Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 276846, 12 pages
doi:10.1155/2008/276846

Research Article

Audiovisual Head Orientation Estimation with Particle Filtering in Multisensor Scenarios

Cristian Canton-Ferrer,[1] Carlos Segura,[2] Josep R. Casas,[1] Montse Pardàs,[1] and Javier Hernando[2]

[1] *Image Processing Group, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain*
[2] *TALP Research center, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain*

Correspondence should be addressed to Cristian Canton-Ferrer, ccanton@gps.tsc.upc.edu

Received 1 February 2007; Accepted 7 June 2007

Recommended by Enis Ahmet Cetin

This article presents a multimodal approach to head pose estimation of individuals in environments equipped with multiple cameras and microphones, such as SmartRooms or automatic video conferencing. Determining the individuals head orientation is the basis for many forms of more sophisticated interactions between humans and technical devices and can also be used for automatic sensor selection (camera, microphone) in communications or video surveillance systems. The use of particle filters as a unified framework for the estimation of the head orientation for both monomodal and multimodal cases is proposed. In video, we estimate head orientation from color information by exploiting spatial redundancy among cameras. Audio information is processed to estimate the direction of the voice produced by a speaker making use of the directivity characteristics of the head radiation pattern. Furthermore, two different particle filter multimodal information fusion schemes for combining the audio and video streams are analyzed in terms of accuracy and robustness. In the first one, fusion is performed at a decision level by combining each monomodal head pose estimation, while the second one uses a joint estimation system combining information at data level. Experimental results conducted over the CLEAR 2006 evaluation database are reported and the comparison of the proposed multimodal head pose estimation algorithms with the reference monomodal approaches proves the effectiveness of the proposed approach.

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" [1], "intelligent house," "intelligent classroom," and "smart conferencing rooms" [2, 3]. 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 [4]. 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.

Previous approaches to estimate the head pose have mostly used video technologies. The first techniques proposed for head orientation estimation rely on facial feature detection. The facial features extracted are compared to a face model to determine the head orientation [5, 6]. These approaches usually require high-resolution images which are not commonly available in the aforementioned scenarios. Global techniques that use the entire image of the face to estimate the head orientation are more suitable in these scenarios. Most of the global techniques produce a classification of the head orientation based on a number of previously learned classes using neural networks [7–10]. An analysis-by-synthesis approach is proposed in [11]. The estimation of head orientation based on audio is a very new and challenging task. An early work on speaker orientation based on acoustic energy was defined in [12], which was using a large microphone array consisting in hundreds of sensors surrounding the environment. The oriented global coherence field (OGCF) method has been proposed in a recent work [13], which is a variation on GCF acoustic localization algorithm.

In scenarios where both audio and video are available, such as Smart Rooms or automatic video conferencing, a multimodal approach can achieve more accurate and robust results. Audio information is only available for the person who is speaking, but this person is usually the center of attention for the system. For this reason, audio information will improve the precision of the head orientation system for the speaking person and will correct errors produced in the video analysis due to the estimation system or to the unavailability of video data (when the person moves away from the camera field of view).

Recently [14], the authors have presented two multimodal algorithms aiming to estimate the head pose using audiovisual information. The proposed architecture combines the results of a former system from the authors based on video [15] and a novel method using exclusively acoustic signals from a small set of microphones. In the monomodal video system, the estimation is performed by fitting a 3D reconstruction of the head combining the views from a calibrated set of cameras. Audio head orientation is based on the fact that the radiation pattern of the human head is frequency dependent. Within this context, we propose a method for estimating the orientation of an active speaker using the ratio of energy in different bands of frequency. The fusion was made both at data level and also at decision level by means of a decentralized Kalman filtering applied to the sequence of the video and audio orientation estimates [16].

Particle filters have proved to be a very useful technique for tracking and estimation tasks when the variables involved do not hold Gaussianity uncertainty models and linear dynamics [17]. They have been successfully used for video object tracking and for audio source localization. Information of audio and video sources has also been effectively combined employing PF strategies for active speaker tracking [18] or audiovisual multiperson tracking [19].

In this article, we propose to use particle filters as a unified framework for the estimation of the head orientation for both monomodal and multimodal case. Regarding particle filter multimodal fusion, two different strategies for combining the audio and video data are proposed. In the first one, information is performed at a decision level combining each monomodal head pose estimation, while the second one uses a joint estimation system combining information at data level.

The remainder of this paper is organized as follows. In Section 2, we present the general architecture of the system that we propose, and we introduce the particle filters that will be the basis of the estimation techniques that we develop in the following sections. In Section 3, the monomodal video head estimation technique is introduced, and in Section 4, we present the audio single modality system for speaker orientation estimation. In Section 5, we propose two methods to fuse audio and video modalities combining the estimations provided by each system at the data and decision levels. In Section 6, the performance obtained by each system is discussed, and we conclude the paper in Section 7.

2. ANALYSIS FRAMEWORK

Nowadays the decreasing cost of audio and visual sensors and acquisition hardware makes the deployment of multisensor systems for distributed audio visual observation commonplace. Intelligent scenarios requires the design of flexible and reconfigurable perception networks feeding data to the perceptual analysis front end [20]. The design of multicamera configurations for continuous room video monitoring consists of several calibrated cameras, connected to dedicated computers, whose fields of view aim to cover completely the scene of interest, usually with a certain amount of overlap allowing for triangulation and 3D data capture for visual tracking, face localization, object detection, person identification, gesture classification, and overall scene analysis. A multimicrophone system for aural room analysis deploys a flexible microphone network comprising microphone arrays, microphone clusters, table top microphones, and close-talking microphones, targeting the detection of multiple acoustic events, voice activity detection, ASR and speaker location and tracking. Also for acoustic sensors, a calibration step is defined, according to the purpose of having a jointly consistent description of the audio-video sensor geometry, and timestamps are added to all the acquired data for temporal synchronization.

The perceptual analysis front end of an intelligent environment consists of a collection of perceptual components detecting and classifying low-level features which can be later interpreted at a higher semantical level. The perceptual component analyzing the audio-visual data for head orientation detection contributes a low-level feature yielding fundamental clues to drive the interaction strategy.

The angle of interest to be estimated for our purposes in a multisensor scenario has been chosen as the orientation of the head onto the xy plane. This angle provides semantical information such as where people is looking at in the scene and it can be used for further analysis such as tracking of attention in meetings [21]. In the next subsection, particle

filters will be introduced as the technological base for all the systems described in this article.

