Probabilistic Head Pose Tracking Evaluation in Single and Multiple Camera Setups

Sileye O. Ba [a] Jean-Marc Odobez [a]

IDIAP–RR 07-21

march 2007

to appear in
Classification of Events, Activities and relationships (CLEAR)
Evaluation and Workshop

[a] IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne

IDIAP Research Institute www.idiap.ch
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

Probabilistic Head Pose Tracking Evaluation in Single and Multiple Camera Setups

Sileye O. Ba Jean-Marc Odobez

march 2007

to appear in
Classification of Events, Activities and relationships (CLEAR) Evaluation and Workshop

Abstract. This paper presents our participation in the CLEAR 07 evaluation workshop head pose estimation tasks where two head pose estimation tasks were to be addressed. The first task estimates head poses with respect to (w.r.t.) a single camera capturing people seated in a meeting room scenario. The second task consisted of estimating the head pose of people moving in a room from four cameras w.r.t. a global room coordinate. To solve the first task, we used a probabilistic exemplar-based head pose tracking method using a mixed state particle filter based on a represention in a joint state space of head localization and pose variable. This state space representation allows the combined search for both the optimal head location and pose. To solve the second task, we first applied the same head tracking framework to estimate the head pose w.r.t each of the four camera. Then, using the camera calibration parameters, the head poses w.r.t. individual cameras were transformed into head poses w.r.t to the global room coordinates, and the measures obtained from the four cameras were fused using reliability measures based on skin detection. Good head pose tracking performances were obtained for both tasks.

1 Introduction

The study of head related-behaviors such as head gestures is of interest in many computer vision related applications. When studying head gestures in a general framework, information about head poses are required. Over at least a decade, many head pose tracking methods have been proposed. However, before being the bases of head gesture studies, the performances of the proposed head pose estimation methods have to be thoroughly investigated. For evaluating head pose estimation methods evaluation databases are required. Efforts have been made to build and make publicly available, a head pose video database with people having their head orientation continuously annotated with a magnetic field location and orientation tracker [4]. Such a database is usefull for comparison of cross-institution head pose evaluation where similar protocols can be used. Therefore, since 2006, the Classification of Events, Activities and Relationships (CLEAR) evaluation workshop has targetted the evaluation of head-pose estimation algorithms. In 2006, the head pose tracking task in the CLEAR evaluation workshop involved estimating the head direction of a person among 8 possible directions (North, North-East, East,..., where the cardinal directions corresponded to the wall of a room). A limitation of this task was that the head directions were annotated by hand and that no precise evaluation was possible. This was remedied in 2007.

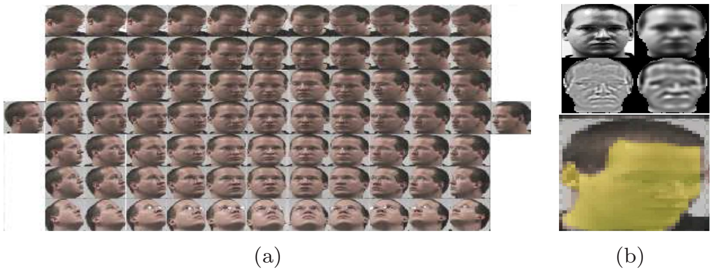
In the 2007 CLEAR evaluation workshop, two head pose estimation tasks were proposed. The first task, called the Augmented Multi-party Interaction (AMI) task, was about estimating people's head pose w.r.t. the view from a single camera. The data for this task consisted of 8 one-minute recordings. In each recording, four people are involved in a meeting and two people among the four have their head pose w.r.t. to a camera view annotated using a magnetic field location and orientation tracker [1]. These annotations were used for evaluation as a head pose ground truth (GT). The second task, called the Computers in the Human Interaction Loop (CHIL) task, involved estimating the head pose of a person w.r.t. to a global room coordinate system using four camera views of the person. For this task head pose ground truth was also generated using a magnetic field location and orientation tracker. The two tasks are interesting in the sense that they cover two common scenarii in computer vision applications. The first scenario occurs in a meeting room in which people are mostly seated. The second, occurs in a seminar room or a lecture theatre in which the head is captured at a much lower resolution and the people are mostly standing and moving. Evaluating head pose tracking algorithms in these two situations is important to understand the behaviors of the algorithms for a wide range of potentially interesting experimental setups.

