A JOINT PARTICLE FILTER TO TRACK THE POSITION AND HEAD ORIENTATION OF PEOPLE USING AUDIO VISUAL CUES

Alessio Brutti and Oswald Lanz

Center of Information Technology - Irst, Fondazione Bruno Kessler
Via Sommarive 18, 38123 Povo di Trento (TN), Italy
{brutti,lanz}@fbk.eu

ABSTRACT

Automatic analysis of interactive people behavior is an emerging field where significant research efforts of the audio and image processing communities converge. In this paper we present a particle filter for jointly tracking the position of multiple people, their head orientation and speaking activity based on audio visual cues. These are integrated with a novel fusion technique that takes into account the spatial distribution of the sensing infrastructure. The resulting system provides real time information about peoples' behavior and activities that can be used to boost the awareness of technology assisted working and living environments.

1. INTRODUCTION

There is an increasing demand from the society to realize electronic systems that assist people in their working and living environment. Applications in the field of Domotics, Ambient Assisted Living, Surveillance and Human-Computer Interaction require such solutions to be sensitive and responsive to the presence of people. Therefore, there is a need to develop perceptual technologies able to provide detailed reports on their behavior and activities. Audio visual analysis hereby offers a convenient framework.

This paper focuses on the joint determination of head position and horizontal orientation, and speech activity of people interacting in indoor environments monitored with multiple cameras and microphones. A vast amount of literature is available on the problem of people tracking and of head location and pose estimation. The CLEAR workshops [1] addressed these tasks and provided a quantitative comparison of several techniques. Many approaches are based on a two step strategy: head detection followed by pose estimation, often using neural networks classifiers (e.g. [2, 3, 4, 5]). A somehow different approach is followed in [6] where pose estimation is computed by Bayesian integration of the responses of multiple face detectors tuned to different views. An alternative, potentially more robust approach proposed in [7, 8, 9, 10, 11] is to estimate jointly location and orientation using a mixed-state particle filter.

The system proposed in this paper follows the joint estimation approach and is based on a particle filter for the integration of multiple sensor information and temporal dynamics. We build upon our previous work on visual and acoustic tracking [12, 13, 14, 15] to come up with an integrated approach that conveniently marries the advantages of each modality.

2. SEQUENTIAL BAYESIAN FRAMEWORK

To analyze the behaviour patterns of interacting people we adopt a Bayesian approach. Such approach turns out to be particularly convenient in a multi-sensor multi-modal setting. It allows to easily link relevant information from different sources by (i) defining a common reference frame representing the features of interest, (ii) modeling the dynamics on the chosen representation, and (iii) implementing, for each modality, a generative model of the measurement process.

For the task addressed in this paper, i.e. estimating the location of people, their focus of attention and speech activity, the representation $\mathbf{x}$ is chosen to be a five dimensional vector composed of the two dimensional position x^p of each person measured on a horizontal reference plane, the horizontal orientations of the head x^h and torso x^t, and a binary variable x^s indicating speech activity or silence. For the temporal evolution we assume that speaker turns can happen suddenly, and that a person can move randomly within the environment and can orient its head in any direction. Thus our motion model relies on independence assumptions. Both movements, though, can be executed only with limited velocity. The resulting dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ at time t does then not depend on x^s and can be expressed as a product of three Gaussians $p(\mathbf{x}_t|\mathbf{x}_{t-1}) = G(x_t^p - x_{t-1}^p|\sigma^p) \cdot G(x_t^h - x_{t-1}^h|\sigma^h) \cdot G(x_t^t - x_{t-1}^t|\sigma^t)$. The observation likelihood $q(\mathbf{z}|\mathbf{x})$ for the multi-source signal $\mathbf{z}$ is the key components of the Bayesian model and is designed in a generative fashion: we first render the hypothesis $\mathbf{x}$ into the sensor domain using a model of the target and the measurement physics and match it then with the real observations. Specific models for the acoustic and visual domain are detailed in the following sections.