2.1. Particle filtering

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 [22] driven by the following state process equation:

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

and the observation equation:

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

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* $p(\theta_t \mid \Omega_{1:t})$, and this can be done recursively in two steps, namely, prediction and update. The prediction step uses the process equation (1) to obtain the prior *pdf* by means of the Chapman-Kolmogorov integral

$$p(\theta_t \mid \Omega_{1:t-1}) = \int p(\theta_t \mid \theta_{t-1}) p(\theta_{t-1} \mid \Omega_{1:t-1}) d\theta_{t-1} \tag{3}$$

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

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

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

Particle filtering (PF) [23] 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 $p(\theta_t \mid \Omega_{1:t})$ with a sum of N_s Dirac functions centered in $\{\theta_t^j\}$, $0 < j \leq N_s$ as

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

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 resampling (SIR) strategy to drive particles across time [24]. This assumption leads to a recursive update of the weights as

$$w_t^j \propto w_{t-1}^j \, p(\Omega_t \mid \theta_t^j). \tag{6}$$

SIR PF circumvents the particle degeneracy problem by resampling with replacement at every time step [23], that is, to dismiss the particles with lower weights and proportionally replicate those with higher weights. In this case, weights are set to $w_{t-1}^j = N_s^{-1}$ for all j, therefore,

$$w_t^j \propto p(\Omega_t \mid \theta_t^j). \tag{7}$$

Hence, the weights are proportional to the likelihood function that will be computed over the incoming data Ω_t. The resampling 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 (5). The most common solution is the Monte Carlo approximation of the expectation

$$\Theta_t = \mathbb{E}[\theta_t \mid \Omega_{1:t}] \approx \sum_{j=1}^{N_s} w_t^j \theta_t^j. \tag{8}$$

Finally, a propagation model is adopted to add a drift to the angles θ_t^j of the resampled particles in order to progressively sample the state space in the following iterations [23]. For complex PF problems involving a high-dimensional state space such as in articulated human body tracking tasks [25], 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 [17] or speech enhancement [26]. Information of audio and video sources have been effectively combined employing PF strategies for active speaker tracking [18] or audiovisual multiperson tracking [19].

2.2. PF applied to multimodal head pose estimation

PF techniques will be applied to the problem under study taking into account a common criteria when designing the implementation of the PF for both audio and video modalities. This common design criterion will allow natural multimodal information fusion strategies at decision and data level as it will be described in Section 5.

An input observation Ω_t may be written as the set

$$\Omega_t = \left[\Omega_t^{\mathrm{A}} \;\; \Omega_t^{\mathrm{V}}\right], \tag{9}$$

where Ω_t^{A} and Ω_t^{V} refer to the audio and video observations, respectively. For both sources, it may happen that these sets are empty depending whether there is audio or video information available or not. Typically, $\Omega_t^{\mathrm{A}} = \varnothing$ when the subject

under study is not speaking and $\Omega_t^{\mathrm{Y}} = \varnothing$ when there is not a projection of the head of the person in any camera. From this data perspective, three analysis possibilities can be devised: audio, video, and audiovisual processing.

The main factor to be taken into account when employing PF is the construction of the likelihood evaluation function that will measure the similarity between the input data set Ω_t and a given pan angle θ_t^j. This function will assign the weights to the particles as stated by (7).

Finally, it must be noted that if more than one person is present in the scene, a PF estimating the head orientation will be assigned for each of them.

3. VIDEO HEAD POSE ESTIMATION

Methods for head pose estimation from video signals proposed in the literature can be classified as feature based or appearance based [27]. Feature based methods [5, 6, 28] 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 [15, 28, 29]. On the contrary, appearance-based methods [8, 30] 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 [7] angle estimation is restricted to steps of 25° while in [31] steps of 45° are employed. When performing a multimodal fusion, informative video outputs are desired, thus preferring data analysis methods providing a real-valued angle output.

This section presents a new approach to multicamera 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, as described in Section 2, thus producing a real-valued output which will increase the precision of our system as compared with classification approaches and will pave the way for the multimodal integration.

3.1. Spatial analysis

Head localization is the first task to be performed before any head orientation estimation process. This objective has been addressed in the literature referred as person localization and tracking [32, 33] or face localization [34]. Here, a head localization algorithm based on our previous research [35] is reviewed.

Prior to any further image analysis, the analyzed scene must be characterized in terms of space disposition and configuration of the foreground volumes, that is, people candidates, in order to select those potential 3D regions where the head of a person could be present. Images obtained from a multiple view camera system allow exploiting spatial redundancies in order to detect these 3D regions of interest [36]. For a given frame in the video sequence, a set of N_{CAM} images are obtained from the N_{CAM} cameras. Each camera is modeled using a pinhole camera model based on perspective projection. Accurate calibration information is available. Foreground regions from input images are obtained using a segmentation algorithm based on Stauffer-Grimson's background learning and substraction technique [37]. It is assumed that the moving objects are human people. Original and segmented images are the input information for the rest of image analysis modules described here.

Once foreground regions are extracted from the set of N_{CAM} original images at time t, a set of M 3D points $\mathbf{x}^k$, $0 \leq k < M$, corresponding to the top of each 3D detected volume in the room is obtained by applying the robust Bayesian correspondence algorithm described in [35]. Information coming from the tracking loop speeds up the process narrowing the search space of these correspondences on time $t + 1$ and allows rejecting false head detections.

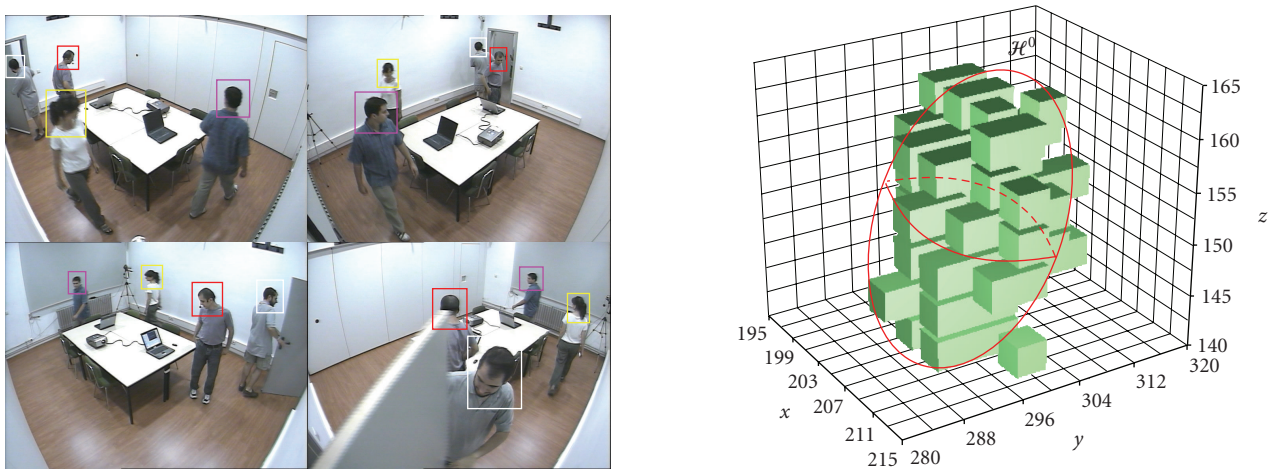
The information given by the established correspondences allows defining a bounding box $\mathcal{B}^k$, centered on each 3D top $\mathbf{x}^k$ with an average size adequate to contain the human head candidate (see an example of this output in Figure 1(a)). Afterwards, a voxel reconstruction [38] is computed on each bounding box $\mathcal{B}^k$, thus obtaining a set of voxels $\mathcal{V}^k$ defining the kth 3D foreground volume candidate as a head. In order to refine and verify whether the set $\mathcal{V}^k$ indeed belongs to an ellipsoidal geometric shape, a template matching evaluation [38] is performed.