In this work we used a probabilistic method based on a mixed state particle filter to perform head pose tracking w.r.t. a single camera view [5]. Applying this method solves the AMI head pose estimation task. To address the CHIL task, the head pose w.r.t. the camera is transformed to be relative to the global room coordinate system using the camera calibration parameters. Then the head pose estimated w.r.t. to the global room coordinate obtained from the four cameras are fused into a single head pose estimate using the percentage of skin present in the estimated bounding box for the head as reliability measure.

The remainder of this paper describes in more details the methods we used to solve the two tasks. Section 2 describes the estimation method in term of the head pose w.r.t. a single camera view. Section 3 describes the method we used to estimate the head pose w.r.t to a global room coordinate to solve the second task. Section 4 gives the results we obtained for the AMI task and Section 5 the results for the CHIL task. Finally, Section 6 provides some concluding remarks.

2 Head Pose Tracking with Respect to a Camera View

In this Section, we summarize the probabilistic method we used to track the head pose of a person w.r.t. a single camera view. This method is more thoroughly described in [5, 3].



(a) (b)

Figure 1: a) Training head pose appearance range (Prima-Pointing Database[6]) and b) texture features from Gaussian and Gabor filters and skin color binary mask.

2.1 Probabilistic Method for Head Pose Tracking

The Bayesian formulation of the tracking problem is well known. Denoting the hidden state representing the object configuration at time t by X_t and the observation extracted from the image by Y_t, the objective is to estimate the filtering distribution $p(X_t|Y_{1:t})$ of the state X_t given the sequence of all the observations $Y_{1:t} = (Y_1, \ldots, Y_t)$ up to the current time. Given standard assumptions, Bayesian tracking effectively solves the following recursive equation:

$$p(X_t|Y_{1:t}) \propto p(Y_t|X_t) \int_{X_{t-1}} p(X_t|X_{t-1}) p(X_{t-1}|Y_{1:t-1}) \mathrm{d}X_{t-1} \qquad (1)$$

In non-Gaussian and non linear cases, this can be done recursively using sampling approaches, also known as particle filters (PF). The idea behind PF consists of representing the filtering distribution using a set of N_s weighted samples (particles) $\{X_t^n, w_t^n, n = 1, ..., N_s\}$ and updating this representation when new data arrives. Given the particle set of the previous time step, configurations of the current step are drawn from a proposal distribution $X_t \sim q(X|X_{t-1}^n, Y_t)$. The weights are then computed as $w_t \propto w_{t-1}^n \frac{p(Y_t|X_t)p(X_t|X_{t-1}^n)}{q(X_t|X_{t-1}^n, Y_t)}$.

Five elements are important in defining a PF: i) a state model which is an abstract representation of the object we are interested in; ii) a dynamical model $p(X_t|X_{t-1})$ governing the temporal evolution of the state; iii) a likelihood model $p(Y_t|X_t)$ measuring the adequacy of the data given the proposed configuration of the tracked object; and iv) a proposal distribution $q(X|X_{t-1}^n, Y_t)$ the role of which is to propose new configurations in high likelihood regions of the state space v) and a sampling mechanism which defines how the filtering distribution will be approximated using particles. These elements are described in the following paragraphs.

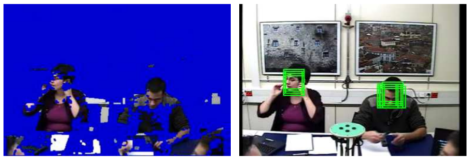
State Space: The state space contains both continuous and discrete variables. More precisely, the state is defined as $X = (S, r, l)$ where S represents the head location and size, and r represents the in-plane head rotation. The variable l labels an element of the discretized set of possible out-of-plane head poses. In addition to the set of poses displayed in Fig. 1(a), 3 additional poses at pan values of -135$\circ$, -180$\circ$, and 135$\circ$ (and a 0$\circ$ tilt) were selected to represent the head from the back and allow head tracking when people are turning their head to the camera.

