07 dataset for head pose estimation [5].

Provided are four camera views from a room's upper corners with 640×480 pixels at 15 frames per second. The dataset contains 15 different persons, whose head poses were captured with a magnetic pose sensor [4] and who showed rotations over the whole range of angle classes. We stayed consistent with training our system on the same video subset, that was officially used during the workshop. To directly distinguish between head detection and pose estimation tasks, manually annotated head bounding boxes for every 5th frame in all camera views are included in the dataset, which automatically lets us assess implied head tracking and alignment. Evaluations only take place on these dedicated selected frames.

With each evaluation video, we initialized our system to randomly spread particles in a spherical radius of 20 cm off the true head position in 3D. We considered this to be a valid assumption for state-of-the-art body tracker and as it showed, our particles converged very fast onto the real head nearby. Aside from initialization, the remaining videos were processed fully automatic. With this, the implicit tracking only showed a mean absolute difference of only 2 cm compared to the annotated head centroids. As can be seen in Table 1, for the

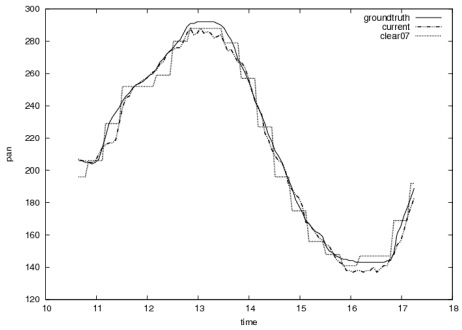

Figure 3. Pan rotation over 7 seconds for the sequence depicted in Figure 1. Shown is the groundtruth head rotation, estimates we gained with our previous system from CLEAR07 [6] and results from this work. As can be seen, the previous system only estimated the pan angle on every 5th frame, because only then annotated head bounding boxes were available within the dataset. Our current implementation includes automatic tracking and alignment, hence we achieve estimates on every frame.

individual camera views, this resulted in a mean difference of 3.6 pixels for the projected 2D head box centers. Their width and height fit with a mean error of 4.0 and 3.9 pixels respectively.

Table 1. Mean and standard deviation of the distance of our hypotheses to the annotated ground truth.

	μ	σ
3D head box centroid [cm]	2.0	0.7
2D head box centroid [px]	3.6	2.0
2D head box width [px]	4.0	3.0
2D head box height [px]	3.9	2.8

Considering head poses, we observed mean errors of 7.2° and 9.3° for pan and tilt estimation. In CLEAR07, we presented a system that simply used the included head box annotations directly, instead of providing a custom tracking and alignment scheme [6]. Table 2 shows both systems' results in contrast. The overall decrease in error by 14.9% and 25.8% thereby emphasizes the advantages of a joint position and orientation state filtering. With such, we experienced that observation likelihoods were sometimes maximized by decreasing ellipse sizes, so that projected 2D image patches concentrated on face regions instead of the whole head with its hair and further background noise.

Furthermore, the joint tracking of both rotation angles in one state, mostly helps to resolve ambiguities that e.g. arise with poses that show downwards

Table 2. Results on head pose estimation, compared to our previous system.

	$\mu_{pan}[^\circ]$	$\mu_{tilt}[^\circ]$
voit07 [6]	8.46	12.54
voit09	7.2	9.3
Δ	14.9%	25.8%

tilting. Here, views from above mostly depict hair in every view, which strongly resembles rear or even profile shots where parts of the face are occluded from long hairstyles. Smoothing the joint pose trajectory over time hereby restricts unlikely and sudden rotations in either direction and removes ambiguities that come along with them.

4 Conclusion

In this paper, we presented a new approach for tracking the head's position and rotation in a 3D environment with multiple low-resolution cameras. We implemented a particle filter framework, that uses artificial neural networks for pose estimation and head shape evaluations by means of histograms of oriented gradients in a joint sampling importance resampling scheme.

We evaluated our implementation on the CLEAR07 multi-view subset for head pose estimation and obtained 7.2° and 9.3° mean errors regarding pan and tilt estimation. In contrast to our previous system we presented for CLEAR07 [6], which respectively showed errors of 8.46° and 12.54°, this approach jointly models the head's position, size and orientation in single state hypotheses. Besides the hereby gained full-automatic head tracking and alignment, both in 3D as well as 2D camera views, this tight integration increased the overall accuracy by 14.9% and 25.8% in horizontal and vertical direction. A comparison to further state of the art results on the same dataset can thereby be found in [5].

Still unaddressed problems in this work include the normalization of camera view distortions and coping with in-plane rotations for a more robust recognition. Since head appearances not only vary with respect to pose and person, but also to the distance of the observing camera, observations from above result in differences due to their high viewing angle the nearer a person gets. Yet other enhancements can be found in using different shape models for individual pose classes, hence obtaining a coarse estimate of head orientations implicitly from the alignment observation, and successively applying neural networks for pose refinement only.

5 Acknowledgments

This work was supported by the FhG Internal Programs under Grant No. 692 026.

References

1. Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection. International Conference on Computer Vision and Pattern Recognition. Proceedings of IEEE Conference Computer Vision and Pattern Recognition, San Diego, USA, 886 – 893 (2005)
2. Viola, P. and Jones, M.: Robust Real-time Object Detection. International Journal of Computer Vision (2001)
3. OpenCV Library: http://sourceforge.net/projects/opencvlibrary
4. Ascension Technology Corporation. http://www.ascension-tech.com/
5. Stiefelhagen, R., Bernardin, K., Bowers, R., Travis, R., Michel, M. and Garofolo, J.: The CLEAR 2007 Evaluation. Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007 (2007)
6. Voit, M., Nickel, K. and Stiefelhagen, R.: Head Pose Estimation in Single- and Multi-view Environments - Results on the CLEAR'07 Benchmarks. Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007 (2007)
7. Ba, S.O. and Odobez, J.-M.: A Probabilistic Head Pose Tracking Evaluation in Single and Multiple Camera Setups. Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007 (2007)
8. Canton-Ferrer, C., Casas, J.R. and Pards, M.: Head Orientation Estimation using Particle Filtering in Multiview Scenarios. Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007 (2007)
9. Arulampalam, S., Maskell, S., Gordon, N. and Clapp, T.: A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. IEEE Transactions on Signal Processing **50** 174–188 (2002)
10. S. Yan, Z. Zhang, Y. Hu, J. Tu and T. Huang: Learning a Person-Independent Representation for Precise 3D Pose Estimation. Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007 (2007)
11. S. Yan, H. Wang, Y. Fu, J. Yan, X. Tang, and T. Huang: Synchronized Submanifold Embedding for Person-Independent Pose Estimation and Beyond. IEEE Transactions on Image Processing (TIP), 18(1):202-210, 2009.
12. Lanz, O. and Brunelli, R.: Joint Bayesian Tracking of Head Location and Pose from Low-Resolution Video. Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007 (2007)