3.2. 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 [39], where the color distribution of skin is estimated from offline hand-selected samples of skin pixels.

Finally, color information is combined with spatial information obtained from the former analysis step. For each pixel classified as skin, $\mathrm{p}_{\mathrm{skin}}^n$, in the view n, $0 \leq n < N_{\mathrm{CAM}}$, we check whether

$$\mathrm{p}_{\mathrm{skin}}^n \in P_n(\mathcal{V}^k), \quad 0 \leq k < M, \tag{10}$$



(a) (b)

FIGURE 1: Example of the outputs from the spatial analysis and model fitting modules. In (a), multiview correspondences among heads are correctly established. The projection of the bounding box $\mathcal{B}^0$ containing the head is depicted in white. In (b), voxel reconstruction is applied to $\mathcal{B}^0$ thus obtaining the voxels belonging to the head (green cubes). Model fitting module result is depicted in red.

where $P_n(\cdot)$ is the perspective projection operator from 3D to 2D coordinates on the view n [36]. In this way, $\mathrm{p}^n_{\mathrm{skin}}$ can be identified as being a projection of a voxel of the set $\mathcal{V}^k$ and therefore correctly handled when establishing orientation of multiple heads and faces in later modules. Let us denote with $\mathcal{S}^k_n$ all skin pixels in the nth view classified as belonging to the kth voxel set. It should be recalled that there could be empty sets $\mathcal{S}^k_n$ due to occlusions or under-performance of the skin detection technique. However, tracking information and redundancy among views would allow to overcome this problem.

3.3. Head model fitting

In order to achieve a good fitting performance, a geometrical 3D configuration of human head must be considered. For our research work, an ellipsoid model of human head shape has been adopted. In spite of this fairly simple approximation compared to more complex geometries of head shape [11], head fitting still achieves enough accuracy for our purposes (see Figure 1(b), e.g.).

Let $\mathcal{H}^k = \{\mathbf{c}^k, \mathbf{R}^k, \mathbf{s}^k\}$ be the set of parameters that define the ellipsoid modelling the kth detected human head candidate where $\mathbf{c}^k$ is the center, $\mathbf{R}^k$ the rotation along each axis centered on $\mathbf{c}^k$ and $\mathbf{s}^k$ the length of each axis. After obtaining the set of voxels $\mathcal{V}^k$ belonging to kth candidate head $\mathcal{H}^k$, the ellipsoid shell modelling it is fit to these voxels. Statistic moment analysis is employed to estimate the parameters of the ellipsoid from the centers of the marked voxels thus obtaining a 3D spatial mean $\overline{\mathcal{V}}^k$ and a covariance matrix $\mathbf{C}_{\mathcal{V}^k}$. The covariance can be diagonalized via an eigenvalue decomposition into $\mathbf{C}_{\mathcal{V}^k} = \mathbf{\Phi}\mathbf{\Delta}\mathbf{\Phi}^\top$, where $\mathbf{\Phi}$ is orthonormal and $\mathbf{\Delta}$ is diagonal. Identification of the defining parameters of the estimated ellipsoid $\mathcal{H}^k$ with moment analysis parameters is then straightforward:

$$\mathbf{c}^k = \overline{\mathcal{V}}^k, \qquad \mathbf{R}^k = \mathbf{\Phi}, \qquad \mathbf{s}^k = \mathrm{diag}(\mathbf{\Delta}). \tag{11}$$

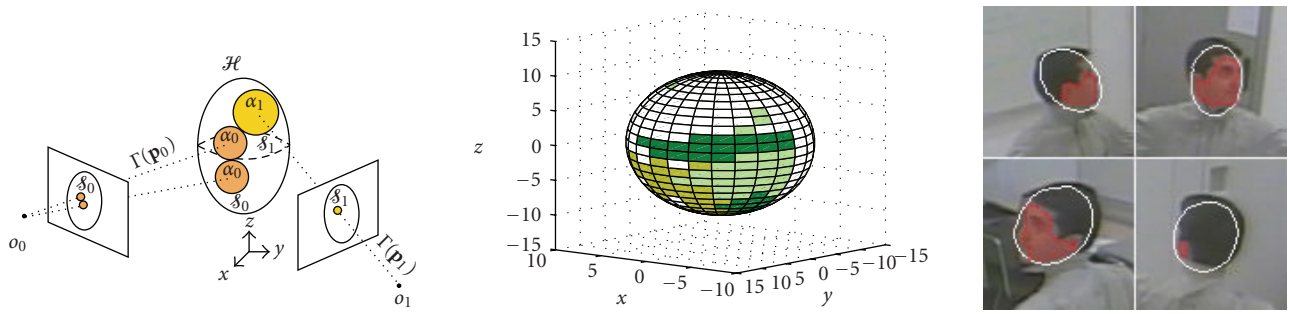
3.4. 3D head appearance generation

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 each person: an ellipsoid estimation $\mathcal{H}^k$ of the head and a set of skin patches at each view belonging to this head $\{\mathcal{S}^k_n\}$, $0 \le n < N_{\mathrm{CAM}}$. The output of this technique is a fusion of color and space information set denoted as Υ^k.

