

Head Orientation Estimation using Particle Filtering in Multiview Scenarios

C.Canton-Ferrer, J.R. Casas, and M.Pardàs

Technical University of Catalonia, Barcelona, Spain,
{ccanton,josep,montse}@gps.tsc.upc.es

Abstract. This paper presents a novel approach to the problem of determining head pose estimation and face 3D orientation of several people in low resolution sequences from multiple calibrated cameras. Spatial redundancy is exploited and the head in the scene is approximated by an ellipsoid. Skin patches from each detected head are located in each camera view. Data fusion is performed by back-projecting skin patches from single images onto the estimated 3D head model, thus providing a synthetic reconstruction of the head appearance. A particle filter is employed to perform the estimation of the head pan angle of the person under study. A likelihood function based on the face appearance is introduced. Experimental results proving the effectiveness of the proposed algorithm are provided for the SmartRoom scenario of the CLEAR Evaluation 2007 Head Orientation dataset.

1 Introduction

The estimation of human head orientation has a wide range of applications, including a variety of services in human-computer interfaces, teleconferencing, virtual reality and 3D audio rendering. In recent years, significant research efforts have been devoted to the development of human-computer interfaces in intelligent environments aiming at supporting humans in various tasks and situations. Examples of these intelligent environments include the “digital office” [3], “intelligent house”, “intelligent classroom” and “smart conferencing rooms”. The head orientation of a person provides important clues in order to construct perceptive capabilities in such scenarios. This knowledge allows a better understanding of what users do or what they refer to. Furthermore, accurate head pose estimation allows the computers to perform Face Identification or improved Automatic Speech Recognition by selecting a subset of sensors (cameras and microphones) adequately located for the task. Being focus of attention directly related to the head orientation, it can also be used to give personalized information to the users, for instance through a monitor or a beamer displaying text or images directly targeting their focus of attention. In synthesis, determining the individuals head orientation is the basis for many forms of more sophisticated interactions between humans and technical devices. In automatic video conferencing, a set of computer-controlled cameras capture the images of one or more individuals adjusting for orientation and range, and compensating for any source motion [22].

In this context, head orientation estimation is a crucial source of information to decide which cameras and microphones are more suited to capture the scene. In video surveillance applications, determination of the head orientation of the individuals can also be used for camera selection. Other applications include control of avatars in virtual environments or input to a cross-talk cancellation system for 3D audio rendering.

Methods for head pose estimation from video signals proposed in the literature can be classified as feature based or appearance based [21]. Feature based methods [2, 14, 17] use a general approach that involves estimating the position of specific facial features in the image (typically eyes, nostrils and mouth) and then fitting these data to a head model. In practice, some of these methods might require manual initialization and are particularly sensitive to the selection of feature points. Moreover, near-frontal views are assumed and high-quality images are required. For the applications addressed in our work, such conditions are usually difficult to satisfy. Specific facial features are typically not clearly visible due to lighting conditions and wide angle camera views. They may also be entirely unavailable when faces are not oriented towards the cameras. Methods which rely on a detailed feature analysis followed by head model fitting would fail under these circumstances. Furthermore, most of these approaches are based on monocular analysis of images but few have addressed the multiocular case for face or head analysis [4, 6, 17]. On the contrary, appearance based methods [20, 24] tend to achieve satisfactory results with low resolution images. However, in these techniques head orientation estimation is posed as a classification problem using Neural Networks, thus producing an output angle resolution limited to a discrete set. For example, in [19] angle estimation is restricted to steps of 25° while in [9] steps of 45° are employed. Data analysis methods providing a real valued angle output will be preferred since produce more information for further analysis modules. Recently, a multimodal head orientation estimation algorithm has been presented by the authors [5] based on the output of this video analysis system.

The remainder of this paper is organized as follows. In the next section, we introduce the Particle Filtering theory that will be the basis of the head orientation estimation technique. In Section 3, the monomodal video head estimation algorithm is presented. Finally, in the following section, the performance obtained by our system is discussed.

