Multimodal Real-Time Focus of Attention Estimation in SmartRooms

C. Canton-Ferrer, C. Segura, M. Pardàs, J.R. Casas, J. Hernando
Technical University of Catalonia
Barcelona, Spain
{ccanton,csegura,montse,josep,javier}@gps.tsc.upc.edu

Abstract

This paper presents an overview of our work on real-time multimodal tracking focus of attention of multiple persons in a SmartRoom scenario. Redundancy among cameras is exploited to generate a 3D discrete reconstruction of the space. This information is fed to a novel low complexity Monte Carlo based tracking scheme. Estimated locations of people in the room are used to automatically determine their head positions. Head orientation of every person is computed using video and audio separately and then a multimodal estimation is produced by combining data at feature level employing a decentralized Kalman filter. Finally, participants' focus attention is estimated by means of two geometric descriptors: the attention cone and the attention map. Experiments conducted over annotated databases yield quantitative results proving the effectiveness of the presented approach.

1. Introduction

Multimodal analysis of signals towards providing reliable and informative data related to human activities has gained a lot of interest in the recent years. This analysis is important to build attentive interfaces aiming at supporting humans in various tasks and situations. Examples of these intelligent environments include the "digital office" [7], "intelligent house", "intelligent classroom" and "smart conferencing rooms" [1, 13]. One important aspect for the analysis and understanding of human-human or human-computer interaction, is to somehow automatically gain knowledge about people's focus of attention, i.e.the knowledge about the targets, objects, or other people with whom they interact. The analysis of focus of attention dynamics can for instance give relevant information about who is talking to whom [27], the roles of people, their dominance and possibly ranks, the structure of interaction [21], as well as about the type of interaction going on (for example discussion vs. presentation by one person).

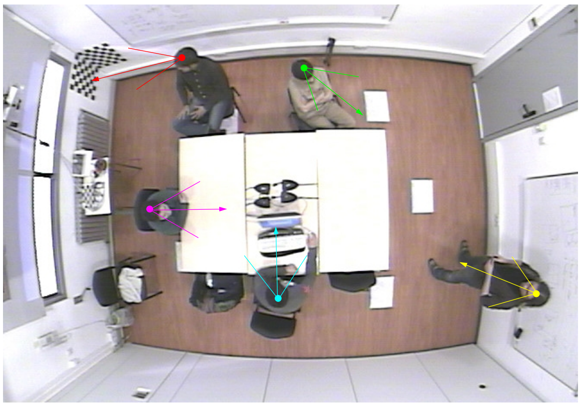


Figure 1. Focus of attention estimation of a group of people may allow the system to identify this gathering as a lecture where somebody is distracted looking through the window and somebody else is checking his email at the computer.

This paper presents a real-time operating system for multi-person focus of attention (FoA) estimation in an indoor scenario equipped with multiple cameras and far-field microphones. A first step to determine someone's FoA is to find out in which direction the person looks. There are two contributing factors in the formation of where a person looks: head orientation and eye orientation. In this work head orientation is considered as a sufficient cue to detect a person's direction of attention. Relevant psychological literature offers a number of convincing arguments for this approach [4] and the feasibility of this approach is demonstrated experimentally in this paper.

Two already developed technologies are combined towards estimating FoA in real-time: person tracking and multimodal head orientation. Person tracking is addressed by a novel Monte Carlo based technique that noticeably reduces the computational load of the process and still yields to satisfactory results. Head orientation is performed in both audio and video domains based on the algorithm presented in [24]. Finally, FoA is addressed by two proposed descriptors that can be computed fast and allow the detection and classification of some cases of interest such as detecting the regions of maximum FoA or recognizing interactions between participants. Presented results show the effectiveness of the proposed system at an average rate of 10 fps.