The procedure of information aggregation we define is based on the assumption that all skin patches $\{\mathcal{S}^k_n\}$ 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 $\mathcal{S}^k_n$ from all N_{CAM} views onto the kth 3D ellipsoid model. Formally, for each pixel $\mathrm{p}^k_n \in \mathcal{S}^k_n$, we compute

$$\Gamma(\mathrm{p}^k_n) \equiv P_n^{-1}(\mathrm{p}^k_n) = \mathbf{o}_n + \lambda\mathbf{v}, \quad \lambda \in \mathbb{R}^+, \tag{12}$$

thus obtaining its back-projected ray in the world coordinate frame passing through p^k_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^k_n onto the surface of the ellipsoid modelling the kth head, (12) is substituted into the equation



(a) (b) (c)

Figure 2: In (a), color and spatial information fusion process scheme. Pixels in the set $\mathcal{S}_n^k$ are back-projected onto the surface of the ellipsoid defined by $\mathcal{H}^k$, generating the set $\mathcal{S}_n^k$ with its weighting term α_n^k. 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.

of an ellipsoid defined by the set of parameters $\mathcal{H}^k$ [36]. It gives a quadratic in λ:

$$a\lambda^2 + b\lambda + c = 0. \tag{13}$$

The case of interest will be when (13) 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 Figure 2(a).

This process is applied to all pixels of a given patch $\mathcal{S}_n^k$ obtaining a set $\mathcal{S}_n^k$ containing the 3D points being the intersections of the back-projected skin pixels in the view n with the kth ellipsoid surface. In order to perform a joint analysis of the sets $\{\mathcal{S}_n^k\}$, 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^k can be estimated by projecting a sphere with radius $r = \max(\mathbf{s}^k)$ 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^k should be estimated for each element in $\mathcal{S}_n^k$ but, since the *far-field* condition

$$\max(\mathbf{s}^k) \ll ||\mathbf{c}^k - \mathbf{o}_n||_2, \quad \forall n, \tag{14}$$

is fulfilled, α_n^k can be considered constant for all intersections in $\mathcal{S}_n^k$. A schematic representation of the fusion procedure is depicted in Figure 2(a). Finally, after applying this process to all skin patches, we obtain a fusion of color and spatial information set $\Upsilon^k = \{\mathcal{S}_n^k, \alpha_n^k, \mathcal{H}^k\}$, $0 \le n < N_{\text{CAM}}$, for every head in the scene. A result of this process is shown in Figure 2(b).

3.5. 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. For the sake of simplicity in the notation, let us assume that only one person is present in the scene, thus $\Upsilon^k \equiv \Upsilon$. The observation Ω_t^{V} will be constructed upon the information provided by the set Υ. 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 $\Omega_t^{\text{V}}(n_\theta, n_\phi)$ 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 Figure 3.

Groundtruth information from a training database is employed to compute an average normalized *template face map* centered at $\theta = 0$, namely, $\tilde{\Omega}^{\text{V}}(n_\theta, n_\phi)$, 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 Figure 4.

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

$$\begin{aligned}\Sigma^{\text{V}}(\theta, \Omega^{\text{V}}(n_\theta, n_\phi)) &= \sum_{k_\theta=0}^{N_\theta} \sum_{k_\phi=0}^{N_\phi} \left(1 - \left(\Omega^{\text{V}}(k_\theta, k_\phi) \cdot \tilde{\Omega}^{\text{V}}\left(k_\theta \ominus \left\lfloor \frac{\theta}{\Delta_\theta} \right\rfloor, k_\phi\right)\right)^2\right), \\ N_\theta &= \left\lfloor \frac{2\pi}{\Delta_\theta} \right\rfloor, \qquad N_\phi = \left\lfloor \frac{\pi}{\Delta_\phi} \right\rfloor,\end{aligned} \tag{15}$$

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 $\Omega^{\text{V}}(n_\theta, n_\phi)$. Finally, the weights of the particles are defined as

$$w_t^j(\theta_t^j, \Omega^{\text{V}}(n_\theta, n_\phi)) = \exp\left(-\beta_{\text{V}}\, \Sigma^{\text{V}}(\theta_t^j, \Omega^{\text{V}}(n_\theta, n_\phi))\right). \tag{16}$$

Inverse exponential functions are used in PF applications in order to reflect the assumption that measurement errors are

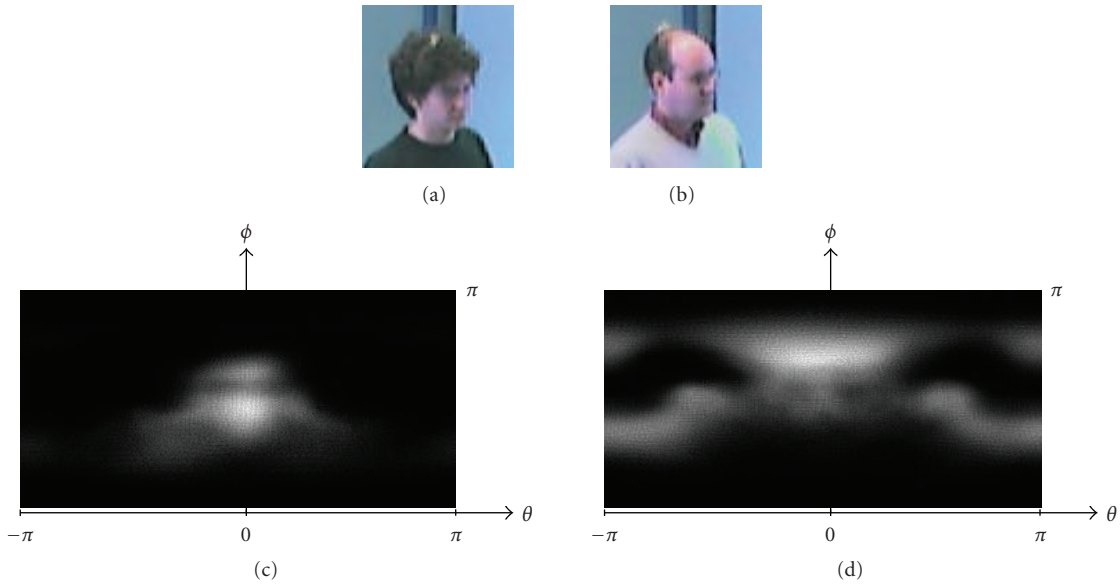


FIGURE 3: Two examples of the Ω_t^V sets containing the visual information that will be fed to the video PF. This set may take different configurations depending on the appearance of the head of the person under study. For our experiments, a quantization step of $\Delta_\theta \times \Delta_\phi = 0.02 \times 0.02$ rads have been employed. These images are courtesy of the University of Karlsruhe.