2 Particle Filtering for Head Orientation Estimation

The estimation of the pan angle θ_t of the head of a person at a given time t given a set of observations $\Omega_{1:t}$ can be written in the context of a state space estimation problem [23] driven by the following state process equation:

$$\theta_t = \mathbf{f}(\theta_{t-1}, \mathbf{v}_t), \quad (1)$$

and the observation equation:

$$\Omega_t = \mathbf{h}(\theta_t, \mathbf{n}_t), \quad (2)$$

where $\mathbf{f}(\cdot)$ is a function describing the evolution of the model and $\mathbf{h}(\cdot)$ an observation function modeling the relation between the hidden variable θ_t and its measurable magnitude Ω_t . Noise components, \mathbf{v}_t and \mathbf{n}_t , are assumed to be independent stochastic processes with a given distribution.

From a Bayesian perspective, the pan angle estimation and tracking problem is to recursively estimate a certain degree of belief in the state variable θ_t at time t , given the data $\Omega_{1:t}$ up to time t . Thus, it is required to calculate the *pdf* $p(\theta_t|\Omega_{1:t})$, and this can be done recursively in two steps, namely prediction and update. The prediction step uses the process equation Eq.1 to obtain the prior *pdf* by means of the Chapman-Kolmogorov integral:

$$p(\theta_t|\Omega_{1:t-1}) = \int p(\theta_t|\theta_{t-1})p(\theta_{t-1}|\Omega_{1:t-1})d\theta_{t-1}, \quad (3)$$

with $p(\theta_{t-1}|\Omega_{1:t-1})$ known from the previous iteration and $p(\theta_t|\theta_{t-1})$ determined by Eq.1. When a measurement Ω_t becomes available, it may be used to update the prior *pdf* via Bayes' rule:

$$p(\theta_t|\Omega_{1:t}) = \frac{p(\Omega_t|\theta_t)p(\theta_t|\Omega_{1:t-1})}{\int p(\Omega_t|\theta_t)p(\theta_t|\Omega_{1:t-1})d\theta_t}, \quad (4)$$

being $p(\Omega_t|\theta_t)$ the likelihood statistics derived from Eq.2. However, the posterior *pdf* $p(\theta_t|\Omega_{1:t})$ in Eq.4 can not be computed analytically unless linear-Gaussian models are adopted, in which case the Kalman filter provides the optimal solution.

Particle Filtering (PF) [1] algorithms are sequential Monte Carlo methods based on point mass (or "particle") representations of probability densities. These techniques are employed to tackle estimation and tracking problems where the variables involved do not hold Gaussianity uncertainty models and linear dynamics. In this case, PF approximates the posterior density $p(\theta_t|\Omega_{1:t})$ with a sum of N_s Dirac functions centered in $\{\theta_t^j\}$, $0 < j \leq N_s$, as:

$$p(\theta_t|\Omega_{1:t}) \approx \sum_{j=1}^{N_s} w_t^j \delta(\theta_t - \theta_t^j), \quad (5)$$

where w_t^j are the weights associated to the particles fulfilling $\sum_{j=1}^{N_s} w_t^j = 1$. For this type of estimation and tracking problems, it is a common approach to employ a Sampling Importance Re-sampling (SIR) strategy to drive particles across time [12]. This assumption leads to a recursive update of the weights as:

$$w_t^j \propto w_{t-1}^j p(\Omega_t|\theta_t^j). \quad (6)$$

SIR PF circumvents the particle degeneracy problem by re-sampling with replacement at every time step [1], that is to dismiss the particles with lower weights and proportionally replicate those with higher weights. In this case, weights are set to $w_{t-1}^j = N_s^{-1}$, $\forall j$, therefore

$$w_t^j \propto p(\Omega_t|\theta_t^j). \quad (7)$$

Hence, the weights are proportional to the likelihood function that will be computed over the incoming data Ω_t . The re-sampling step derives the particles depending on the weights of the previous step, then all the new particles receive a starting weight equal to N_s^{-1} that will be updated by the next likelihood evaluation.