Dynamical Model: The dynamics governs the temporal evolution of the state, and is defined as

$$p(X_t|X_{t-1}) = p(r_t|r_{t-1}, l_t) p(l_t|l_{t-1}, S_t) p(S_t|S_{t-1}, S_{t-2}) \,. \qquad (2)$$

The dynamics of the in-plane head rotation r_t and discrete head pose l_t variables are learned using head pose GT training data. Notice that the roll dynamics depend on the out-of-plane appearance (in plane rotation dynamics is different for frontal and profile poses). Head location and size dynamics are modelled as second order auto-regressive processes.

Observation Model: The observation model $p(Y|X)$ measures the likelihood of the observation for a given state . The observations $Y = (Y^{tex}, Y^{skin}, Y^{sil})$ are composed of texture features, skin color



(a) (b)

Figure 2: a) Silhouette binary features (foreground mask), b) heads detected from the silhouette features.

features (see Fig. 1(b)) and silhouette features (see Fig. 2(a)). Texture features are represented by the output of three filters (a Gaussian and two Gabor filters at different scales) applied at locations sampled from image patches extracted from each frame and preprocessed by histogram equalization to reduce light variations effects. Skin color features are represented by a binary skin mask extracted using a temporally adapted skin color model. The silhouette features are represented by a binary mask which is obtained from foreground segmentation using a temporally adapted background model as presented in [7]. Assuming that, given the state vector, the texture, skin color, and silhouette features are independent, the observation likelihood is modeled as:

$$p(Y|X = (S, r, l)) = p_{tex}(Y^{tex}(S, r)|l)p_{skin}(Y^{skin}(S, r)|l)p_{sil}(Y^{sil}) \tag{3}$$

where $p_{tex}(\cdot|l)$ and $p_{skin}(\cdot|l)$ are pose dependent models learned from the Prima-Pointing Database[6] and $p_{sil}(\cdot)$ is the silhouette likelihood model learned from training data (provided with the two tasks). For a given hypothesized configuration X, the parameters (S, r) define an image patch on which the features are computed, while for the pose dependent models, the exemplar index l selects the appropriate appearance likelihood model.

Proposal Distribution: The role of the proposal distribution is to suggest candidate states in interesting regions of the state space. As a proposal distribution, we used a mixture between the state dynamics $p(X_t|X_{t-1})$ and a head detector $p(X_t|\hat{\mathcal{X}}_t^d((Y_t)))$ based on the silhouette features according to the formula:

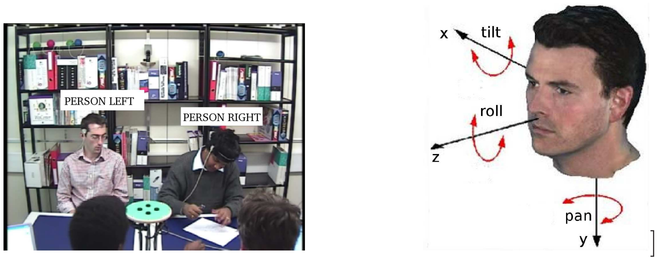
$$q(X_t|X_{t-1}^n, Y_t) = (1-\alpha)p(X_t|X_{t-1}^n) + \alpha p(X_t|\hat{\mathcal{X}}_t^d((Y_t))) \tag{4}$$

where $\alpha < 1$ is a mixture weight and $\hat{\mathcal{X}}_t^d((Y_t)) = \{\hat{X}_i^d(Y_t),\ i = 1, ..., N_t^d\}$ is the set candidate head states obtained from the detection procedure illustrated by the green boxes in Fig. 2(b). Qualitatively, the particles drawn from the second mixture components are randomly sampled around the detected head locations. More information about the proposal function can be found in [3]. The state dynamics are used to enforce temporal continuity in the estimated filtering distribution while the head detector's role is to allow automatic re-initialization after short-term failure and to avoid the tracker being trapped in local maxima of the filtering distribution.