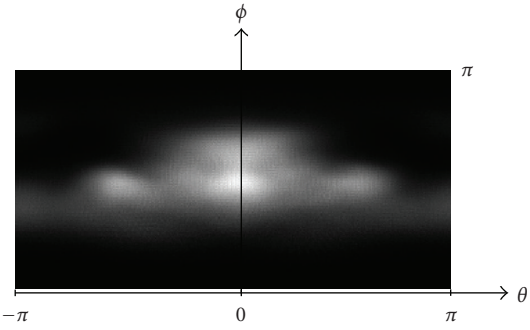


FIGURE 4: Template face map obtained from an annotated training database for 10 different subjects.

Gaussian [17]. 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. The value of β_V is noncrucial and its value allows a faster convergence of the tracking system when $\beta > 1$ [25]. It has been empirically fixed at $\beta_V = 50$.

4. MULTIMICROPHONE HEAD POSE ESTIMATION

In this section, we present a new monomodal approach for estimating the head orientation from acoustic signals, which makes use of the frequency dependence of the head radiation pattern. The proposed method is very efficient in terms of computational load due to its simplicity and also does not require a large aperture microphone array as previous works [12]. All results described in this work were derived using only a set of four T-shaped 4-channel microphone clusters. However, it is not necessary that the microphone clusters have a specific geometry nor to be located at a predefined position.

The acoustic speaker orientation approach presented in this work consists essentially in finding a candidate source location and classifying it as speech or nonspeech, compute the high/low band ratio described in the following sections for each microphone, and finally compute a likelihood evaluation function in order to implement a PF. Since the aim of this work is to determine head orientation, we will assume that the active speaker's locations are known beforehand and they are the same as those used in video. Robust speaker localization in multimicrophone scenario based on SRP-PHAT algorithm has been addressed in our previous research [40].

4.1. Head radiation

Human speakers do not radiate speech uniformly in all directions. In general, any sound source (e.g., a loudspeaker) has a radiation pattern determined by its size and shape and the frequency distribution of the emitted sound. Like any acoustic radiator, the speaker's directivity should increase with frequency and mouth aperture. Infact, the radiation pattern is time-varying during normal speech production, being dependent on lip configuration. There are works that try to simulate the human radiation pattern [41] and other works that accurately measure the human radiation pattern, showing the differences for male and female speaker and using different languages [42].

Figure 5(a) shows the A-weighted typical radiation pattern of a human speaker in horizontal plane passing through his mouth. This radiation pattern shows an attenuation of -2 dB on the side of the speaker (90° or 270°) and -6 dB at his back. Similarly, the vertical radiation pattern is not uni-

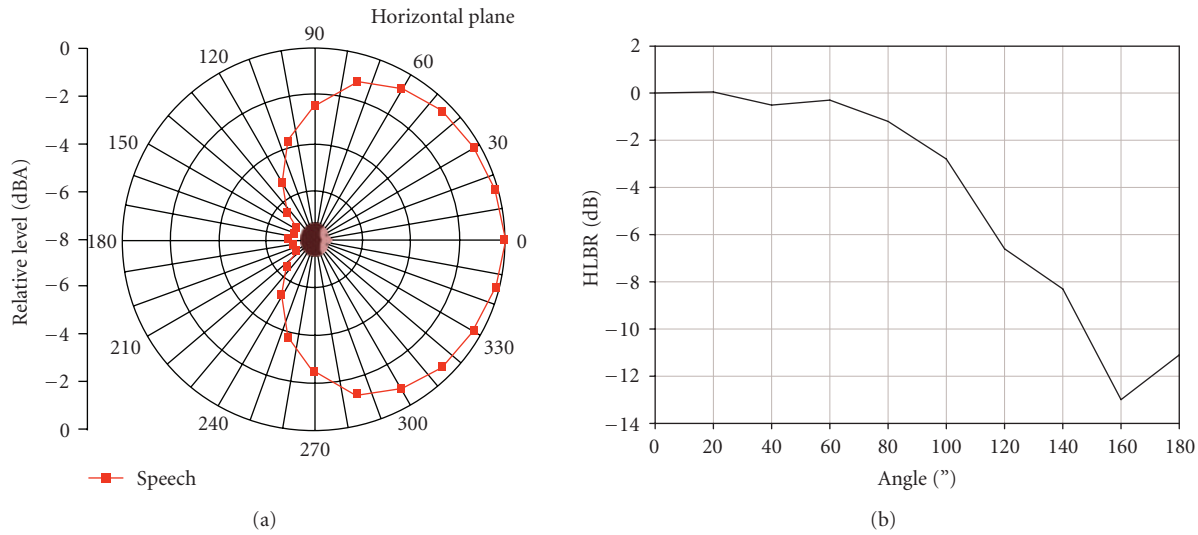


FIGURE 5: In (a), A-weighted head radiation diagram in the horizontal plane. In (b), HLBR of the head radiation pattern.

form, for example, there is about −3 dB attenuation above the speaker head.

The knowledge of the human radiation pattern can be used to estimate the head orientation of an active speaker by simply computing the energy received at each microphone and searching the angle that best fits the radiation pattern with the energy measures. However, this simple approach has several problems since the microphones should be perfectly calibrated and different attenuation at each microphone due to propagation must be accounted for, requiring the use of sound propagation models. In our approach, we propose to keep the computational simplicity using acoustic energy normalization to solve the aforementioned problems.

The energy radiated at 200 Hz by an active speaker is low directional. However, for frequencies above 4 kHz the radiation pattern is highly directive [42]. Based on this fact, we define the high/low band ratio (HLBR) of a radiation pattern as the ratio between high and low bands of frequencies of the radiation pattern and can be observed in Figure 5(b).

Instead of computing the absolute energy received at each microphone, we propose the computation of the HLBR of the acoustic energy. This value is directly comparable across all microphones since, after this normalization, the effects of bad calibration and propagation losses are cancelled.

4.2. High/low band ratio estimation

As for the video case, we assume that the active speaker's location is known beforehand and determined by $\mathbf{c}$ and the vector $\mathbf{r}_i$ from the speaker to each microphone m_i is calculated. The projection of the vector $\mathbf{r}_i$ on the xy plane forms an angle θ_i with the x-axis. Let ρ_i be the value of the HLBR of the acoustic energy at each microphone m_i. The values ρ_i are normalized with a softmax function [43], which is widely used in neural networks, when the output units of a neural network have to be interpreted as posterior probabilities. The softmax normalized HLBR values $\overline{\rho_i}$ are given by

$$\overline{\rho_i} = \frac{e^{k\cdot\rho_i}}{\sum_{k=1}^{n} e^{k\cdot\rho_k}}, \tag{17}$$

where k is a design factor. In our experiments, k is set to 20.

The definition of the softmax function ensures that $\overline{\rho_i}$ lie between 0 and 1 and that their sum is equal to 1.