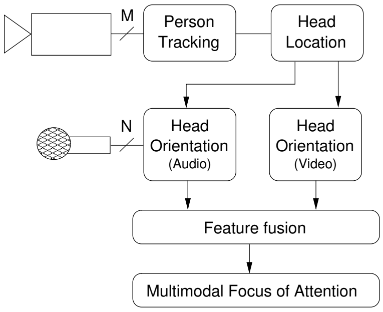


Figure 2. System flowchart.

2. System description

According to flowchart in Fig.2, the proposed system comprises three main signal processing modules: person tracking, head orientation estimation and focus of attention estimation. For a given frame in the video sequence, a set of M images are obtained from the M cameras (see a sample in Fig.3a). Each camera is modeled using a pinhole camera model based on perspective projection with camera calibration information available. Foreground regions from input images are obtained using a segmentation algorithm based on Stauffer-Grimson's background learning and substraction technique [25] as shown in Fig.3b.

Redundancy among cameras is exploited by means of a Shape-from-Silhouette (SfS) technique [12]. This process generates a discrete occupancy representation of the 3D space (voxels). A voxel is labeled as foreground or background by checking the spatial consistency of its projection on the M segmented silhouettes. The data obtained with this 3D reconstruction is corrupted by spurious voxels introduced due to wrong segmentation, camera calibration inaccuracies, etc. A connectivity filter is introduced in order to remove these voxels and the final 3D binary reconstruction is shown in Fig.3c.

These data is then fed to the person tracking module that will produce a number of hypothesis about the centroid location of the multiple targets (people) in the scenario. The 3D position of their heads is directly inferred by selecting the upper voxels of the connected component associated to a given centroid position. These head locations are the main input required by the head orientation algorithms (both audio and video). Estimations from these two algorithms are combined to produce a more robust multimodal estimation. Focus of attention is addressed by combining the head orientation estimation of the multiple tracked people. Finally, this information might be fed to a higher semantic analysis module.

3. Multi-person Tracking

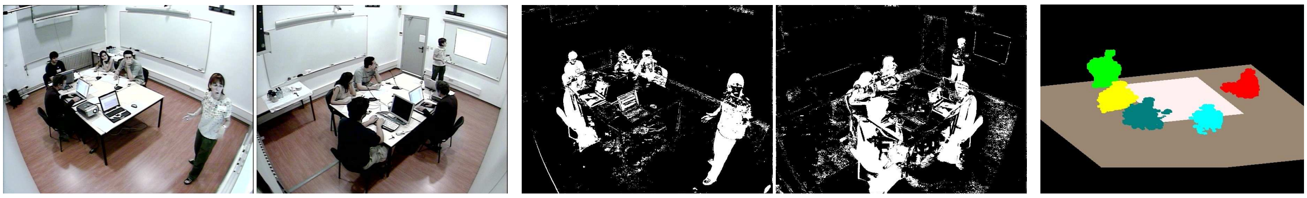
Detecting and tracking a group of people present in an indoor scenario provides relevant data towards activity recognition and understanding. Robust, multi-person tracking systems are employed in a wide range of applications, including SmartRoom environments, surveillance for security, health monitoring, as well as providing location and context features for human-computer interaction. As an example, a person standing next to a white board and several people located around a table provides enough evidence to guess that people attend a presentation or lecture. Furthermore the location and number of people in a room also is a useful feature for activity classification.

A number of methods for camera based multi-person 3D tracking has been proposed in the literature [17, 22]. A common goal in these systems is robustness under occlusions created by multiple objects present in the scene when estimating the position of a target. Single camera approaches [22] have been widely employed but are more vulnerable to occlusions, rotation and scale changes of the target. In order to avoid these drawbacks, multi-camera tracking techniques exploit spatial redundancy among different views and provide 3D information as well. Integration of features extracted from multiple cameras has been proposed in terms of multi-view histograms [17] or voxel reconstructions [18] among others.

Filtering techniques are employed to add temporal consistency to tracks. Kalman filtering approaches have been extensively used to track a single object under Gaussian uncertainty models and linear dynamics [20]. However, these methods do not perform accurately when facing noisy scenes or rapidly maneuvering targets. Particle filtering has been applied to cope with these situations since it can deal with multi-modal *pdf*s and is able to recover from lost tracks [5].