Sampling Method: In this work, we use Rao-Blackwellization (RB) which is, a process in which we apply the standard PF algorithm to the tracking variables S and r while applying an exact filtering step to the exemplar variable l. In the current case, this means that the samples are given by: $X^i = (S^i, r^i, \pi^i(l), w^i)$ instead of $X^i = (S^i, r^i, l^i, w^i)$, where $\pi^i(l)$ represents the posterior distribution of the pose variable given all the other variables and the observations. The method theoretically results in a reduced estimation variance, as well as a reduction of the number of samples. For more details about the RB procedure, the reader is referred to [5].

3 Head Pose w.r.t to a Global Room Reference

In section 2 we presented a method to track and estimate head poses w.r.t. to a camera viewing direction. Using the head pose w.r.t. to a camera view, we can estimate it w.r.t. to a global room reference using matrix transformations.



(a) Sample view of the AMI evaluation data.

(b) Head attached basis w.r.t which head rotations are defined.

Figure 3: Example views of two persons in the evaluation data.

3.1 Head Pose Estimation w.r.t to a Global Room Reference using a Single Camera View

Head poses can be equivalently represented using Euler angles or a rotation matrixes. Denoting by R_{vdir}^{head} the current pose expressed in the local viewing direction coordinate reference system. Then, given the correction matrix R_{cam}^{vdir} that expresses the local reference vectors into the camera reference vector basis (this matrix depends on the image position of the head), and the calibration matrix R_{3D}^{cam} that provides the rotation of the camera reference w.r.t. the global 3D reference, the current pose w.r.t the global reference can simply be computed as: $R_{3D}^{head} = R_{3D}^{cam} R_{cam}^{vdir} R_{vdir}^{head}$. The pose w.r.t. to the global room reference is then obtained using the Euler angle representation of the matrix R_{3D}^{head}.

3.2 Head Pose Estimation w.r.t to a Global Room Coordinate using Multiple Camera Views

The method described in Section 3.1 allows us to estimate the head pose of a person w.r.t. the global room reference using only a single camera. When multiple camera views are available, head poses can be estimated from these sources by defining a procedure to fuse the estimates obtained from the single camera views. Such a fusion procedure is usually based on a measure that assesses how reliable the estimates are for each camera. Assuming the head is correctly tracked, the amount of skin pixels in the head bounding box can be used as a reliability measure. In general, a high percentage of skin pixels is characteristic of near frontal head poses for which head pose estimation methods are known to be more reliable [2], while a low percentage of skin pixels in the head location means the head appears either as a near profile head pose or from the back. Thus, we defined a camera fusion procedure as follows. After tracking the person in each of the camera views and having estimated the head pose w.r.t. to the global room reference, the final head pose is estimated by averaging the estimates from the two cameras for which the percentage of skin in the head bounding box is higher.

4 The AMI Head Pose Estimation Task

The AMI head pose estimation task consisted of tracking head poses w.r.t a single camera in a meeting room. In this section we describe the evaluation data and protocols for this task, then give the results achieved by our head pose tracking algorithm.

4.1 Evaluation Data and Protocols

The AMI data is composed of 8 meetings recorded from a single camera view. Four people are involved in each meeting. Among the four people, two which are always visible are used for head pose estimation evaluation. Figure 3(a) shows an example view of the two people denominated *person right* and *person left*. The evaluation data consists of 1 minute recordings of 16 people. The head pose annotations of the 16 people used for evaluation were obtained using a magnetic field 3D location

error (in degrees)	1R	1L	2R	2L	3R	3L	average
pointing vector	15. 6	17. 5	16. 0	14. 8	8. 4	11. 6	14. 0
pan	9. 9	13. 4	4. 9	12. 9	4. 4	7. 4	8. 8
tilt	11. 2	9. 5	14. 7	6. 7	6. 8	7. 5	9. 4
roll	10. 4	8. 1	13. 7	8. 2	7. 2	11. 5	9. 8