4.3. Speaker orientation likelihood evaluation

In this work, the HLBR of the head radiation pattern (see Figure 5(b)) has been used as the likelihood evaluation function of the PF. From the values of $\overline{\rho_i}$, we compute a continuous approximation of the HLBR of the head radiation pattern as

$$W(\theta) = \sum_{i=0}^{N_{\text{MICS}}} \overline{\rho_i} * \exp\left(-\left(\frac{|\theta - \theta_i|}{\pi} C\right)^2\right), \tag{18}$$

where the constant C in the interpolation function (18) is a measure of confidence of the $\overline{\rho_i}$ and θ_i estimation.

In this work, C has been chosen as

$$C = \frac{\eta}{\epsilon}, \tag{19}$$

where η is the likelihood of the SRP-PHAT acoustic localization algorithm, and ϵ is a threshold dependent on the number of microphones used [40].

In order to maintain the parallelism with the video counterpart, a cost function is defined as follows, being Ω^{A} the audio observations $W(\theta)$:

$$\Sigma^{\text{A}}(\theta, \Omega^{\text{A}}) = 1 - W(\theta). \tag{20}$$

Finally, the weights of the particles are defined as the visual likelihood evaluation function:

$$w_t^j(\theta_t^j, \Omega^{\mathrm{A}}) = \exp\left(-\beta_{\mathrm{A}}\Sigma^{\mathrm{A}}(\theta, \Omega^{\mathrm{A}})\right). \quad (21)$$

$\beta_{\mathrm{A}} = 100$ provided satisfactory results.

5. MULTIMODAL INTEGRATION

Multimodal head orientation tracking is based on the audio and video technologies described in the previous sections. In our framework, it is expected to have far more observations from the video modality than from the audio modality since persons in the SmartRoom are visible by the cameras during most of the video frames. Moreover, the audio system can estimate the person's head orientation only if she/he is speaking. Hence, the presented approach relies primarily on the video system and the audio information is incorporated to the corresponding video estimates in a multimodal fusion process. This is achieved by first synchronizing the audio and video estimates and fusing the two sources of information.

The combination of audio and video information with particle filters has been addressed in the past for speaker tracking applications. In [19, 44] a multiple people tracking system was based on integrated audio and visual state and observation likelihood components. Thus, the combined probability for audio and video data is obtained by multiplying the corresponding probabilities from the audio and video source, assuming independent estimations by the complementary modalities. In a different context, in [25], the same approach is used for combining different data for articulated body tracking. In [45] multiple speakers were tracked with a set of independent PFs, one for each person. Each PF used a mixture proposal distribution, in which the mixture components were derived from the output of single-cue trackers. In [18] the joint audio visual probability for speaker tracking was computed as a weighted average of the single modality probabilities.

In this paper, we will report the advantages of the two modalities fusion at the data level by comparing it to a decision level fusion. The first decision level fusion that we will consider will be based on two independent PF for the audio and video modalities. Thus, the estimated angle will be computed as a linear combination of the audio and video estimations. A second strategy will also consider two independent particle filters, but the estimated angle will be computed as a joint expectation over the audio and video particles. These two simple strategies will be compared to the data level fusion that we will approach computing the combined probability for the audio and video data as in [19, 44].

5.1. Decision level fusion

Two strategies are presented to perform an information fusion at decision level.

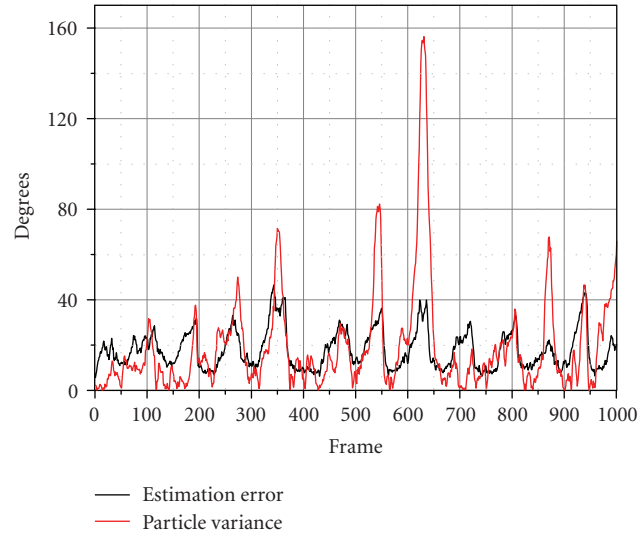


Figure 6: Pan angle estimation error is correlated with the dispersion of the particle thus allowing the construction of multimodal estimators.

(i) *Linear combination of monomodal angle estimations*

The pan angle estimation provided by the audio and video particle filters, Θ_t^{A} and Θ_t^{V}, respectively, are linearly combined to produce $\Theta_t^{\mathrm{AV}_1}$ according to the formula

$$\Theta_t^{\mathrm{AV}_1} = \frac{1}{1/\sigma_t^{\mathrm{A}^2} + 1/\sigma_t^{\mathrm{V}^2}}\left(\frac{1}{\sigma_t^{\mathrm{A}^2}}\Theta_t^{\mathrm{A}} + \frac{1}{\sigma_t^{\mathrm{V}^2}}\Theta_t^{\mathrm{V}}\right), \quad (22)$$

where $\sigma_t^{\mathrm{A}^2}$ and $\sigma_t^{\mathrm{V}^2}$ refer to the variance of the audio and video estimations after a normalization process. Moreover, this variance figure (related to the dispersion of the particles) can be understood as a magnitude related with the estimation error. This effect is depicted in Figure 6 shown as a correlation between the pan angle estimation error and the variance.

(ii) *Particle combination*

A decision level fusion may be performed before the expectation is taken at each monomodal PF (see (8)). Indeed, particles generated by each monomodal PF contain information about the sampled audio and video *pdf*s: $p(\theta_t \mid \Omega_{1:t}^{\mathrm{A}})$ and $p(\theta_t \mid \Omega_{1:t}^{\mathrm{V}})$. A joint expectation can be computed over the particles coming from audio and video PFs as

$$\Theta_t^{\mathrm{AV}_2} = \mathbb{E}\left[\theta_t \mid \Omega_{1:t}^{\mathrm{A}}, \Omega_{1:t}^{\mathrm{V}}\right] \approx \sum_{j=1}^{N_s}\left(w_t^{\mathrm{A},j}\theta_t^{\mathrm{A},j} + w_t^{\mathrm{V},j}\theta_t^{\mathrm{V},j}\right), \quad (23)$$

enforcing

$$\sum_{j=1}^{N_s} w_t^{\mathrm{A},j} + \sum_{j=1}^{N_s} w_t^{\mathrm{V},j} = 1. \quad (24)$$

(a) (b)

Figure 7: Images from two experimental cases. In (a), speaker is bowing his head towards the laptop and video-based head orientation estimation does not produce an accurate result (red vector) while audio estimation (green vector) generates a more accurate output. Estimation reliability is proportional to vector length. In (b), an example where both estimators output a correct result.