The aim of tracking is then to recursively estimate the posterior distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ of the representation conditioned on a sequence of sensory observations $\mathbf{z}_{1:t}$. At each iteration this is done in two steps, by first propagating the posterior obtained at the previous time $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$ according to the dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and then updating it with the information contained in the new observation using the likelihood model $q(\mathbf{z}_t|\mathbf{x}_t)$ of the observation process

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto q(\mathbf{z}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})\, d\mathbf{x}_{t-1}.$$

When the observation likelihood is complex, like in our case, the posterior cannot be expressed in closed form and one has to resort to approximations. The particle filter maintains a compressed representation of the posterior by means of a set of representative sample states, the particles. At each iteration t, a new set of representative particles, $\{x_i\}$, is i.i.d.-sampled from the motion prior mixture evaluated over pre-

Research partly funded by the Autonomous Province of Trento, Italy, under project JRP PUMALAB

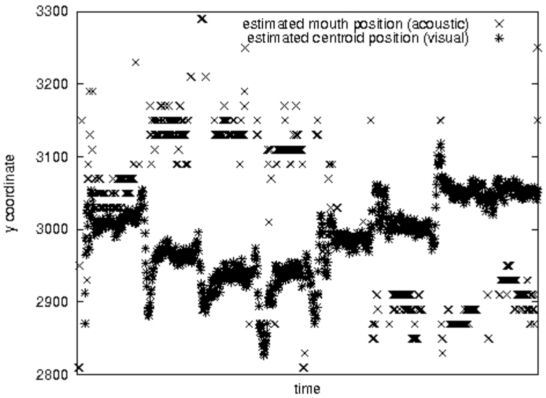


Figure 1: The plot shows the estimated horizontal y-displacement (in mm) of the speech source while a person turns around at a fixed position (at $y \approx 3m$) and speaks towards the four walls and four corners of the room. There is an evident offset of the speech source from the body center of about 20 cm. This plot shows also that the proposed acoustic and visual likelihoods provide estimates that are sufficiently accurate to exploit this offset for jointly estimating the orientation of an acoustic source.

vious particles. Then, the particle likelihoods are computed on the new observation $\mathbf{z}_t$ and used as importance weights $\pi_i = q(\mathbf{z}_t|x_i)$. To focus on likely trajectories, particles are periodically resampled according to their weights.

3. AUDIO VISUAL LIKELIHOOD

Following a generative approach, in this section we describe likelihood functions to estimate the position and head orientation of a person by audio visual means. While audio visual localization and tracking has been subject of extensive research in the past, not much attention has been paid to the joint estimation of the speaker orientation. Here we focus on this latter, exploiting that (i) the acoustic signal is directional (we use this fact in Sec. 4), (ii) the head orientation is directly observable in an image by its color and body shape pattern, and (iii) the location of the speech source has a horizontal offset from the body along the head orientation. Assuming that the individual modalities support sufficiently accurate localization of the speech source and the body center they can be explicitely conditioned to each other by this offset, suggesting that a joint approach may be most convenient. Fig. 1 shows that with the acoustic and visual likelihoods proposed in this section this assumption is met in our sensor setup.

3.1 Acoustic likelihood

Given a sound source in spatial position $\mathbf{p}$ and two microphones with 3D coordinates $\mathbf{s}_1$ and $\mathbf{s}_2$, the direct wavefronts reach the two sensors with a certain time delay which is referred to as Time Difference of Arrival (TDOA) by $\tau(\mathbf{p}) = (\|\mathbf{s}_1 - \mathbf{p}\| - \|\mathbf{s}_2 - \mathbf{p}\|)/c$, where c is the speed of sound and $\|\cdot\|$ is the Euclidean norm. Since this equation maps a hypothesis $\mathbf{p}$ into its corresponding time delay $\tau(\mathbf{p})$ it can be interpreted as a rendering function into the TDOA measurement domain. Knapp and Carter [16] introduced the generalized cross correlation phase transform (GCC-PHAT), which is the most popular method for TDOA estimation. Denoting by z_1 and z_2 the digitized sequences captured by the two microphones over a small time window, GCC-PHAT is formulated as follows:

$$\mathrm{GCC}(\tau) = FFT^{-1}\left\{ \frac{FFT(z_1)\cdot FFT^{*}(z_2))}{|FFT(z_1)|\cdot|FFT(z_2)|} \right\}$$

where τ is the time lag in samples between the signals. For each time lag, GCC-PHAT evaluates the similarity between the two signals and, in ideal conditions, it presents a dominat peak for τ equal to the actual TDOA. Alternative approaches to TDOA estimation adopt multiple microphone set up [22] or information theory [21].