Nevertheless, particle filtering turns out to be computationally demanding and hence not suitable for real-time performing algorithms. This paper proposes a method that aims at decreasing computation time by means of a novel tracking technique based on the seminal particle filtering principle presented in [18]. Particles no longer sample the state space but instead a magnitude whose expectancy produces the centroid of the tracked person: the surface voxels. The likelihood evaluation relying on occupancy information is computed on local neighborhoods thus dramatically decreasing the computation load of the overall algorithm.

Multiple targets tracking is addressed by assigning a tracker to every one. In order to achieve the most independent set of trackers, a 3D blocking method modeling the interaction between targets is considered. This strategy defines an exclusion region where no particles from other trackers may fall, following the idea introduced by [18].



(a) (b) (c)

Figure 3. In (a), a sample of multiview original images. In (b), foreground segmentation of the input images employed by the SfS algorithm. In (c), example of the binary 3D voxel reconstruction used in this paper (false colors are employed to depict various people).

3.1. Particle Filtering Background

Particle Filtering (PF) is an approximation technique for estimation problems where the variables involved do not hold Gaussianity uncertainty models and linear dynamics. The current tracking scenario can be tackled by means of this algorithm to estimate the 3D position of a person $\mathbf{x}_t = (x, y, z)_t$ at time t, taking as observation a set of colored voxels representing the 3D scene up to time t denoted as $\mathbf{z}_{1:t}$. Multiple people might be tracked assigning a PF to each target and defining an interaction model to ensure track coherence.

For a given target $\mathbf{x}_t$, PF approximates the posterior density $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ with a sum of N_s Dirac functions:

$$p\left(\mathbf{x}_t|\mathbf{z}_{1:t}\right) \approx \sum_{j=1}^{N_s} w_t^j \delta(\mathbf{x}_t - \mathbf{x}_t^j), \qquad (1)$$

where w_t^j are the weights associated to the particles and $\mathbf{x}_t^j$ their positions. For this type of tracking problem, a Sampling Importance Re-sampling (SIR) PF is applied to drive particles across time [5]. Assuming importance density to be equal to the prior density, weight update is recursively computed as:

$$w_t^j \propto w_{t-1}^j \, p(\mathbf{z}_t|\mathbf{x}_t^j). \qquad (2)$$

SIR PF avoids the particle degeneracy problem by re-sampling at every time step. In this case, weights are set to $w_{t-1}^j = 1/N_s, \forall j$, therefore

$$w_t^j \propto p(\mathbf{z}_t|\mathbf{x}_t^j). \qquad (3)$$

Hence, the weights are proportional to the likelihood function that will be computed over the incoming volume $\mathbf{z}_t$. The re-sampling step derives the particles depending on the weights of the previous step, then all the new particles receive a starting weight equal to $1/N_s$ which will be updated by the next volume likelihood function.

Finally, the best state at time t of target m, $\mathbf{X}_t^m$, is derived based on the discrete approximation of Eq.1. The most common solution is the Monte Carlo approximation of the expectation as

$$\mathbf{X}_t^m = \mathbb{E}\left[\mathbf{x}_t|\mathbf{z}_{1:t}\right] \approx \frac{1}{N_s}\sum_{j=1}^{N_s} w_t^j \mathbf{x}_t^i. \qquad (4)$$

3.2. Sparse Sampling

PF approach to tracking defines a set of instances of the position of the tracked person, the particles, and a formulation to measure the fitness of these hypothesis with relation to the observable data. However, the evaluation of this likelihood function may be computationally expensive. An alternative to PF is devised by reviewing the estimation of the state $\mathbf{X}_t$ in Eq.4. Centroid of the person may be alternatively extracted by computing the expectation over all the surface voxel positions. By randomly selecting a given number of voxels on this surface, it is still possible to obtain an enough accurate estimation of $\mathbf{X}_t$. We define the *sparse sampling* (SS) algorithm as a method to recursively estimate $\mathbf{X}_t$ from an evolving set of samples placed on the surface of the tracked person. Since we are no longer exploring the state space, we will talk about *samples* instead of *particles*.