5.2. Data level fusion

Video PF estimates the head orientation angle taking into account that the frontal part of the face defines the orientation. On the other hand, audio PF estimated this angle by exploiting the fact that the maximum of the HLBR function of the head radiation pattern corresponds to the mouth region. Multimodal information fusion at data level has been done by taking into account that speech is produced by the frontal part of the head. This correlation between the two modalities is modeled in this work by defining a joint likelihood function $p(\theta_t \mid \Omega^{\mathrm{A}}_{1:t}, \Omega^{\mathrm{V}}_{1:t})$ which exploits the dependence between audio and video sources. In this article, multimodal weights have been defined as

$$w_t^{\mathrm{MM},j}(\theta_t^j, \Omega_t^{\mathrm{A}}, \Omega_t^{\mathrm{V}}) = \exp\left(-\beta_{\mathrm{MM}}\left(\lambda_{\mathrm{A}}\Sigma^{\mathrm{A}}(\theta_t^j, \Omega_t^{\mathrm{A}}) + \lambda_{\mathrm{V}}\Sigma^{\mathrm{V}}(\theta_t^j, \Omega_t^{\mathrm{V}})\right)\right), \quad (25)$$

where λ_{A} and λ_{V} are empirically estimated weighting parameters controlling the influence of each modality. After comparing the performance of the monomodal estimators (see Section 6), parameters λ_{A} and λ_{B} have been set for our experiments as $\lambda_{\mathrm{A}} = 0.6$, $\lambda_{\mathrm{V}} = 0.4$ providing satisfactory results. The convergence parameter has been set at $\beta_{\mathrm{MM}} = 100$.

6. RESULTS

In order to evaluate the performance of the proposed algorithms, we employed the CLEAR 2006 head pose database [31] containing a set of scenes in an indoor scenario were a person is giving a talk, for approximately 15 minutes. In order to provide meaningful and comparable results among mono- and multimodal approaches, the subject under study in this evaluation database is always speaking, that is, there is always audio and video information available. The analysis sequences were recorded with 4 fully calibrated cameras with a resolution of 720×576 pixels at 25 fps and 4 microphone cluster arrays with a sampling frequency of 44 KHz. All audio and video sensors were synchronized. Head localization is assumed to be available since the aim of our research is at estimating its orientation. Nevertheless, results on head localization have been specifically reported by the authors in [15, 46]. Even though a more complete database might be devised, this is the only existing database designed for this task up to authors knowledge.

Table 1: Quantitative results for the four presented systems showing that multimodal approaches outperform monomodal approaches.

Method	PMAE (°)	PCC (%)	PCCR (%)
Video	59.52	24.68	64.21
Audio	47.84	31.84	71.90
MM Feature Fusion Type 1	49.09	28.21	73.29
MM Feature Fusion Type 2	44.04	34.54	75.27
MM Data Fusion	**30.61**	**48.99**	**83.69**

The metrics proposed in [31] for head pose evaluation have been adopted: the pan mean average error (PMAE), that measures precision of the head orientation angle in terms of degrees; the pan correct classification (PCC), which shows the ability of the system to correctly classify the head position within 8 classes spanning 45° each; and the pan correct classification within a range PCC, which shows the performance of the system when classifying the head pose within 8 classes allowing a classification error of ±1 adjacent class.

For all the experiments conducted in this article, a fixed number of particles have been set for every PF, $N_s = 100$. Experimental results proved that employing more particles does not report in a better performance of the system.

The four systems presented in this paper (video, audio, and multimodal fusion at decision and data level) have been evaluated and these 3 measures computed in order to compare their performance. Table 1 summarizes the obtained results where multimodal approaches almost always outperform monomodal techniques as expected. Improvements achieved by multimodal approaches are twofold. First, error in the estimation of the angle (PMAE) decreases due to the combination of estimators and, secondly, classification performance scores (PCC and PCC) increase since failures in one modality are compensated by the other. Compared to the results provided by the CLEAR 2006 evaluation [31], our system would be ranked on the 2nd position over 5 participants. Visual results are provided in Figure 7 showing that

multimodal approaches allow enhancing results when one modality fails.

7. CONCLUSIONS AND FUTURE WORK

The use of particle filters has been proved to be useful as a unified framework for the estimation of the head orientation for both monomodal and multimodal cases in terms of accuracy and robustness over the CLEAR 2006 evaluation database. In monomodal head pose estimation, good results have been obtained with a video estimation based on a 3D reconstruction of the head and, especially, with a novel audio estimator based on the directivity characteristics of the head radiation pattern. In multimodal head pose estimation, slightly better results have been obtained by a linear combination of those monomodal estimators and even better results have been reached by particle combination at a decision level. However, in the current scenario, the use of a joint particle filter for fusion of video and audio streams at data level has yielded the best results, achieving a relative 42% reduction of the classification error rate from the best monomodal estimation.

Future research lines aim at designing adaptive modality weighting algorithms in the multimodal data level fusion estimator to automatically set values for λ_A and λ_B. Analysis of the produced data towards tracking attention of multiple people in meetings and understanding behaviors of individuals is under study.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Andrey Temko for fruitful discussions.

REFERENCES

[1] M. Black, F. Berard, A. Jepson, et al., "The digital office: overview," in *Proceedings of the AAAI Spring Symposium on Intelligent Environments*, pp. 98–102, Palo Alto, Calif, USA, March 1998.

[2] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox, "Room with a rear view: meeting capture in a multimedia conference room," *IEEE Multimedia*, vol. 7, no. 4, pp. 48–54, 2000.

[3] "CHIL-Computers in the Human Interaction Loop," http://chil.server.de/.

[4] C. Wang, S. Griebel, and M. Brandstein, "Robust automatic video-conferencing with multiple cameras and microphones," in *Proceedings of IEEE International Conference on Multi-Media and Expo (ICME '00)*, vol. 3, pp. 1585–1588, New York, NY, USA, July-August 2000.

[5] P. Ballard and G. C. Stockman, "Controlling a computer via facial aspect," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 4, pp. 669–677, 1995.

[6] T. Horprasert, Y. Yacoob, and L. S. Davis, "Computing 3-D head orientation from a monocular image sequence," in *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 242–247, Killington, Vt, USA, October 1996.