It is a known fact that particle filters do not behave well with sharply peaked and irregular likelihoods [17]. Therefore we compute the acoustic likelihood using a smoothed version of GCC-PHAT computed as follows. For a given hypothesis $\mathbf{x}$ an audio source is hypothesized at $\mathbf{p} = (x, y, z)$ where (x, y) are the particle coordinates shifted by an offset of 20 cm along the direction of the particle orientation, and z is fixed to 90% of target height (which is assumed to be known apriori). We then compute for each microphone pair the interval of TDOAs which map inside a sphere centered at this point (in the experiments of Sec. 5 the radius of the sphere is set to 50 cm). The highest GCC-PHAT response in this interval is found and weighted with the relative distance to the source location. Taking its exponential we get the likelihood computed on a microphone pair, which is now a smooth function of $\mathbf{x}$. This likelihood is used both for tracking and speech activity detection. If the acoustic likelihoods on a target's particle set exceed a given threshold (empirically set to $\exp(2) = 7.4$ in our implementation) the event is marked as active speech for that person, and the acoustic likelihoods are used in the calculation of the particle weights. Otherwise, the audio scores are neglected and only the visual likelihoods are considered in the filter update. The presence of multiple targets is dealt with by removing the GCC-PHAT measurements associated to the active speaker [15] so that they do not compromise the speech activity detection associated to other targets.

3.2 Visual likelihood

Following again a generative approach, we define a visual likelihood that builds upon a rendering function $g(\mathbf{x})$ to map a hypothesis $\mathbf{x}$ into a set of visual features (color histograms in our case). The rendered features are then scored against features extracted from the actual observation using an appropriate distance function d (Bhattacharyya distance).

The rendering model has two components: a coarse 3D shape model assembled from five cone truncs representing a standing person, and a body part- and viewpoint-based representation of the targets color pattern, in form of head, torso and legs histograms. To obtain its image projection for a state $\mathbf{x}$ we fit a silhouette template around the segment joining the image projection of the two 3D points representing the targets center of feet and top of head under $\mathbf{x}$. To account for the change in the targets profile width we rescale the template width according to the relative orientation θ_t of the torso to the camera by $0.7 + 0.3 \cdot |\cos\theta_t|$. In addition, the head patch is shifted horizontally by a quantity equal to $0.2 \cdot |\sin\theta_h| \cdot w_h$ where θ_h is the relative orientation of the head to the camera under $\mathbf{x}$ and w_h is the width of the head patch. This makes the shape projection model sensitive to head orientation. The projected shape is decomposed into three body parts: head, torso and legs. Within each of these parts, the appearance of the target is described by a RGB color histogram ($8 \times 8 \times 8$ bins). To obtain the histograms under $\mathbf{x}$ we follow a view-based rendering approach. Given a set of pre-acquired key views of the target, we extract the color histograms for each body part, and generate the histograms for a new view by interpolation. To compute the interpolation weights for a given particle, the orientation of the body part θ with respect to the camera is taken into account. The set of neighboring model views $\mathscr{V}$ is identified and the histogram is computed by $\sum_{v \in \mathscr{V}} w_v(\theta) \cdot h_v$, where h_v indicates the histogram of key view v. The interpolation weights account linearly for the angular offset between the two orientations $w_v(\theta) = 2\cos^{-1}\langle \vec{\theta}, \vec{\theta}_v \rangle / \pi$ where $\vec{\theta}, \vec{\theta}_v$ denote the 3D versors oriented according to θ and key view v. The visual likelihood for $\mathbf{x}$ is then computed by matching the color histograms extracted from the head, torso and leg patches identified by the shape projection with the interpolated model histograms using Bhattacharyya-coefficient based distance.