Table 1: Head pose estimation performance for the person left (L) and right (R) in the three test meetings of the AMI data. The last column gives the average pose estimation errors.

and orientation tracker [1], called a flock of bird, that was attached to each person's head. After calibration of the flock of birds to the camera, the outputs of the sensors were transformed to generate the head pose annotations. Among the 16 people available, the data (video recordings and head pose annotations) of 10 people were used as development data, and that of the 6 remaining people were used as test data.

As performance measures, we used four metrics: the head pointing vector error, and the absolute head pan, tilt, and roll estimation error. The head pointing vector is the normal unit vector of the z axis of the basis attached to the head (see Fig 3(b)). It defines the head pointing direction. The head pointing vector error is the absolute angle between the ground truth head pointing vector and the estimate. The head pan is defined as the rotation w.r.t to the y- axis of the basis attached to the head (see Fig 3(b)), the head tilt is the rotation w.r.t. to the x-axis and the head roll is the rotation w.r.t. to the z-axis. The estimation errors for these angles are the absolute differences between the head pose ground truth and estimates.

4.2 Results for the AMI Task

To solve the AMI task, we initialized the head localization manually before applying the head pose tracking method described in Section 2. The performances of the algorithm for the 6 persons in the test set are given in Table 1. Over the whole AMI evaluation dataset, our head pose tracking method achieves an average estimation error of 14° for the pointing vector estimation, 8.8° for head pan estimation, 9.4° for the head tilt estimation and for 9.8° for the head roll estimation. An analysis of the errors according to each individual shows significant variability of the performances due to variations in appearance or sitting attitude. For the head pointing vector estimation, the lowest estimation errors is achieved with the person sitting to the right side in the third test meeting (3R in Table 1) while the highest errors are obtained with person 1L (cf Table 1). This shows that some people are much better tracked than others, which is most probably due to the fact that they are better represented by the appearance models than others.

Fig. 4 gives sample images of head pose tracking results. This figure can be analyzed in parallel with Table 1. The first row of Fig. 4 shows that for person 1L (left person), head localization problems occur in some frames. In the last row of Fig. 4, we can observe that the person sitting to the right side, for whom the best tracking performance are achieved, has his head always correctly localized, even in difficult conditions. This illustrates the correlation between good head pose estimation performance and good head localization performance.

5 The CHIL Head Pose Estimation Task

The CHIL task consisted of estimating the head pose of a person w.r.t to a global room coordinate system using single or multiple camera views. In the following subsections, we describe the evaluation data and protocols and show the results using the algorithm described in Section 3.

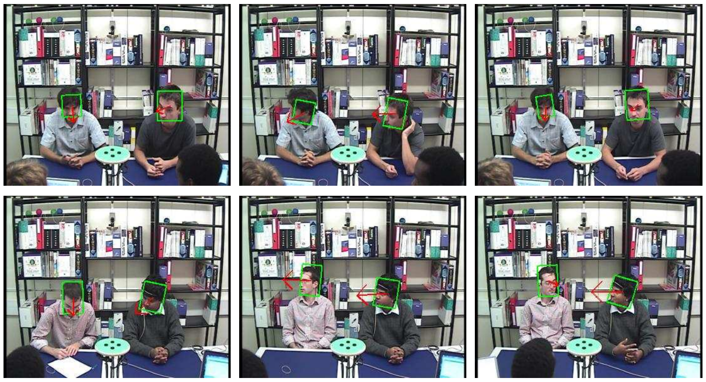
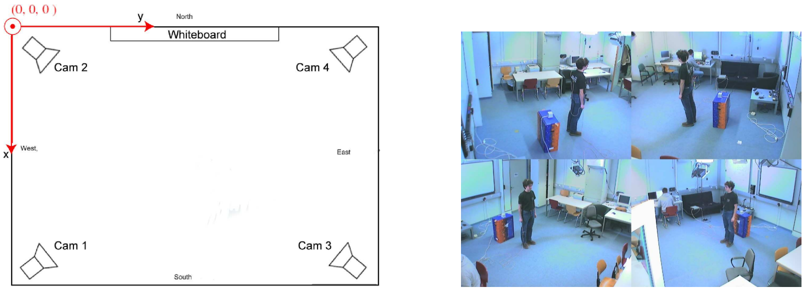


