

# The AIT 3D Audio / Visual Person Tracker for CLEAR 2007

Nikos Katsarakis, Fotios Talantzis,  
Aristodemos Pnevmatikakis and Lazaros Polymenakos

Athens Information Technology, Autonomic and Grid Computing,  
P.O. Box 64, Markopoulou Ave., 19002 Peania, Greece  
{nkat, fota, apne, lcp}@ait.edu.gr  
<http://www.ait.edu.gr/research/RG1/overview.asp>

**Abstract.** This paper presents the Athens Information Technology system for 3D person tracking and the obtained results in the CLEAR 2007 evaluations. The system utilizes audiovisual information from multiple acoustic and video sensors. The proposed system comprises a video and an audio subsystem whose results are suitably combined to track the last active speaker. The video subsystem combines in 3D a number of 2D face localization systems, aiming at tracking all people present in a room. The audio subsystem uses an information theoretic metric upon an ensemble of microphones to estimate the active speaker.

## 1 Introduction

Three dimensional person tracking from multiple synchronized audiovisual sensors has many applications, like surveillance, security, smart spaces [1], pervasive computing, and human-machine interfaces [2] to name a few. In such trackers, body motion is the most widely used video cue, while speech is the audio cue. As speech is not always present for all people present in the monitored space, a stand-alone audio tracker cannot provide continuous tracks. A video tracker on the other hand can loose track of the people due to clutter from other people and the background. In this case the audio cue can help resolve the tracks.

In this paper an audiovisual approach towards 3D tracking is employed. The stand-alone video tracker aims at tracking all people present in the monitored space from the synchronized recordings of multiple calibrated cameras [3] to produce 3D tracks from multiple 2D face trackers [4]. The audio tracker aims at finding the active speaker. It employs an information theoretic approach [5] for direction-of-arrival estimation, as this can be combined using multiple clusters of microphones [6]. The multimodal tracker combines the two into a system that tracks the last active speaker, even after he or she has stopped talking.

This paper is organized as follows: In sections 2 to 4 the audio, video and multimodal combination modules of the tracker are detailed. The results on CLEAR 2007 evaluations are presented and discussed in section 5. Finally, in section 6 the conclusions are drawn, followed by some indications for further work.

## 2 Audio 3D Tracker

An audio-based localization system is typically employed in a reverberant environment and it considers  $M$  microphones arranged in  $P$  pairs. The sound source that the system attempts to locate and track is assumed to be in the far field of the microphones. Therefore, we can approximate the spherical wavefront emanating from the source as a plane wavefront of sound waves arriving at the microphone pairs in a parallel manner. The discrete signal recorded at the  $m$ -th microphone ( $m = 1, 2$ ) of the  $p$ -th pair at time  $k$  is:

$$x_{mp}(k) = h_{mp}(k) * s(k) + n_{mp}(k), \quad p = 1, 2, \dots, P \quad (1)$$

where  $s(k)$  is the source signal,  $h_{mp}(k)$  is the room impulse response between the source and  $m$ -th microphone,  $n_{mp}(k)$  is additive noise, and  $*$  denotes convolution. The length of  $h_{mp}(k)$ , and thus the number of reflections, is a function of the reverberation time  $T_{60}$  (defined as the time in seconds for the reverberation level to decay to 60 dB below the initial level) of the room and expresses the main problem when attempting to track an acoustic source. This is because when the system is used in reverberant environments, the source location can be estimated in a spurious location created by the reflections.

Most of the localization systems are required to operate in real time. Therefore we assume that data at each sensor  $m$  are collected over  $t$  frames of data  $\mathbf{x}_{mp}^{[t]} = [x_{mp}(tL), x_{mp}(tL+1), \dots, x_{mp}(tL+L-1)]$  of  $L$  samples. So, we can have a representation of the microphone data until time frame  $t$  by forming the array:

$$\mathbf{x}_{1:t} = \begin{bmatrix} \mathbf{x}_{11}^{[1]} & \mathbf{x}_{11}^{[2]} & \dots & \mathbf{x}_{11}^{[t]} \\ \mathbf{x}_{21}^{[1]} & \mathbf{x}_{21}^{[2]} & \dots & \mathbf{x}_{21}^{[t]} \\ \vdots & \vdots & & \vdots \\ \mathbf{x}_{1p}^{[1]} & \mathbf{x}_{1p}^{[2]} & \dots & \mathbf{x}_{1p}^{[t]} \\ \mathbf{x}_{2p}^{[1]} & \mathbf{x}_{2p}^{[2]} & \dots & \mathbf{x}_{2p}^{[t]} \end{bmatrix} \quad (2)$$

Since the microphones of each pair reside in different spatial locations, their corresponding recordings will be delayed with respect to each other by a relative time delay  $\tau_p$ . The first aim of localization systems is to retrieve this Time Delay Estimate (TDE) between the two microphones of each pair  $p$ . This typically performed by looking for the delay that maximizes some criterion amongst some set of candidate delays. Using all estimated  $\tau_p$  the localizer can then provide an estimate of the source location. Traditional systems typically do this by converting  $\tau_p$  to a line along which the estimated source position is. The problem of localization then reduces to finding the location which minimizes the distance to each intersection points of the bearing lines [7]. In the context of the present work we use an alternative approach. We still

collect all  $\tau_p$  but these serve as a feed to a Particle Filter (PF) that has an embedded Voice Activity Detection (VAD) functionality. PFs allow us to integrate the properties of human motion while the VAD module helps us to deal with the silence periods existing in the utterances of the active speaker.

The following paragraphs describe separately the sub-systems of the audio tracker. First the general PF framework for localization is presented. Then we briefly describe the TDE estimation method and finally we review the VAD system.