Essentially, the proposed algorithm follows the PF analysis loop (re-sampling, propagation, evaluation and estimation). Being our volume a discrete representation, samples are constrained to occupy a single voxel and move with displacements on the 3D discrete orthogonal grid. By defining the appropriate likelihood function, samples attain high weights when placed on the surface while the re-sampling block is constrained to place the newly created samples on the foreground voxels. With this process, we define a recursive way to obtain a sparsely sampled version of the surface of the target and, therefore, its centroid.

3.2.1 Likelihood evaluation

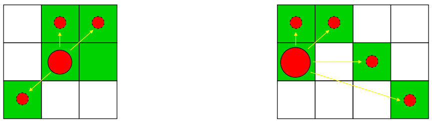
Function $p(\mathbf{z}_t|\mathbf{x}_t^j)$ can be defined as the likelihood of a sample belonging to the surface of a target. Let $\mathcal{C}(\mathbf{x}_t^j, q)$ be a neighborhood over a connectivity q domain on the 3D orthogonal grid around a sample placed in voxel $\mathbf{x}_t^j$. Then, we define the occupancy neighborhood around $\mathbf{x}_t^j$ as $\mathbf{O}_t^j = \mathbf{V}_t^b \cap \mathcal{C}(\mathbf{x}_t^j, q)$. For a given sample j occupying a voxel, its likelihood may be formulated as

$$p(\mathbf{z}_t|\,\mathbf{x}_t^j) = 1 - \left| \frac{2\|\mathbf{O}_t^j\|}{\|\mathcal{C}(\mathbf{x}_t^j, q)\|} - 1 \right|, \qquad (5)$$

where $\|\cdot\|$ is the number of occupied voxels of the enclosed volume. This expression measures the likelihood of a sample being placed in a surface voxel, attaining its maximum value when the half of its neighborhood is occupied. In our research $q = 26$ provided accurate results.

3.2.2 3D Discrete Re-sampling

The re-sampling step has been defined according to the condition that every sample is assigned to a foreground voxel. In other words, re-sampling has usually been defined as a process where some noise is added to the position of the re-sampled particles according to their weights [5]. The higher the weight, the more replicas will be created. In our current tracking scenario, re-sampling adds some *discrete* noise to samples only allowing motion within the 3D discrete positions of adjacent foreground voxels as depicted in Fig.4a. Then, non populated foreground voxels are assigned to re-sampled samples. In some cases, there are not enough adjacent foreground voxels to be assigned, then a connectivity search finds closer non-empty voxels to be assigned as shown in Fig.4b.



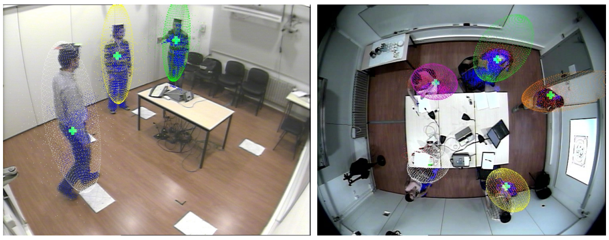
(a) (b)

Figure 4. Discrete re-sampling example (in 2D).

3.3. Tracking Performance

In order to evaluate the performance of the proposed algorithm, we collected a set of multi-view scenes in an indoor scenario involving up to 6 people, for a total of approximately 25 min. The analysis sequences were recorded with 5 fully calibrated and synchronized wide angle lense cameras in the SmartRoom at UPC with a resolution of 720x576 pixels at 25 fps (see a sample in Fig.3). The test environment is a 5m by 4m room with occluding elements such as tables and chairs. Groundtruth data was labeled manually allowing a quantitative measure of tracker's performance. It should be noted that the employed test database has been included in the CLEAR07 Evaluation [2].

