MULTIMODAL HEAD ORIENTATION TOWARDS ATTENTION TRACKING IN SMARTROOMS

C. Segura, C. Canton-Ferrer, A. Abad, J.R. Casas, J. Hernando

Signal Theory and Communications Department
Technical University of Catalonia (UPC)
Barcelona, Spain

ABSTRACT

This paper presents a multimodal approach to head pose estimation and 3D gaze orientation of individuals in a SmartRoom environment equipped with multiple cameras and microphones. We first introduce the two monomodal approaches as reference. 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. Two multimodal information fusion schemes working at data and decision levels are analyzed in terms of accuracy and robustness of the estimation. Experimental results conducted over the CLEAR 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.

Index Terms— Data fusion, Head orientation, Speaker orientation, multi-camera image analysis

1. INTRODUCTION

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. The head orientation of a person provides important clues in order to give a better service in such scenarios. This knowledge allows a better understanding of what users do or what they refer to. In applications that require human-computer interaction, accurate head pose estimation can 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. Moreover, other technologies such as Face Identification or Automatic Speech Recognition could exploit available head orientation information and improve their performance by selecting a subset of sensors (cameras and microphones) adequately located for the task.

Previous approaches to estimate the head pose have mostly used video technologies [1, 2]. 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 [3], 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 [4], which is a variation on GCF acoustic localization algorithm.

This material is based upon work partially supported by the IST programme of the EU through the IP IST-2004-506909 CHIL, by TEC2004-01914 and TIN-2005-08852 projects of the Spanish Government.

In this paper we present two multimodal fusion algorithms aiming to estimate the head pose using audiovisual information. The proposed architecture combines data and features extracted from a former system from the authors based on video [5] 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 remainder of this paper is organized as follows. In the next section we introduce the monomodal video head pose estimation. In Section 3, we present the audio single modality system for speaker orientation estimation. In Section 4 we propose two methods to fuse audio and video modalities combining the estimations provided by each system at the data and decision levels. In the following section, the performance obtained by each system is discussed and we conclude the paper in Section 6.

2. VIDEO HEAD POSE ESTIMATION

Methods for head pose estimation proposed in the literature [1] use to follow 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 and high-quality images are available. For the applications addressed in our work, such conditions are usually difficult to satisfy. Methods which rely on a detailed feature analysis followed by head model fitting would fail under these circumstances.

Most of the existing approaches are based on monocular analysis of images but few have addressed the multi-ocular case for face or head analysis [5]. In this context, appearance-based approaches [2] tend to achieve satisfactory results with low resolution images. However, since head orientation estimation is posed as a classification problem, output angle resolution is limited to a discrete set. Typically, 8 categories are employed [6] thus leading to a resolution of 45^{o}. When performing a multimodal fusion, informative video outputs are desired, thus preferring data analysis methods providing a real valued angle output. The next subsection reviews the monomodal visual approach presented in [5].

2.1. Multi-view Head Pose Estimation

Since the aim of this work is to determine head orientation, we separate this task from the task of head localization. Therefore, the 3D position of the head of the person of interest is assumed to be known and determined by a bounding box $\mathcal{B}$, already available as an input to the head orientation algorithm. Automatic 3D head detection in multi-view sequences has been addressed in our previous research [7]. The center and size of the bounding box $\mathcal{B}$ allow defining an ellipsoid model of the head $\mathcal{H}$ as shown in Fig.1a.

Color information within $\mathcal{B}$ is processed to extract skin colored pixels in every image by mean of a classifier that learns the statistics of the skin color. Let us denote with $\mathcal{S}_n$ all pixels classified as skin in the n-th view. It should be noted that there could be empty sets $\mathcal{S}_n$ due to occlusions or poor performance of the skin classifier. An example of skin classification is shown in Fig.1a.