4. LIKELIHOOD INTEGRATION IN UNEVENLY DISTRIBUTED SENSOR NETWORKS

In this section we propose a method to consistently integrate multi-sensor observations in a Bayesian framework that accounts for the spatial distribution of the sensors in the monitored environment. We claim that this is an important aspect in asymmetric deployments when the signal source is directional and/or the sensitivity of the signals to variations in the state variables (i.e. likelihood sharpness) varies significantly with sensor position and/or modality. We give an example to support this claim at the end of this section.

To derive the method from a theoretically grounded model we consider the ideal case in which we have a spatially dense population of sensors, each one providing a conditionally independent measurement $z(\mathbf{s})$ indexed by its position $\mathbf{s}$. To account for directional sources, we introduce a model $e(\mathbf{s}|\mathbf{x})$ of the emission pattern of the target under state $\mathbf{x}$, whose purpose is to suppress the contributions of those measurements that are out of the influence range of the emitted signal under $\mathbf{x}$. The likelihood of a dense measurement $\mathbf{z} = \{z(\mathbf{s})\}_{\mathbf{s}}$ can then be evaluated by

$$-\log q(\mathbf{z}|\mathbf{x}) \propto \int e(\mathbf{s}|\mathbf{x}) l(z(\mathbf{s})|\mathbf{x}) d\mathbf{s} \qquad (1)$$

where l denotes the local log-likelihood evaluated on the sensor at $\mathbf{s}$.

In practice, however, we do have only a finite number of sensors deployed at $\{\mathbf{s}_i\}_i$ in the environment to compute an approximation of the above integral. To do so in a Monte Carlo fashion we interpret their measurements $\{z(\mathbf{s}_i)\}$ as i.i.d. samples of a importance density $s(\mathbf{s}|\mathbf{x})$, such that

$$\int e(\mathbf{s}|\mathbf{x}) l(z(\mathbf{s})|\mathbf{x}) d\mathbf{s} \approx \sum e(\mathbf{s}_i|\mathbf{x}) l(z(\mathbf{s}_i)|\mathbf{x}) / s(\mathbf{s}_i|\mathbf{x}). \qquad (2)$$

To apply this method the values of s at the sensor positions $\mathbf{s}_i$ have to be estimated. This can be done by a density estimation technique using a kernel K that accounts for the expected correlation of the measurements under $\mathbf{x}$.

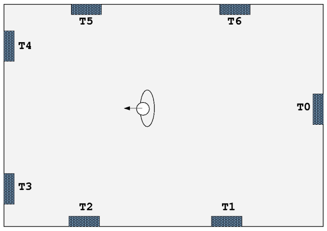


Figure 2: Microphone arrays in the experiment room

Although the proposed technique may most conveniently be adopted in a multi-modal setting (where K has to be designed to jointly consider all modalities, which may not be straightforward) in this paper we focus on its application to the acoustic component of the likelihood (i.e. we set $e/s \equiv 1$ for the visual contributions in Eq. 2). This choice is also motivated by the fact that acoustic estimates are generally more affected by the sensing and room geometry than their visual counterparts, and, as a consequence, it allows us to highlight the advantages deriving from the method. Fig. 3(a) shows a plot of the density s, the emission pattern e and the weighting factor e/s we use to approximate the joint acoustic likelihood in the experiments in Sec. 5. The kernel accounts for the radial distribution of the microphone pairs $\mathbf{s}_i$ around the source hypothesized in $\mathbf{x}$ (here $\langle \cdot, \cdot \rangle$ is the internal product, $[\cdot]^y$ is the y coordinate, $\mathscr{N}$ is the density of the normal distribution)

$$K(\mathbf{s}, \mathbf{s}_i|\mathbf{x}) = \mathscr{N}(\cos^{-1} \frac{\langle \mathbf{s}-\mathbf{x}, \mathbf{s}_i - \mathbf{x} \rangle}{\|\mathbf{s}-\mathbf{x}\| \|\mathbf{s}_i - \mathbf{x}\|};\ 0,\ \sigma_k) \qquad (3)$$