Metrics proposed in [6] for multi-person tracking evaluation have been adopted. These metrics, being used in international evaluation contests [2] and adopted by several research projects such as the European CHIL [1] or the U.S. Vace [3] allow objective and fair comparisons. Two metrics employed are: the **M**ultiple **O**bject **T**racking **P**recision (*MOTP*), which shows tracker's ability to estimate precise object positions, and the **M**ultiple **O**bject **T**racking **A**ccuracy (*MOTA*), which expresses its performance at estimating the number of objects, and at keeping consistent trajectories. *MOTP* scores the average metric error when estimating multiple target 3D centroids, while *MOTA* evaluates the percentage of frames where targets have been missed, wrongly detected or mismatched.



(a) (b)

Figure 5. Tracking performance. In (a), a tracking scenario with 3 people and, in (b), with 5 people.

Two parameters drive the performance of the algorithm: the voxel size ν and the number of samples. In order to achieve a real-time performance, voxel size was set to $\nu = 4$ cm and 300 particles were used. The performance of this algorithm incresases proportionally with the number particles and with the inverse of the voxel size as shown in [10]. The obtained performance indicators were *MOTP*=110 mm and *MOTA*=78%. A visual example of the tracker's performance is shown in Fig.5.

4. Multimodal Head Orientation

4.1. Multi-camera Estimation

Methods for head pose estimation proposed in the literature [8] 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 [9]. In this context, appearance-based approaches [28] 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 [2] 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.

Let us assume that the 3D position of the head of the person of interest is available from the tracking module and determined by a bounding box $\mathcal{B}$, already available as an input to the head orientation algorithm. The center and size of the bounding box $\mathcal{B}$ allow defining an ellipsoid model of the head $\mathcal{H}$ as shown in Fig.6a.

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 [16]. 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.6a.

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

$$\Gamma(\mathrm{p}_n) \equiv P_n^{-1}(\mathrm{p}_n) = \mathbf{o}_n + \lambda\mathbf{v}, \qquad \lambda \in \mathbb{R}^+, \qquad (6)$$

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.6c. 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.

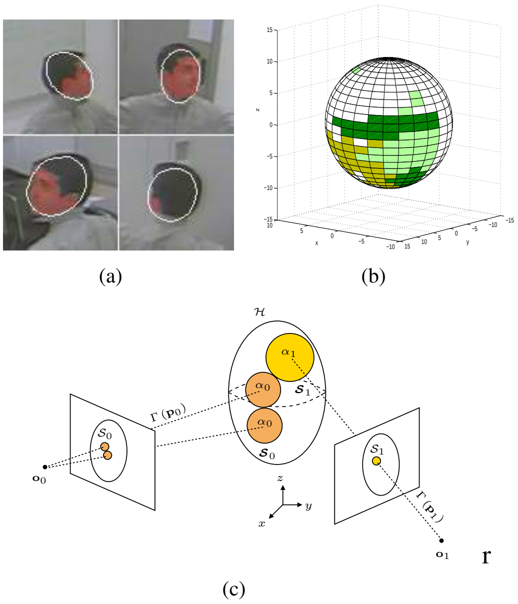
4.1.1 Orientation Estimation

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

$$\mathbf{d}_V = \frac{1}{\sum_{n=0}^{N-1} |\boldsymbol{\mathcal{S}}_n|} \sum_{n=0}^{N-1} \alpha_n \sum_{\mathbf{p}_n \in \boldsymbol{\mathcal{S}}_n} (\mathbf{p}_n - \mathbf{c}), \qquad (7)$$

$$\hat{\theta}_V = \tan^{-1}\left(\mathbf{d}_{V_y} / \mathbf{d}_{V_x}\right), \qquad (8)$$

where $|\boldsymbol{\mathcal{S}}_n|$ denotes the number of elements (3D intersections) in the set and $\mathbf{c}$ is the center of the head $\mathcal{H}$.