The best state at time t , Θ_t , is derived based on the discrete approximation of Eq.5. The most common solution is the Monte Carlo approximation of the expectation

$$\Theta_t = \mathbb{E}[\theta_t | \Omega_{1:t}] \approx \sum_{j=1}^{N_s} w_t^j \theta_t^j. \quad (8)$$

Finally, a propagation model is adopted to add a drift to the angles θ_t^j of the re-sampled particles in order to progressively sample the state space in the following iterations [1]. For complex PF problems involving a high dimensional state space such as in articulated human body tracking tasks [10], an underlying motion pattern is employed in order to efficiently sample the state space thus reducing the number of particles required. Due to the single dimension of our head pose estimation task, a Gaussian drift is employed and no motion models are assumed.

PF have been successfully applied for a number of tasks in both audio and video such as object tracking tasks with cluttered backgrounds [15]. Information of audio and video sources have been effectively combined employing PF strategies for active speaker tracking [18] or audiovisual multi-person tracking [11].

3 Video Head Pose Estimation

This section presents a new approach to multi-camera head pose estimation from low-resolution images based on PF. A spatial and color analysis of these input images is performed and redundancy among cameras is exploited to produce a synthetic reconstruction of the head of the person. This information will be used to construct the likelihood function that will weight the particles of this PF based on visual information. The estimation of the head orientation will be computed as the expectation of the pan angle thus producing a real valued output which will increase the precision of our system as compared with classification approaches.

For a given frame in the video sequence, a set of N images are obtained from the N cameras. Each camera is modeled using a pinhole camera model based on perspective projection. Accurate calibration information is available. Bounding boxes describing the head of a person in multiple views are used to segment the interest area where the colour module will be applied. Center and size of the bounding box allow defining an ellipsoid model $\mathcal{H} = \{\mathbf{c}, \mathbf{R}, \mathbf{s}\}$ where \mathbf{c} is the center, \mathbf{R} the rotation along each axis centered on \mathbf{c} and \mathbf{s} the length of each axis. Colour information is processed as described in the following subsection.

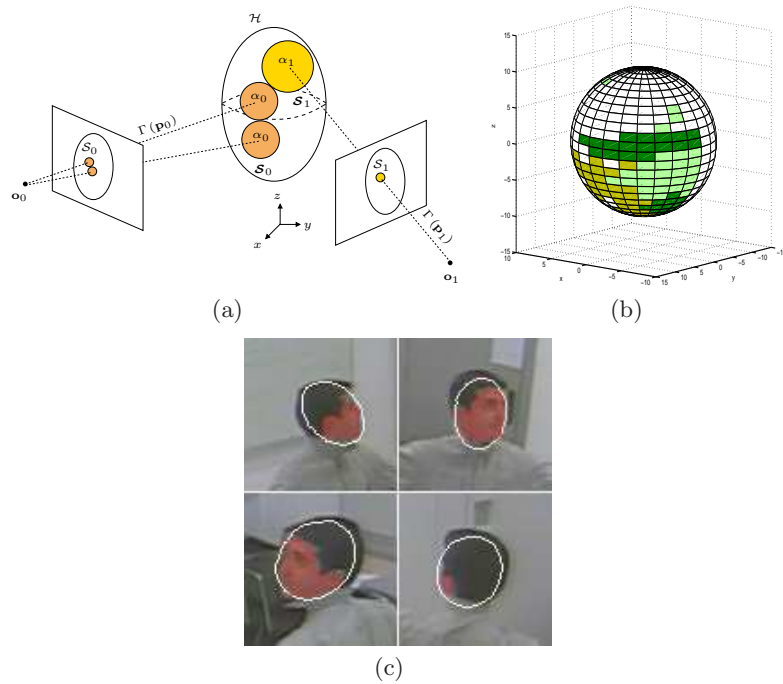


Fig. 1. In (a), color and spatial information fusion process scheme. Pixels in the set \mathcal{S}_n are back-projected onto the surface of the ellipsoid defined by \mathcal{H} , generating the set \mathcal{S} with its weighting term α_n . In (b), result of information fusion obtaining a synthetic reconstruction of face appearance from images in (c) where the skin patches are plot in red and the ellipsoid fitting in white.