and

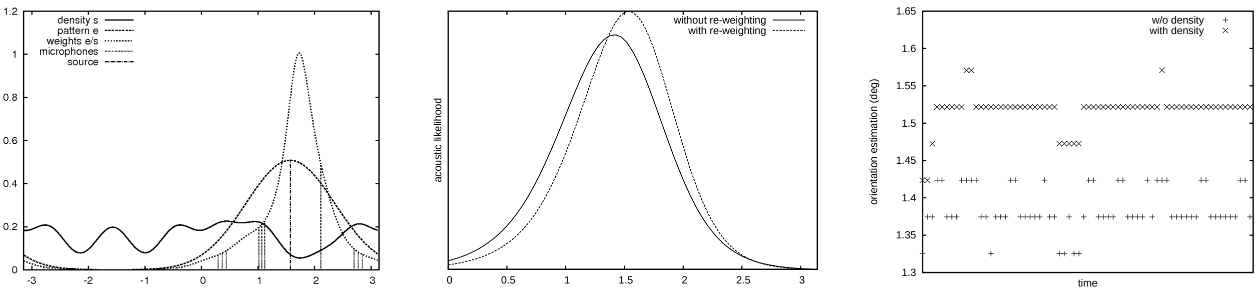
$$e(\mathbf{s}|\mathbf{x}) = \mathscr{N}(\tan^{-1} \frac{[\mathbf{s}-\mathbf{x}]^y}{[\mathbf{s}-\mathbf{x}]^x};\ x^h,\ \sigma_e) \qquad (4)$$

models the acoustic radiation pattern of the directed speech of a person. Note in the figure that the microphones are unevenly distributed around the position of the directed sound source: three close microphone pairs are available at the left (T4 in Fig. 2) but only one at the right (T3). In addition, the reverberation of the room injects more *noise* in the signals recieved by the microphone pairs at the left than in the single microphone pair at the right. By neglecting the density s the joint likelihood will be heavily biased towards the (noisy) estimates obtained from the three microphone pairs at the right. By considering the distribution of the microphones around the source location by means of s the contribution of the three pairs are properly down-weighted and the bias is attenuated (see Fig 3(b)).

5. EXPERIMENTAL RESULTS

To validate the approach, three sequences of graded difficulty have been acquired in our lab. The room dimensions are $6.0 \times 4.8 \times 5.0$m. The acoustic sensor set up consists of a distributed microphone network which includes 7 microphone arrays distributed in the monitored environment (see Fig. 2). Since each array consists of 3 microphones spaced at 20 cm, the overall number of available microphone pairs for GCC-PHAT computation is 21. The reverberation time of the room is 0.7 s and the sampling rate is 44.1 kHz. Four firewire cameras with a field of view of about 90 deg are installed in the corners of the room. They deliver RGB images of size 512×384 at a rate equal to 15 Hz.

In the first sequence (90 s) a person located at the center of the room turns around in steps of about 45 deg and



(a) (b) (c)

Figure 3: Source orientation estimation with unevenly distributed microphone arrays. (a): importance density s, emission pattern e, and weigthing function e/s for the directional acoustic source in Fig 2. (b): joint acoustic likelihood in Eq. 2 computed using radial kernel (Eq. 3), and without re-weighting (uniform s). (c) shows the orientation estimation based on acoustic likelihood. Notice how the estimation is biased towards the three microphone pairs at the left if their contributions are not re-weighted according to their radial distribution around the source.

speaks towards the four walls and four corners of the room. The ground truth (position and head orientation) for this sequence has been generated manually. The second sequence (95 s) involves two people moving in the room and speaking to each other in turn. Both people continue to look to each other while moving: this way we were able to generate the references for evaluation in an automatic way by tracking the spatial position of both targets and using the ray connecting the two targets to compute a reference for their head orientations over time. The walking trajectories hereby were computed offline with the color based particle filter described in this paper using high quality color models acquired offline and a high number of particles. The quality of the references obtained this way are comparable to manual labelling. The third sequence (131 s) is the most challenging one, with two people moving quickly while speaking in turn. After about 60 s one person sits down and continues to speak to the other which moves around. After that, a loudspeaker is turned on and emits the speech of a person at a comparable volume for about 40 s in the central area of the room (a coherent noise source in view of our evaluation). The references for evaluation have been generated in the same manner as for the second sequence. However, since the targets moved much faster than in the second sequence the references in this case may be less reliable, also because visually following a moving target does not necessarily mean that a persons head orientation is tightly aligned with the gaze direction: if the direction changes quickly, as happening in this sequence, one tends to delay the movement of the head while gazing correctly.