(a) (b)

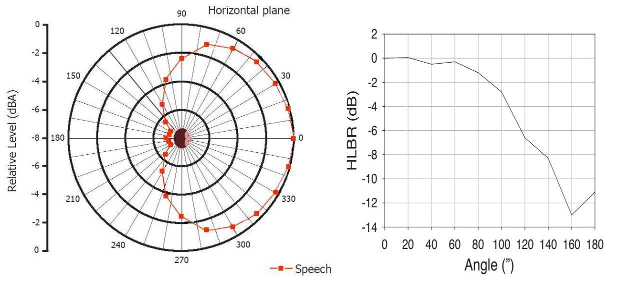
(c)

Figure 6. 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.

4.2. Multi-microphone Estimation

In this section we present a 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 [23]. All results described in this work were derived using only a set of four T-shaped 4-channel microphone clusters. It will assumed that the active speaker's location is known beforehand.

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 [19] 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 [14]. Fig.7a 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 uni-



(a) (b)

Figure 7. In (a), A-weighted head radiation diagram in the horizontal plane. In (b), HLBR of the head radiation pattern.

form, 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 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 [14]. 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.7b.

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.

It must be noted that this technique is intended for single person orientation estimation. When more than a person is speaking the audio modality can not provide an accurate estimation. Nevertheless, multimodality alleviates this problem by a relying more on the video modality.

4.2.1 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.

$$W(\theta) = \sum_{i=0}^{N_{MICS}} \delta(\theta - \theta_i) \cdot w_i, \tag{9}$$

$$\hat{\theta}_A = \underset{\theta}{\operatorname{argmax}}\, G(\theta) * W(\theta). \tag{10}$$

4.3. Multi-modal Estimation and Performance

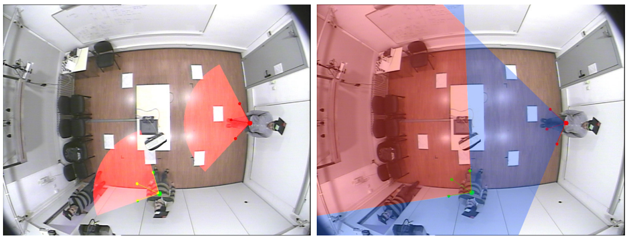
Video and audio head orientation estimations, $\hat{\theta}_V$ and $\hat{\theta}_A$, are combined by means of decentralized a Kalman filter [15]. Using both input streams allowed the system to overcome video occlusions among head participants, specially the speaker. This solution has been adopted instead of other more sophisticated ones [11] in order to keep an affordable computational load.

In order to evaluate the performance of the proposed algorithms, we employed the CLEAR head pose database [2] 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 [2] 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. 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.

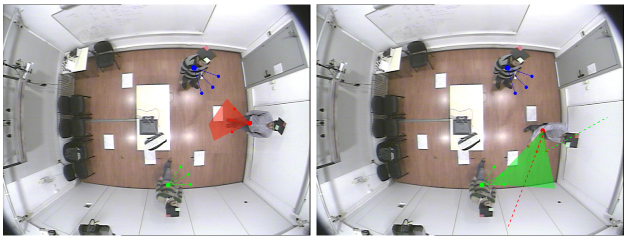
Method	PMAE (^{o})	PCC (%)	PCCR (%)
Video	57.23	32.88	71.39
Audio	53.14	28.47	69.17
Multimodal	48.53	38.19	73.47

Table 1. Quantitative results for the three presented systems.



(a) (b)

Figure 8. Focus of attention descriptors. In (a), the attention cones generated by two people. In (b), the attention map generated by the intersection of two cones: blue denotes the areas with one intersection and red the areas with two intersections.