In order to estimate face orientation, we assume that all skin patches $\{\mathcal{S}_n\}$, $0 \leq n < N$, are projections of a region of the surface of the estimated ellipsoid defining the head of a person. Hence, color and space information are combined to produce a synthetic reconstruction of the head and face appearance in 3D. This is accomplished by back-projecting the skin pixels of $\mathcal{S}_n$ from all N views onto the 3D ellipsoid model. Formally, for each pixel $\mathrm{p}_n \in \mathcal{S}_n$, we compute

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

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 $\mathbf{v}$. Term $P_n(\cdot)$ is the perspective projection operator from 3D to 2D coordinates on the view n. A scheme of this process is shown in Fig.1c. This information is considered by the set $\boldsymbol{\mathcal{S}}_n$ containing the 3D points. An associated weighting factor α_n takes into account the actual surface of the ellipsoid represented by a single pixel in view n in order to quantize the effect of the different distances from the center of the object to each camera. These weights are normalized such that $\sum_{n=0}^{N-1} \alpha_n = 1$. Finally, after applying this process to all skin patches we obtain a set $\Omega = \{\boldsymbol{\mathcal{S}}_n, \alpha_n, \mathcal{H}\}$, $0 \leq n < N$, combining color and spatial information. Tracking over time is performed by a Kalman filter. An example of this fusion is shown in Fig.1b.

2.2. Head and Face Orientation

Head and face orientation is computed from the set Ω. The angle to be estimated for our purposes in the SmartRoom scenario has been chosen as a direction onto the xy plane. The orientation angle $\hat{\theta}_V$ is estimated by the computation of the weighted centroid of the fusion data Ω as

$$\mathbf{d}_V = \frac{1}{\sum_{n=0}^{N-1} |\boldsymbol{\mathcal{S}}_n|} \sum_{n=0}^{N-1} \alpha_n \sum_{\mathbf{p}_n \in \boldsymbol{\mathcal{S}}_n} (\mathbf{p}_n - \mathbf{c}), \tag{2}$$

$$\hat{\theta}_V = \tan^{-1}\left(\mathbf{d}_{V_y} / \mathbf{d}_{V_x}\right), \tag{3}$$

where $|\boldsymbol{\mathcal{S}}_n|$ denotes the number of elements (3D intersections) in the set and $\mathbf{c}$ is the center of the head $\mathcal{H}$. Results for this technique have been reported in [5].

3. MULTI-MICROPHONE HEAD POSE ESTIMATION

In this section we present a new monomodal approach for estimating the head orientation from acoustic signals. 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 [3]. All results described in this work were derived using only a set of four T-shaped 4-channel microphone clusters. Since the aim of this research is to determine head orientation, we will assume that the active speaker's location is known beforehand. Robust speaker localization in multi-microphone scenario has been addressed in our previous reseach [8].

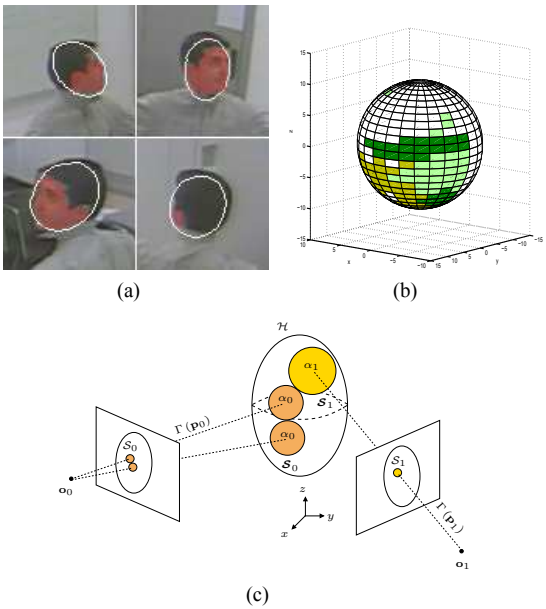


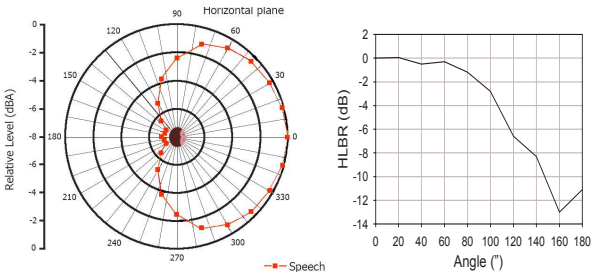
Fig. 1. In (a) skin patches are plotted in red and the ellipsoid fitting in white and in (b), result of information fusion obtaining a synthetic reconstruction of face appearance from images. In (c), color and spatial information fusion process scheme. Pixels in the set $\mathcal{S}_n$ are back-projected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

3.1. 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. However, 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 [9] and other works that accurately measure the human radiation pattern, showing the differences for male and female talker and using different languages as English and French [10].