The presented multi-modal particle filter has been implemented as an extension of the *SmarTrack* system [18, 19] (`http://tev.fbk.eu/smartrack`) and runs comfortably in real time on a modern workstation on the evaluation sequences. It automatically detects people as they enter the monitored room, acquires their visual signature, and tracks their position, head orientation and speech activity using the audio visual likelihoods and fusion technique proposed in this paper. See `http://pumalab.fbk.eu/amm` for videos of some of the evaluation runs and to assess the quality of the references. We also intend to make the sequences publicly available for evaluation purpose.

The proposed algorithm is evaluated using the MOT scoring tool adopted in the CLEAR 2007 evaluation campaign [20]. The tool was extended to evaluate head orientation as well by means of the average absolute angular error. In order to better appreciate the impact of using acoustic and visual likelihoods jointly, we report on the performance of the system during speech segments only. Although the behaviour of the filter is conditioned also by what happens during silence, it can quickly adapt when new acoustic observations are available, making this kind of analysis reasonable. When two targets are present, only the speaking one is evaluated. For comparison, two mono-modal trackers operating on either visual [18, 19] and acoustic [15] observations are also evaluated on the acquired sessions. The acoustic tracker is designed for a single source and is therefore not evaluated on session 3 where an interfering speech source is present.

	modality	run 1	run 2	run 3	run 4	run 5
Session 1	video	68 0.167	62 1.38	68 0.185	68 0.175	62 1.35
	audio	105 0.421	104 0.421	108 0.414	105 0.417	105 0.420
	audio-video	77 0.182	77 0.194	75 0.164	77 0.183	82 0.253
Session 2	video	62 0.922	81 0.867	62 0.908	94 0.440	63 0.802
	audio	141 0.485	141 0.489	143 0.474	142 0.491	140 0.479
	audio-video	64 0.391	64 0.414	54 0.346	65 0.374	63 0.404
Session 3	video	65 1.60	49 1.56	36 0.278	46 2.63	81 1.17
	audio-video	54 1.20	40 0.506	55 1.19	38 0.225	35 1.243

Table 1: Experimental results for each session and each modality. The first row reports the localization precision (mm) while the second row shows the precision in the head orientation estimation (rad). The MOTA index is not reported in the table because it is always above 95%.

Table 1 reports on the performance for each session and for each modality. The table shows the results obtained on 5 independent runs on each modality. Since the visual appearance model of the targets are acquired on-the-fly (color histograms, but also target height is estimated) their quality may change significantly from run to run. It is clear that the quality of the model has a major impact on the filters ability to visually track the head orientation; this can be observed in the table, where the performance of the visual tracker exhibits a high variance in the head orientation error. This is most evident in session 1 (run 2 and 5). Conversely, adding the acoustic likelihood, which is not target dependent, ensures more uniform performance among different runs. It is worth noting that the acoustic observations do not significantly improve the estimation performance when good visual models are available. Concerning localization performance, the visual likelihood alone can already provide accurate estimations of the target position.

The results achieved in session 2 seem to confirm the trend observed in session 1. It is worth underlining that references here are obtained automatically and therefore only qualitative assessments can be derived from those numbers. Finally, session 3 shows that the system is not robust to the presence of interfering sound sources which are not dealt with in the current implementation. As a matter of fact, the sound irradiated by the interferer conditions all GCC-PHAT measurements, weakening the acoustic likelihood. A possible solution may be to handle the interferer as a non human target, thus going for a multi-target multi-speaker framework. In this session the performance on head orientation estimation is further deteriorated by the fact that the quality of the automatically extracted target model associated to the second person is low in 4 cases out of 5.

6. CONCLUSION

We have presented an integrated approach to audio visual tracking of interacting people to determine their spatial positions, head orientations, and speaking activities. It is based on a particle filter that operates in real time, thus providing a suitable online tool for higher level analysis about the behaviour of interacting people. Despite the enriched reports provided by a multi-modal analysis, the integrated use of both acoustic and visual cues have shown to provide also more robust and accurate results than their single modality counterparts. Our plans are to further investigate on the correlations that exist across space, time and modality building on the integration technique presented in this paper.