Figure 4: Sample images of head pose tracking results in the AMI data: first row show images of the first test meeting, second row shows images for the third test meeting.



(a) Top view of the room. At the top left corner is the global room reference and at the four corners are the cameras.

(b) Sample images from the 4 camera recordings. First row, camera 4 and 2, second row camera 1 and 3

Figure 5: Top view of the seminar room and sample images from the four camera views

5.1 Evaluation Data and Protocols

The CHIL data involved 15 people recorded in a seminar room. The 15 people were wearing a magnetic field location and orientation tracker, which provided their head pose w.r.t a global room reference. Four cameras located in the upper corners of the room were used to record the whole scene during three minutes in which the people had to move around and orient their head towards all the possible directions. Fig. 5(a) shows a top view of the room with its reference and the four cameras and Fig. 5(b) shows a sample image of all four camera views. In each of the recordings, head location annotations were provided every 5 frames. The recorded data were split into 10 videos with their corresponding annotations for training and 5 videos for testing. Only the frames for which head location annotations were available were used for head pose estimation evaluation. Similar to the AMI task, the error measures used for evaluation were the head pointing vector, pan, tilt and roll errors in degrees.

5.2 Results for the CHIL Task

To solve the CHIL task we used two methods. The first method, denoted CHIL-M1, is based on head pose tracking with respect to a single camera view as described in Section 2. Then the pose w.r.t. the camera are transformed into pose w.r.t. to the global room reference using the methodology described in Section 3.1. For this method, only one camera (cam 3 in Fig. 5(b)) was considered. The second

method	pointing vector	pan	tilt	roll
CHIL-M1	30.0°	24.1°	14.0°	7.3°
CHIL-M2	19.4°	15.0°	10.0°	5.3°

Table 2: Head pose estimation average errors over the whole test set of the CHIL data using single camera (CHIL-M1) and the fusion of the four cameras (CHIL-M2).

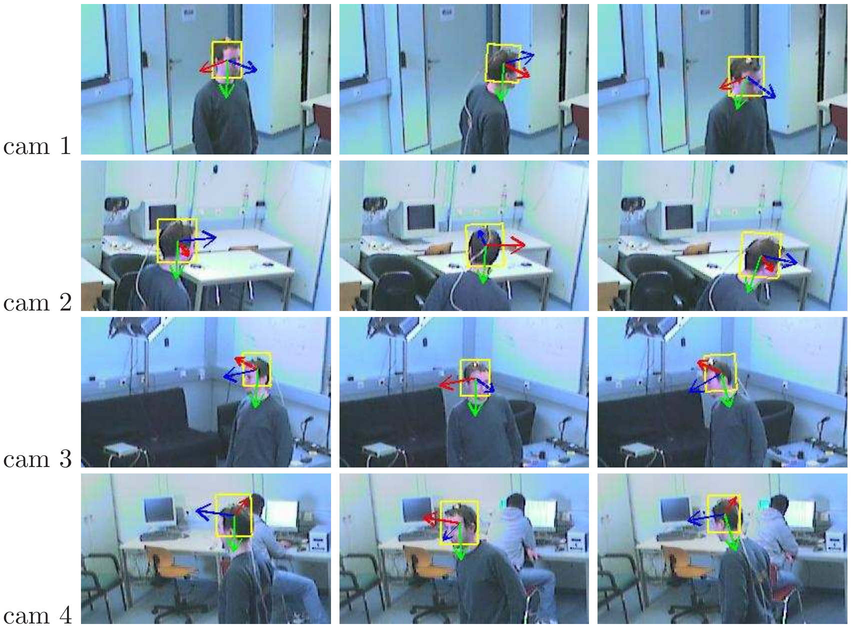


Figure 6: Sample images (cropped from the original images for better visualization) of head pose tracking results for CHIL-M2. Each row displays sample images for the corresponding camera. Images of the same column correspond to a single time frame recorded from the 4 camera views.