(a) (b)

Figure 9. Focus of attention analysis. In (a), two attendees are listening to the lecturer and their attention cones meet at the speaker location. In (b), two attendees are detected to be interacting when their attention cones have a high overlay and their orientation vectors pass close to each other centroid.

5. Focus of Attention Estimation

Once the positions of the people inside the room and the orientation of their heads has been computed, focus of attention (FoA) information may be estimated. This problem has been typically addressed in the literature using neural networks [27] or HHMs [26]. However, in order to keep an affordable complexity of the system, two geometric descriptors of the FoA are introduced and, from its properties, a situation analysis can be derived.

The spatial region where the attention of a person is drawn is tightly correlated with the orientation of his head and the horizontal and vertical span of his eyes perception [4]. In this paper only horizontal FoA is estimated since most of the relevant activity is concentrated in this plane and the visual span angle is empirically set to $\delta = 30^o$. The *attention cone* $\mathcal{K}$ can be defined as the cone with an opening angle δ, the apex centered at the head centroid $\mathbf{x}$ and the orientation in the z-plane set to the head orientation estimation θ. A depiction of the attention cones of several people can be found in Fig.9a. In order to analyze the global focus of attention of a group of people, we define the *attention map* as the cumulative intersection of all the attention cones in the z-plane. An example can be found in Fig.9b. Although these two FoA descriptors are less sophisticated than statistical or HMM based approaches, still capture the underlying information of the group attention and allow detecting and recognizing some basic events.

Two particular cases can be detected out of these two FoA descriptors:

- **Region of Interest detection:** Intersection of a the gaze cones of a number of attendees at a certain region of the space may denote that there is something relevant there. For instance, when a presentation is performed, the gaze cones of most of the participants meet at the beamer projection area or at the presenter. Given an attention map $\mathcal{M}$, the *regions of interest* may be defined as those fulfilling $\mathcal{M} > \alpha$, being α the minimum number of cones intersecting at a given point of the room. An example is depicted in Fig.9a with $\alpha = 3$ in the lecture scenario and another example is shown in Fig.9b with $\alpha = 2$ where two people are interested in the area near the door when somebody is entering.
- **Interaction detection:** Interaction between two people can be detected when their corresponding attention cones have a high overlap and the orientation vectors of each person pass close to the cone origin of the other as shown in Fig.9b. A cost matrix is computed at every frame between all persons in the room and those pairs fulfilling this criterium are labelled as interacting.

5.1. Performance

In order to assess the performance of the proposed focus of attention analysis algorithm, a short 5 minutes database was collected involving up to 5 people in a SmartRoom scenario. The room setup consisted in 5 calibrated cameras with a resolution of 720x576 pixels at 25 fps and 4 T-Shaped microphones sampling at a 44KHz. Groundtruth information regarding the regions of interest and the interaction between individuals in the room was hand-labeled. Performance of the overall system obtained an 85% of correctly detected events and a 15% of false positive detections.

Computational load of the overall system is proportional to the number of targets to be analyzed in the scene. Performance of the whole system attained an average speed of 15 fps when only one person is present in the scene and 6 fps when there are 5 people. A distributed processing system has been employed consisting in 5 off-the-shelf machines with a 2.2GhZ processor.

6. Conclusions and Future Work

This paper presented a multi-person focus of attention tracking system that combine two technologies, namely person tracking and head orientation estimation. This system was intended for real-time operation hence some considerations have been made towards reducing its complexity like the sparse sampling technique introduced in the multi-person tracking module. Focus of attention is addressed by defining the attention cone and the attention map that allows detecting regions of interest and recognizing interactions and behaviors among attendants.

Future research lines within the scope of this paper include defining more sophisticated automatic human behavior analysis techniques based on focus of attention estimation. Further validation of the proposed system over larger databases with focus of attention annotations is under study. Combination of this system with outputs coming from other signal processing modules such as event detection or speech activity detection are under study.

References

[1] CHIL-Computers In the Human Interaction Loop. http://chil.server.de.

[2] CLEAR Evaluation. http://www.clear-evaluation.org.