3.1 Color Analysis

Interest regions provided as a bounding box around the head provide 2D masks within the original images where skin color pixels are sought. In order to extract skin color like pixels, a probabilistic classification is computed on the RGB information [16] where the color distribution of skin is estimated from offline hand selected samples of skin pixels.

Let us denote with \mathcal{S}_n all skin pixels in the n -th view. It should be recalled that there could be empty sets \mathcal{S}_n due to occlusions or under-performance of the skin detection technique. However, tracking information and redundancy among views would allow to overcome this problem.

3.2 3D Head Appearance Generation

Combination of both color and space information is required in order to perform a high semantic level classification and estimation of head orientation. Our information aggregation procedure takes as input the information generated from

the low level image analysis for the person under study: an ellipsoid estimation \mathcal{H} of the head and a set of skin patches at each view belonging to this head $\{\mathcal{S}_n\}$, $0 \leq n < N_{\text{CAM}}$. The output of this technique is a fusion of color and space information set denoted as \mathcal{T} .

The procedure of information aggregation we define is based on the assumption that all skin patches \mathcal{S}_n are projections of a region of the surface of the estimated ellipsoid defining the head of a person. Hence, color and space information can be combined to produce a synthetic reconstruction of the head and face appearance in 3D. This fusion process is performed for each head separately starting by back-projecting the skin pixels of \mathcal{S}_n from all N_{CAM} views onto the 3D ellipsoid model. Formally, for each pixel $p_n \in \mathcal{S}_n$, we compute

$$\Gamma(p_n) \equiv P_n^{-1}(p_n) = \mathbf{o}_n + \lambda \mathbf{v}, \quad \lambda \in \mathbb{R}^+, \quad (9)$$

thus obtaining its back-projected ray in the world coordinate frame passing through p_n in the image plane with origin in the camera center \mathbf{o}_n and director vector \mathbf{v} . In order to obtain the back-projection of p_n onto the surface of the ellipsoid modelling the head, Eq.9 is substituted into the equation of an ellipsoid defined by the set of parameters \mathcal{H} [13]. It gives a quadratic in λ ,

$$a\lambda^2 + b\lambda + c = 0. \quad (10)$$

The case of interest will be when Eq.10 has two real roots. That means that the ray intersects the ellipsoid twice in which case the solution with the smaller value of λ will be chosen for reasons of visibility consistency. See a scheme of this process on Fig.1(a).

This process is applied to all pixels of a given patch \mathcal{S}_n obtaining a set \mathcal{S}_n containing the 3D points being the intersections of the back-projected skin pixels in the view n with the ellipsoid surface. In order to perform a joint analysis of the sets $\{\mathcal{S}_n\}$, each set must have an associated weighting factor that takes into account the real surface of the ellipsoid represented by a single pixel in that view n . That is, to quantize the effect of the different distances from the center of the object to each camera. This weighting factor α_n can be estimated by projecting a sphere with radius $r = \max(\mathbf{s})$ on every camera plane, and computing the ratio between the appearance area of the sphere and the number of projected pixels. To be precise, α_n should be estimated for each element in \mathcal{S}_n but, since the *far-field* condition

$$\max(\mathbf{s}) \ll \|\mathbf{c} - \mathbf{o}_n\|_2, \quad \forall n, \quad (11)$$

is fulfilled, α_n can be considered constant for all intersections in \mathcal{S}_n . A schematic representation of the fusion procedure is depicted in Fig.1(a). Finally, after applying this process to all skin patches we obtain a fusion of color and spatial information set $\mathcal{T} = \{\mathcal{S}_n, \alpha_n, \mathcal{H}\}$, $0 \leq n < N_{\text{CAM}}$, for the head of the person under study. A result of this process is shown in Fig.1(b).