REFERENCES

[1] Proceedings of CLEAR'06 and CLEAR'07 Workshops: Classification of Events, Activities and Relationships, Springer LNCS Series, 2006, 2007

[2] Stiefelhagen, R., Yang, J., Waibel, A.: Modeling Focus of Attention for Meeting Indexing based on Multiple Cues, *IEEE Transactions on Neural Networks*, July 2002, Vol. 13, Number 4, pp. 928-938

[3] Gourier, N., Maisonnasse, J, Hall, D., Crowley, J.L.: Head Pose Estimation on Low Resolution Images, *Proc. of CLEAR'06 Workshop*, Springer LNCS Series, 2006

[4] Voit, M., Nickel, K., Stiefelhagen, R.: Multi-View Head Pose Estimation using Neural Networks, *in Proc. of the 2nd Canadian Conference on Computer and Robot Vision CRV'05*, pp. 347-352, 2005

[5] Canton-Ferrer, C., Casas, J.R., Pardas, M.: Fusion of multiple viewpoint information towards 3d face robust orientation detection, *in IEEE International Conference on Image Processing*, Vol. 2, pp 366- 369, 2005

[6] Zhang, Z., Hu, Y., Liu, M., Huang, T.: Head Pose Estimation in Seminar Room using Multi View Face Detectors, *Proc. of CLEAR'06 Workshop, Southampton, UK*, Springer LNCS Series, 2006

[7] Ba, S.O., Odobez, J.M.: A probabilistic framework for joint head tracking and pose estimation, *in Proc. of ICPR 2004*, Vol. 4, pp 264-267

[8] Ba, S. O., Odobez, J.M.: A Rao-Blackwellized mixed state particle filter for head pose tracking, *in ACM-ICMI Workshop on Multi-modal Multi-party Meeting Processing (MMMP)*, Trento, Italy, 2005, pp 9-16

[9] Canton-Ferrer, C., Segura, C., Casas, J.R., Pards, M., Hernando, J.: Audiovisual head orientation estimation with particle filtering in multisensor scenarios, *EURASIP Journal on Advances in Signal Processing*, No. 32, 2008

[10] Nickel, K., Gehrig, T., Stiefelhagen, R., McDonough, J.: A joint particle filter for audio-visual speaker tracking. *Int. Conf. Multimodal Interfaces*, 2005

[11] Fallon, M., Godsill, S., Blake, A.: Joint Acoustic Source localization and orientation estimation using sequential MonteCarlo, *Conference on Digital Audio Effects*, Montral, Canada, 2006

[12] Brunelli, R., Brutti, A., Chippendale, P., Lanz, O., Omologo, M., Svaizer, P., Tobia, T.: A Generative Approach to Audio-Visual Person Tracking. *Multimodal Technologies for Perception of Humans*, Springer LNCS 4122, pp. 55-68

[13] Lanz, O., Brunelli, R.: Dynamic Head Location and Pose from Video, *IEEE Conference on Multisensor Fusion and Integration*, September 3-6, 2006

[14] Brutti, A., Omologo, M., Svaizer, P.: Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. *Proc. of Interspeech*, 2005

[15] Brutti, A., Omologo, M., Svaizer, P.: A Sequential Monte Carlo approach for tracking of overlapping acoustic sources. *Proc. of EUSIPCO*, 2009

[16] Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976

[17] Sullivan, J., Rittscher, J.: Guiding Random Particles by Deterministic Search, *Int. Conf. Computer Vision*, 2001

[18] Lanz, O.: Approximate Bayesian Multibody Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006

[19] Lanz, O., Messelodi, S.: A sampling algorithm for occlusion robust multi target detection. *Int. Conf. Advanced Video and Signal based Surveillance AVSS*, 2009

[20] Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the CLEAR MOT metrics. J. Image Video Process., 2008(3), 2008.

[21] Talantzis, F., Constantinides, A.G., Polymenakos, L.C.: Estimation of direction of arrival using information theory. *IEEE Signal Processing Letters, Vol. 12(8)*, 2008

[22] Chen, J., Benesty, J., Huang, Y.: Robust time delay estimation exploiting redundancy among multiple microphones. *IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 11(6)*, 2003