Fig.2a 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 -2dB on the side of the speaker (90^0 or 270^o) and -6dB at his back. Similarly, the vertical radiation pattern is not uniform, e.g. there is about -3dB 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



(a) (b)

Fig. 2. In (a), A-weighted head radiation diagram in the horizontal plane. In (b), HLBR of the head radiation pattern.

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, thus requiring the use of sound propagation models. In our approach, we propose to keep the computational simplicity by using acoustic energy normalization to solve the aforementioned problems.

The energy radiated at 200Hz by an active speaker is low directional, however, for frequencies above 4kHz the radiation pattern is highly directive [10]. We make use of this fact to define the High/Low Band Ratio (HLBR) of a radiation pattern. The HLBR of a radiation pattern is defined as the ratio between high and low bands of frequencies of the radiation pattern and can be observed in Fig.2b.

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

3.2. Orientation Estimation

As for the visual 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. Each vector $\mathbf{r}_i$ forms an angle θ_i with the x-axis in the xy plane. We define a function $W(\theta)$ that relates the HLBR of acoustic energy at each microphone, denoted by w_i with each angle θ_i. Weights w_i are normalized fulfilling $\sum_{i=1}^{n} w_n = 1$. The estimated speaker orientation can be computed by searching the angle that maximizes the correlation between the HLBR of a radiated pattern $G(\theta)$ and the HLBR of the acoustic energy measured at each microphone.

$$W(\theta) = \sum_{i=0}^{N_{MICS}} \delta(\theta - \theta_i) \cdot w_i, \quad (4)$$

$$\hat{\theta}_A = \underset{\theta}{\operatorname{argmax}}\, G(\theta) * W(\theta). \quad (5)$$

Finally, a Kalman filter is employed to smooth the estimation.

4. MULTI-MODAL 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.

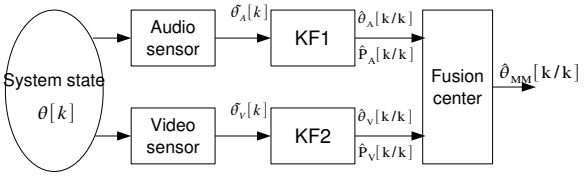


Fig. 3. Structure of the decentralized Kalman filter. The fusion center combines the local estimates to compute a global estimate of the system state.

Two methods for combining Audio and Video single modalities are proposed. First, combining the estimations at a Decision Level by means of a decentralized Kalman filter, and secondly, fusing the two sources of information at Data Level.

4.1. Decision Level Fusion

The decentralized Kalman filter [11] is used for the fusion of audio and video position estimates. As shown in Fig.3, the system can be divided in two modules associated with the audio and video systems. Each modality computes a local a-posteriori estimate $\hat{\theta}_A[k|k]$, $\hat{\theta}_V[k|k]$ of the person head orientation using a local Kalman filter (KF1 and KF2, respectively), based on the corresponding observations $\tilde{\theta}_A[k]$, $\tilde{\theta}_V[k]$. These partial estimates are then combined to provide a global state estimate $\hat{\theta}_{MM}[k|k]$ at the fusion step.

The global estimate of the system state is obtained by weighting the global and local state estimate with the global error covariance matrix $\mathbf{P}_{MM}[k|k]$ and their counterparts $\mathbf{P}_A[k|k]$ and $\mathbf{P}_V[k|k]$ at the audio and video systems.

4.2. Data Level Fusion

Multimodal data fusion at data level has been achieved taking into account that speech is produced by the frontal part of the head. We propose a modification of the presented monomodal video technique in order to include the HLBR of the acoustic energy function $W(\theta)$ from Eq.4. Vectors $\mathbf{r}_i$ going from the head center to each microphone intersect the ellipsoid head model $\mathcal{H}$ in several 3D points defined by the set $\boldsymbol{\mathcal{A}}$. The points having lowest HLBR of acoustic energy are rejected since we expect them to be associated with the microphones behind the focus of attention of the speaker. The weighted centroid of the points in the set $\boldsymbol{\mathcal{A}}$ with respect to the center of the head model $\mathcal{H}$, $\mathbf{c}$, can be defined as:

$$\mathbf{d}_A = \frac{1}{M} \sum_{i=1}^{M} w_i \left(\boldsymbol{\mathcal{A}}_i - \mathbf{c} \right). \quad (6)$$

Finally, the orientation angle $\hat{\theta}_{MM}$ is estimated by the computation of the centroid $\mathbf{d}_{MM}$ that is an average of the previously computed video $\mathbf{d}_V$ and audio $\mathbf{d}_A$ centroids:

$$\mathbf{d}_{MM} = \frac{1}{2} \left(\mathbf{d}_V + \mathbf{d}_A \right), \quad (7)$$

$$\hat{\theta}_{MM} = \tan^{-1} \left(\mathbf{d}_{MM_y} / \mathbf{d}_{MM_x} \right). \quad (8)$$

(a) (b)

Fig. 4. 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 were both estimators outputs a correct result.