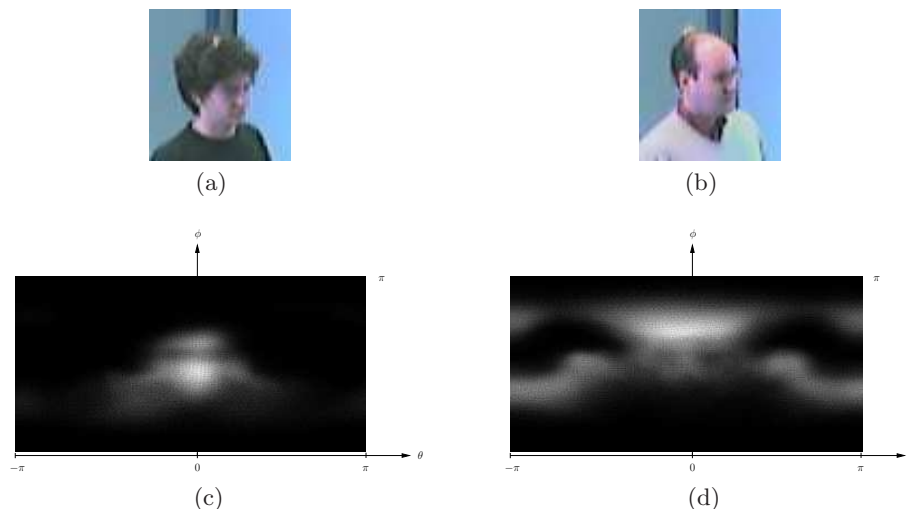


Fig. 2. Two examples of the Ω_t^V sets containing the visual information that will be fed to the video PF. This set may take different configurations depending on the appearance of the head of the person under study. For our experiments a quantization step of $\Delta_\theta \times \Delta_\phi = 0.02 \times 0.02$ rads have been employed.

3.3 Head Pose Video Likelihood Evaluation

In order to implement a PF that takes into account visual information solely, the visual likelihood evaluation function must be defined. The observation Ω_t^V will be constructed upon the information provided by the set \mathcal{Y} . The sets \mathcal{S}_n containing the 3D Euclidean coordinates of the ray-ellipsoid intersections are transformed on the plane $\theta\phi$, in elliptical coordinates with origin at \mathbf{c} , describing the surface of \mathcal{H} . Every intersection has associated its weight factor α_n and the whole set of transformed intersections is quantized with a 2D quantization step of size $\Delta_\theta \times \Delta_\phi$. This process produces the visual observation $\Omega_t^V(n_\theta, n_\phi)$ that might be understood as a *face map* providing a planar representation of the appearance of the head of the person. Some examples of this representation are depicted in Fig.2.

Groundtruth information from a training database is employed to compute an average normalized *template face map* centered at $\theta = 0$, namely $\tilde{\Omega}^V(n_\theta, n_\phi)$. That is, the appearance that the head of a person would have if there were no distorting factors (bad performance of the skin detector, not enough cameras seeing the face of the person, etc.). This information will be employed to define the likelihood function. The computed template face map is shown in Fig.3.

A cost function is defined as a sum-squared difference function $\Sigma^V(\theta, \Omega^V(n_\theta, n_\phi))$ and is computed using

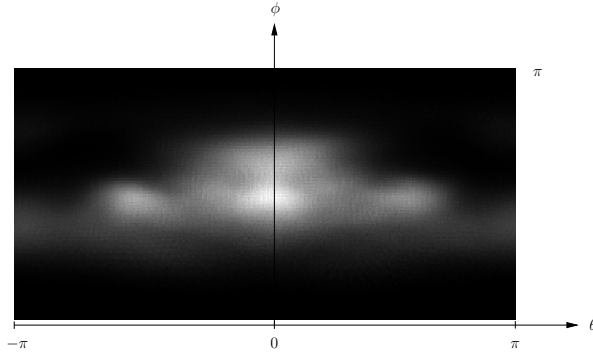


Fig. 3. Template face map obtained from an annotated training database for 10 different subjects.