[7] R. Rae and H. J. Ritter, "Recognition of human head orientation based on artificial neural networks," *IEEE Transactions on Neural Networks*, vol. 9, no. 2, pp. 257–265, 1998.

[8] M. Voit, K. Nickel, and R. Stiefelhagen, "Neural network-based head pose estimation and multi-view fusion," in *Proceedings of the 1st International CLEAR Evaluation Workshop (CLEAR '06)*, vol. 4122 of *Lecture Notes on Computer Science*, pp. 291–299, Southampton, UK, April 2006.

[9] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, "Head pose estimation on low resolution images," in *Proceedings of the 1st International CLEAR Evaluation Workshop (CLEAR '06)*, vol. 4122 of *Lecture Notes on Computer Science*, pp. 270–280, Southampton, UK, April 2006.

[10] L. Zhao, G. Pingali, and I. Carlbom, "Real-time head orientation estimation using neural networks," in *Proceedings of IEEE International Conference on Image Processing (ICIP '02)*, vol. 1, pp. 297–300, Rochester, NY, USA, September 2002.

[11] X. L. C. Brolly, C. Stratelos, and J. B. Mulligan, "Model-based head pose estimation for air-traffic controllers," in *Proceedings of IEEE International Conference on Image Processing (ICIP '03)*, vol. 2, pp. 113–116, Barcelona, Spain, September 2003.

[12] J. M. Sachar and H. F. Silverman, "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 4, pp. 65–68, Montreal, Canada, May 2004.

[13] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 2337–2340, Lisbon, Spain, September 2005.

[14] C. Segura, C. Canton-Ferrer, A. Abad, J. R. Casas, and J. Hernando, "Multimodal head orientation towards attention tracking in smart rooms," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, vol. 2, pp. 681–684, Honolulu, Hawaii, USA, April 2007.

[15] C. Canton-Ferrer, J. R. Casas, and M. Pardàs, "Fusion of multiple viewpoint information towards 3D face robust orientation detection," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 2, pp. 366–369, Genova, Italy, September 2005.

[16] H. R. Hashemipour, S. Roy, and A. J. Laub, "Decentralized structures for parallel Kalman filtering," *IEEE Transactions on Automatic Control*, vol. 33, no. 1, pp. 88–94, 1988.

[17] M. Isard and A. Blake, "CONDENSATION—conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[18] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI '05)*, pp. 61–68, Torento, Italy, October 2005.

[19] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 601–616, 2007.

[20] J. R. Casas, R. Stiefelhagen, K. Bernardin, et al., "Multicamera/multi-microphone system design for continuous room monitoring," Deliverable CHIL-WP4-D4.1-V2.1-2004-07-08-CO, CHIL—IP506909—Computers in the Human Interaction Loop, July 2004.

[21] R. Stiefelhagen, "Tracking focus of attention in meetings," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI '02)*, pp. 273–280, Pittsburgh, Pa, USA, October 2002.

[22] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*, Springer, New York, NY, USA, 2nd edition, 1997.

[23] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[24] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proceedings F—Radar and Signal Processing*, vol. 140, no. 2, pp. 107–113, 1993.

[25] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 185–205, 2005.

[26] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 173–185, 2002.

[27] C. Wang and M. Brandstein, "Robust head pose estimation by machine learning," in *Proceedings of IEEE International Conference on Image Processing (ICIP '00)*, vol. 3, pp. 210–213, Vancouver, BC, Canada, September 2000.

[28] Y. Matsumoto and A. Zelinsky, "An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 499–504, Grenoble, France, March 2000.

[29] M.-Y. Chen and A. Hauptmann, "Towards robust face recognition from multiple views," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 2, pp. 1191–1194, Taipei, Taiwan, June 2004.

[30] Z. Zhang, Y. Hu, M. Liu, and T. Huang, "Head pose estimation in seminar rooms using multi view face detectors," in *Proceedings of the 1st International CLEAR Evaluation Workshop (CLEAR '06)*, vol. 4122 of *Lecture Notes on Computer Science*, pp. 299–304, Southampton, UK, April 2006.

[31] "CLEAR Evaluation Campaign," 2006, http://www.clear-evaluation.org/.

[32] O. Lanz, "Approximate Bayesian multibody tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1436–1449, 2006.

[33] B. Wu, V. K. Singh, R. Nevatia, and C.-W. Chu, "Speaker tracking in seminars by human body detection," in *Proceedings of the 1st International CLEAR Evaluation Workshop (CLEAR '06)*, vol. 4122 of *Lecture Notes on Computer Science*, pp. 119–126, Southampton, UK, April 2006.

[34] A. Pnevmatikakis and L. Polymenakos, "2D person tracking using Kalman filtering and adaptive background learning in a feedback loop," in *Proceedings of the 1st International CLEAR Evaluation Workshop (CLEAR '06)*, vol. 4122 of *Lecture Notes on Computer Science*, pp. 151–160, Southampton, UK, April 2006.

[35] C. Canton-Ferrer, J. R. Casas, and M. Pardàs, "Towards a Bayesian approach to robust finding correspondences in multiple view geometry environments," in *Proceedings of the 5th International Conference on Computational Science (ICCS '05)*, vol. 3515 of *Lecture Notes in Computer Science*, pp. 281–289, Atlanta, Ga, USA, May 2005.

[36] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, 2004.

[37] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, vol. 2, pp. 252–258, Fort Collins, Colo, USA, June 1999.

[38] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, "Articulated body posture estimation from multi-camera voxel data," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 1, pp. 455–460, Kauai, Hawaii, USA, December 2001.

[39] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002.

[40] A. Abad, C. Segura, D. Macho, J. Hernando, and C. Nadeu, "Audio person tracking in a smart-room environment," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, Lisboa, Portugal, September 2005.

[41] P. C. Meuse and H. F. Silverman, "Characterization of talker radiation pattern using a microphone array," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, vol. 2, pp. 257–260, Adelaide, SA, Australia, April 1994.

[42] W. T. Chu and A. C. Warnock, "Detailed directivity of sound fields around human talkers," Tech. Rep., Institute for Research in Construction, Ontario, Canada, 2002.

[43] A. Tuerk and S. J. Young, "Polynomial softmax functions for pattern classification," 2001.

[44] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 5, pp. 881–884, Montreal, Canada, May 2004.

[45] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 485–494, 2004.

[46] A. Lopez, C. Canton-Ferrer, and J. R. Casas, "Multi-person 3D tracking with particle filters on voxels," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, Honolulu, Hawaii, USA, April 2007.