5. RESULTS

In order to evaluate the performance of the proposed algorithms, we employed the CLEAR head pose database [6] containing a set of scenes in an indoor scenario were a person is giving a talk, for a total of approximately 15 min. The analysis sequences were recorded with 4 fully calibrated cameras and 4 microphone cluster arrays, with all both sensors synchronized.

The metrics proposed in [6] for head pose evaluation have been adopted: the **P**an **M**ean **A**verage **E**rror (*PMAE*), that measures precision of the head orientation angle in terms of degrees; the **P**an **C**orrect **C**lassification (*PCC*), which shows the ability of the system to correctly classify the head position within 8 classes spanning 45^o each; and the **P**an **C**orrect **C**lassification within a **R**ange *PCC*, shows the performance of the system when classifying the head pose within 8 classes allowing a classification error of ± 1 adjacent class.

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 Evaluation [6], our system would be ranked on the 2nd position over 5 participants. Visual results are provided in Fig.4 showing that multimodal approaches allow enhancing results when one modality fails.

Method	PMAE (^{o})	PCC (%)	PCCR (%)
Video	47.33	32.88	71.39
Audio	53.14	28.47	69.17
MM Feature Fusion	48.53	28.92	73.47
MM Data Fusion	**35.79**	**38.96**	**83.25**

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

6. CONCLUSIONS

This paper presents and compares head pose estimation techniques based on both video and audio modalities and then combined in two different multimodal fusion schemes. Moreover, a novel head orientation estimator based on audio information is introduced.

These techniques allow integrating information from two sources in order to enhance the estimation of the head orientation angle by decreasing its estimation error and improving the classification rate. In the current scenario, it has been shown that a simple Data Level fusion technique outperformed a sophisticated Decision Level fusion scheme. Quantitative results proved the effectiveness of our approach achieving a relative 41.45% reduction of the classification error rate from the best monomodal estimation (video) to the best multimodal estimation (data fusion).

Future research within this topic involve analysis of the data towards tracking attention of multiple people in meetings and understanding behaviors of individuals. Also, more efficient fusion schemes are under research.

7. REFERENCES

[1] X. Brolly, C. Stratelos, and J. Mulligan, "Model-based head pose estimation for air-traffic controllers," in *Proc. ICIP*, 2003, pp. 113–116.

[2] M. Voit, K. Nickel, and R. Stiefelhagen, "Neural network-based head pose estimation and multi-view fusion," in *Proc. CLEAR Evaluation Workshop*, 2006.

[3] J.M. Sachar and H.F. Silverman, "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array," in *Proc. ICASSP*, 2004, vol. 4, pp. 65–68.

[4] 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 *Proc. Interspeech*, 2005.

[5] C. Canton-Ferrer, J.R. Casas, and M. Pardas, "Fusion of multiple viewpoint information towards 3D face robust orientation detection," in *Proc. ICIP*, 2005, vol. 2, pp. 366–369.

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

[7] C. Canton-Ferrer, J. R. Casas, and M. Pardàs, "Towards a Bayesian approach to robust finding correspondences in multiple view geometry environments," in *Lecture Notes on Computer Science*, 2005, vol. 3515, pp. 281–289.

[8] A. Abad, C. Segura, D. Macho, J. Hernando, and C. Nadeu, "Audio person tracking in a SmartRoom environment," in *Proc. Interspeech*, 2005.

[9] P.C. Meuse and H.F. Silverman, "Characterization of talker radiation pattern using a microphonearray," in *Proc. ICASSP*, 1994, vol. 2, pp. 257–260.

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

[11] H. R. Hashemipour, S. Roy, and J. Laub, "Decentralized structures for parallel Kalman ltering," *Automatic Control, IEEE Tran. on*, vol. 33, no. 1, pp. 88–93, 1988.