method, denoted CHIL-M2, used the head poses w.r.t. to the global room reference estimated by four cameras and fused the estimates into a single one using the fusion procedure described in Section 3.2. In the following experiments the initial head locations were again provided manually.

Table 2 gives the average head pose estimation errors for the whole CHIL test data using the two methods. From the results, we can conclude that the method based on multiple camera fusion outperforms the method that used a single camera view. The improvements can be explained by the camera selection being implicitly embedded into the fusion process. More precisely, when using only one camera, large errors are produced when the tracked persons are showing the back of their head. On the contrary, in the fusion scheme, only the two cameras with the highest reliability measure -usually the ones that the person is facing- are selected to estimate the head pose, thus providing good results in almost all conditions.

Fig. 6 shows the head pose tracking results for one person and illustrates the usefulness of the fusion procedure. In the second column for instance, camera 3 and 4 were automatically selected to provide the pose results.

6 Conclusion

In this paper we described our participation to the two head pose estimation tasks of the CLEAR07 Evaluation and Workshop. We proposed to use an exemplar-based representation of head appearances embedded into a mixed state particle filter framework. This method allowed us to estimate the head orientation of a person w.r.t. to a single camera view, thus solving the first task. The second task was solved by transforming the rotation matrix defining the pose w.r.t. the camera using the camera calibration parameters of a camera to obtain the head pose w.r.t a global room reference. This procedure was improved by fusing the single camera estimates using skin color as a camera fusion reliability measure. Good performances were achieved by the methods we proposed in solving both tasks. In term of future work, we plan to define the head localization component of the state space of our mixed state particle filter directly in the three-dimensional space rather than in the image plane.

Acknowledgment:
This work was partly supported by the Swiss National Center of Competence in Research and Interactive Multimodal Information Management (IM2), and the European union 6th FWP IST Integrated Project AMI (Augmented Multi-Party Interaction, FP6-506811). This research was also funded by the U.S. Government VACE program.
The authors also thank Hayley Hung and Dr. Daniel Gatica-Perez from IDIAP Research Institute for their fruitfull discussions and usefull comments.

References

[1] Ascencion Technology. Flock of Birds.

[2] S. Ba and J. Odobez. Evaluation of Multiple Cues Head-Pose Tracking Algorithms in Indoor Environments. In *International Conference on Multimedia and Expo (ICME)*, Amsterdam, July 2005.

[3] Sileye O. Ba. *Joint Head Tracking and Pose Estimation for Visual Focus of Attention Recognition*. PhD thesis, Ecole Polytechnique Federale de Lausanne, 2007.

[4] Sileye O. Ba and Jean Marc Odobez. A Video Database for Head Pose Tracking Evaluation. Technical Report 04, IDIAP Research Institute, 2005.

[5] S.O. Ba and J.-M. Odobez. A Rao-Blackwellized Mixed State Particle Filter for Head Pose Tracking. In *ICMI Workshop on Multi-Modal Multi-Party Meeting Processing (MMMP)*, pages 9–16, 2005.

[6] N. Gourier, D. Hall, and J. L. Crowley. Estimating Face Orientation from Robust Detection of Salient Facial Features. In *Pointing 2004, ICPR international Workshop on Visual Observation of Deictic Gestures*, pages 183–191, 2004.

[7] Jian Yao and Jean-Marc Odobez. Multi-Layer Background Subtraction Based on Color and Texture. In *CVPR 2007 Workshop on Visual Surveillance (VS2007)*, 2007.