[3] VACE-Video Analysis and Context Extraction. http://www.ic-arda.org.

[4] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[5] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *Signal Processing, IEEE Tran. on*, 50(2):174–188, 2002.

[6] K. Bernardin, A. Elbs, and R. Stiefelhagen. Multiple object tracking performance metrics and evaluation in a smart room environment. In *Proc. IEEE Int. Workshop on Vision Algorithms*, pages 53–68, 2006.

[7] M. Black, F. Brard, A. Jepson, W. Newman, W. Saund, G. Socher, and M. Taylor. The Digital Office: Overview. In *Proc. Symposium on Intelligent Environments*, pages 98–102, 1998.

[8] X. Brolly, C. Stratelos, and J. Mulligan. Model-based head pose estimation for air-traffic controllers. In *Proc. IEEE Int. Conf. on Image Processing*, pages 113–116, 2003.

[9] C. Canton-Ferrer, J. Casas, and M. Pardas. Fusion of multiple viewpoint information towards 3D face robust orientation detection. In *Proc. IEEE Int. Conf. on Image Processing*, volume 2, pages 366–369, 2005.

[10] C. Canton-Ferrer, J. Salvador, and J. Casas. Multi-person tracking strategies based on voxel analysis. In *Proc. Classification of Events, Activities and Relationships Evaluation and Workshop (CLEAR)*, Lecture Notes in Computer Science, 2008.

[11] C. Canton-Ferrer, C. Segura, J. Casas, M. Pardas, and J. Hernando. Audiovisual head orientation estimation with particle filters in multisensor scenarions. *EURASIP Journal on Advances in Signal Processing*, 2007.

[12] G. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3D voxel reconstruction of human motions. In *Proc IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 714–720, 2000.

[13] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox. Room with a rear view: Meeting capture in a multimedia conference room. *IEEE Multimedia Magazine*, 7(4):48–54, 2000.

[14] W. T. Chu and A. Warnock. Detailed directivity of sound fields around human talkers. Technical report, Institute for Research in Construction, 2002.

[15] H. R. Hashemipour, S. Roy, and J. Laub. Decentralized structures for parallel Kalman filtering. *Automatic Control, IEEE Tran. on*, 33(1):88–93, 1988.

[16] M. Jones and J. Rehg. Statistical color models with application to skin detection. *Int. Journal of Computer Vision*, 46(1):81–96, 2002.

[17] O. Lanz. Approximate Bayesian multibody tracking. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 28(9):1436–1449, 2006.

[18] A. Lopez, C. Canton-Ferrer, and J. Casas. Multi-person 3D tracking with particle filters on voxels. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2007.

[19] P. Meuse and H. Silverman. Characterization of talker radiation pattern using a microphone array. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 257–260, 1994.

[20] I. Mikic, S. Santini, and R. Jain. Tracking objects in 3d using multiple camera views. In *Proc. Asian Conf. on Computer Vision*, pages 234–239, 2000.

[21] K. Otsuka, Y. Takemae, J. Yamamoto, and H. Murase. A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions and utterances. In *Proc. Int. Conf. on Multimodal Interfaces*, 2005.

[22] A. Pnevmatikakis and L. Polymenakos. 2D person tracking using Kalman filtering and adaptive background learning in a feedback loop. In *Proc. Classification of Events, Activities and Relationships Evaluation and Workshop (CLEAR)*, volume 4122 of *Lecture Notes on Computer Science*, pages 151–160, 2007.

[23] J. Sachar and H. Silverman. A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 4, pages 65–68, 2004.

[24] C. Segura, C. Canton-Ferrer, J. Casas, and J. Hernando. Multimodal head orientation towards attention tracking in smartrooms. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007.

[25] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 252–259, 1999.

[26] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling people's focus of attention. In *Proc. IEEE Int. Workshop on Modeling People*, 1999.

[27] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. on Neural Networks*, 13(4), 2002.

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