$$\Sigma^V(\theta, \Omega^V(n_\theta, n_\phi)) = \quad (12)$$

$$= \sum_{k_\theta=0}^{N_\theta} \sum_{k_\phi=0}^{N_\phi} \left(1 - \left(\Omega^V(k_\theta, k_\phi) \cdot \tilde{\Omega}^V\left(k_\theta \ominus \left\lfloor \frac{\theta}{\Delta_\theta} \right\rfloor, k_\phi\right) \right)^2 \right),$$

$$N_\theta = \left\lfloor \frac{2\pi}{\Delta_\theta} \right\rfloor, \quad N_\phi = \left\lfloor \frac{\pi}{\Delta_\phi} \right\rfloor, \quad (13)$$

where \ominus is the circular shift operator. This function will produce small values when the value of the pan angle hypothesis θ matches the angle of the head that produced the visual observation $\Omega^V(n_\theta, n_\phi)$. Finally, the weights of the particles are defined as

$$w_t^j(\theta_t^j, \Omega^V(n_\theta, n_\phi)) = \exp\left(-\beta_V \Sigma^V(\theta_t^j, \Omega^V(n_\theta, n_\phi))\right). \quad (14)$$

Inverse exponential functions are used in PF applications in order to reflect the assumption that measurement errors are Gaussian [15]. It also has the advantage that even weak hypotheses have finite probability of being preserved, which is desirable in the case of very sparse samples [10]. The value of β_V is not critical, and it has been empirically fixed at $\beta_V = 50$ to allow some useful bias towards lower cost solutions.

4 Results and Conclusions

In order to evaluate the performance of the proposed algorithms, we employed the CLEAR 2007 head pose database containing a set of scenes in an indoor scenario where a person is moving his/her head. 10 video segments of approximately 2 minutes each are used for the training phase and 5 segments are used

for the testing. The analysis sequences were recorded with 4 fully calibrated and synchronized cameras with a resolution of 640x480 pixels at 25 fps. Head localization is assumed to be available since the aim of our research is at estimating its orientation. Groundtruth information of the pan/tilt/roll angles of the head is available for the evaluation of the algorithm. However, since our algorithm performs an estimation of the pan angle, only result of this magnitude will be reported.

For all the experiments conducted in this article, a fixed number of particles has been set, $N_s = 100$. Experimental results proved that employing more particles does not report in a better performance of the system. Results reported in Table 1 quantitatively prove the effectiveness of the presented method at estimating the pan angle orientation of the head of a person.

Table 1. Results of the proposed method for the CLEAR Head Pose Testing Database.

	Pan Mean Error
person01b	22.91°
person02	31.41°
person03	17.40°
person04	11.83°
person05	17.22°
Average	20.48°

Orientation estimation depends on the detection of skin patches thus being sensitive to its performance. Typically, skin detection underperforms when the face is being illuminated by a coloured light, i.e. a beamer. Other effects to be considered are the hair style, the presence of beard or baldness. Nevertheless, the proposed particle filter strategy is able to cope with such effects in most of the cases.

Future work aims at continuing the multimodal head orientation approach already presented by the authors in [5]. Focus of attention in multi-person meetings based on the information retrieved from head orientation estimation of multiple people is under study.

References

1. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. on Signal Processing*, vol. 50:2, pp. 174–188, 2002.

2. Ballard, P., Stockman, G.C.: Controlling a computer via facial aspect. *IEEE Trans. on Systems, Man and Cybernetics*, vol. 25:4, pp. 669–677, 1995.
3. Black, M., Brard, F., Jepson, A., Newman, W., Saund, W., Socher, G., Taylor, M.: The Digital Office: Overview. *Proc. Spring Symposium on Intelligent Environments*, pp. 98–102, 1998. 72
4. Canton-Ferrer, C., Casas, J. R., Pardàs, M.: Fusion of Multiple Viewpoint Information Towards 3D Face Robust Orientation Detection. *Proc. IEEE Int. Conf. on Image Processing*, vol. 2, pp. 366–369, 2005.
5. Canton-Ferrer, C., Segura, C., Casas, J.R, Pardàs, M., Hernando, J.: Audiovisual Head Orientation Estimation with Particle Filters in Multisensor Scenarios. *EURASIP Journal on Advances in Signal Processing*, to be published on November 2007.
6. Chen, M., Hauptmann, A.: Towards Robust Face Recognition from Multiple Views. *Proc. IEEE Int. Conf. on Multimedia and Expo*, 2004.
7. IP CHIL-Computers in the Human Interaction Loop: <http://chil.server.de>
8. Chiu, P., Kapuskar, A., Reitmeier, S., Wilcox, L.: Room with a rear view: Meeting capture in a multimedia conference room. *IEEE Multimedia Magazine*, vol. 7:4, pp. 48–54, 2000.
9. CLEAR Evaluation Workshop, 2006.
10. Deutscher, J., Reid, I.: Articulated Body Motion Capture by Stochastic Search. *Int. Journal of Computer Vision*, vol.61:2, pp. 185–205, 2005.
11. Gatica-Perez, D., Lathoud, G., Odobez, J.-M., McCowan, I.: Audiovisual Probabilistic Tracking of Multiple Speakers in Meetings. *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15:2, pp. 601–616, 2007.
12. Gordon, N.J., Salmond, D.J., Smith, A.F.M.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. on Radar and Signal Processing*, vol.140:2, pp. 107–113, 1993.
13. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*. 2nd Edition. Cambridge University Press. 2004.
14. Horprasert, T., Yacoob, Y., Davis, L.S.: Computing 3-D head orientation from a monocular image sequence. *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pp. 242–247, 1996.
15. Isard, M., Blake, A.: CONDENSATION—Conditional Density Propagation for Visual Tracking. *Int. Journal of Computer Vision*, vol.29:1, pp. 5–28, 1998.
16. Jones, M., Rehg, J.: Statistical Color Models with Application to Skin Detection. *Int. Journal of Computer Vision*, vol. 46:1, pp. 81–96, 2002.
17. Matsumoto, Y., Zelinsky, A.: An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 499–504, 2000.
18. 72 Nickel, K., Gehrig, T., Stiefelhagen, R., McDonough, J.: A joint particle filter for audio-visual speaker tracking. *Proc. IEEE Int. Conf. on Multimodal Interfaces*, pp. 61–68, 2005.
19. Rae, R., Ritter, H. J.: Recognition of Human Head Orientation Based on Artificial Neural Networks. *IEEE Tran. on Neural Networks*, vol. 9, pp. 257–265, 1998.
20. Voit, M., Nickel, K., Stiefelhagen, R.: Neural Network-based Head Pose Estimation and Multi-view Fusion. *Proc. CLEAR Workshop, LNCS*, vol. 4122, pp. 299–304, 2006.
21. Wang, C., Brandstein, M.: Robust head pose estimation by machine learning. *Proc. IEEE Int. Conf. on Image Processing*, vol. 3, pp. 210–213, 2000.

22. Wang, C., Griebel, S., Brandstein, M.: Robust automatic video-conferencing with multiple cameras and microphones. Proc. IEEE Int. Conf. on Multimedia and Expo, vol. 3, pp. 1585–1588, 2000.
23. West, M., Harrison, J.: Bayesian forecasting and dynamic models. Springer-Verlang, New York, 2nd Ed., 1997.
24. Zhang, Z., Hu, Y., Liu, M., Huang, T.: Head Pose Estimation in Seminar Rooms Using Multi View Face Detectors. Proc. CLEAR Workshop, LNCS, vol. 4122, pp. 299–304, 2006.