## 2.1 State-Space Estimation Using Particle Filters

Assuming a first order model for the source dynamics, the source state at any frame  $t$  is given as:

$$\mathbf{a}_t = [X_t, Y_t, Z_t, \dot{X}_t, \dot{Y}_t, \dot{Z}_t]^T \quad (3)$$

where  $s_a = [X_t, Y_t, Z_t]$  is the source location and  $[\dot{X}_t, \dot{Y}_t, \dot{Z}_t]$  the corresponding source velocity. If we could calculate the conditional density  $q(\mathbf{a}_t | \mathbf{x}_{1:t})$ , we could then find the source location by choosing the state that is more likely given the sensor data until frame  $t$ . We can perform this by using [8]:

$$q(\mathbf{a}_t | \mathbf{x}_{1:t}) \propto q(\mathbf{x}_t | \mathbf{a}_t) q(\mathbf{a}_t | \mathbf{x}_{1:t-1}) \quad (4)$$

where we have set  $\mathbf{x}_t \equiv \mathbf{x}_{1:t}$ . Also,  $q(\mathbf{x}_t | \mathbf{a}_t)$  is the likelihood and  $q(\mathbf{a}_t | \mathbf{x}_{1:t-1})$  is known as the prediction density and it is given as [8]:

$$q(\mathbf{a}_t | \mathbf{x}_{1:t-1}) = \int q(\mathbf{a}_t | \mathbf{a}_{t-1}) q(\mathbf{a}_{t-1} | \mathbf{x}_{1:t-1}) d\mathbf{a}_{t-1} \quad (5)$$

where  $q(\mathbf{a}_t | \mathbf{a}_{t-1})$  is the state transition density, and  $q(\mathbf{a}_{t-1} | \mathbf{x}_{1:t-1})$  is the prior filtering density. The solution to (4) and (5) can be found using a Monte-Carlo simulation of a set of particles with associated discrete probability masses that estimate the source state. For this we require a model of how the source propagates from  $\mathbf{a}_{t-1}$  to  $\mathbf{a}_t$ . To keep consistent with the literature we will use the Langevin model [9]. We also need a function to measure the likelihood of the microphone data, which is [9]:

$$q(\mathbf{x}_t | \mathbf{a}_t) = \prod_{p=1}^P q_p(\mathbf{x}_t | \mathbf{a}_t) \quad (6)$$

where

$$q_p(\mathbf{x}_t | \mathbf{a}_t) = \max\left(\mathbf{R}_p^{[t]}(\tau_p^{[a_t]}), p_0\right)^2 \quad (7)$$

and  $p_0$  is some prior probability that none of the potential source locations is the true one.  $\mathbf{R}_p^{[t]}$  is a TDE function evaluated only at a set of candidate delays given as:

$$\tau_p^{[a_t]} = \left( \|s_{a_t} - \mathbf{m}_{1p}\| - \|s_{a_t} - \mathbf{m}_{2p}\| \right) / c \quad (8)$$

where  $\mathbf{m}_{mp}$  denotes the location of the  $m$ -th microphone at the  $p$ -th pair and  $c$  is the speed of sound.

There are occasions where reverberation or noise sources can trap the particles in a spurious location. For this we use an external PF  $\mathbf{e}_t$  that has the same architecture as the main one  $\mathbf{a}_t$  but it is initialized at every frame  $t$ . If these new particles estimate a source location that is  $d_e$  away from the main PF for a significant amount of time  $T_e$  then we reset the locations of the particles of the main PF to those of the external.

## 2.2 Time Delay Estimation

A variety of methods like GCC [10] exist for TDE. For any pair  $p$  the GCC-PHAT variant  $R_t(\tau)$  is defined as the cross correlation of  $\mathbf{x}_{1p}^{[t]}$  and  $\mathbf{x}_{2p}^{[t]}$ , filtered by a weighting function  $g$  for a range of delays  $\tau$ . Let  $X_{1p}^{[t]}(\omega)$ ,  $X_{2p}^{[t]}(\omega)$  and  $G(\omega)$  denote the  $L$ -point discrete Fourier transforms of the microphone signals and the  $g$  function for the  $p$ -th pair at frame  $t$ . Since the TDE analysis is independent of the data frame, we drop  $t$  to express frames simply as  $X_{mp}^{[t]}(\omega)$  for any  $t$ . In the context of our model, and for any set of frames, we may then write  $X_{1p}(\omega) = X_{2p}(\omega)e^{-j\omega\tau_p}$ . Thus, the problem is essentially to estimate the correct value of  $\tau_p$  for which the two recordings are synchronized. We can express  $R_t(\tau)$  as:

$$R_t(\tau) = \frac{1}{2\pi} \sum_{\omega} G(\omega) X_{1p}(\omega) X_{2p}^*(\omega) e^{j\omega\tau} \quad (9)$$

with

$$G(\omega) = 1 / |X_{1p}(\omega) X_{2p}^*(\omega)| \quad (10)$$

Ideally,  $R_t(\tau)$  exhibits a global maximum at the lag value which corresponds to the correct  $\tau$ . Thus, an estimation of  $\tau_p$  can be obtained by  $\tau_p = \arg \max_{\tau} R(\tau)$ .

## 2.3 Voice Activity Detection

The VAD employed for the purposes of our system is presented in [11]. It is a conceptually simple system that operates fast enough to be used in conjunction with real-time implementations like our speaker localization system. More particularly the used VAD extends the use of statistical models in speech detection by employing a decision-directed parameter estimation method for the likelihood ratio test that defines the presence of speech. This is combined with an effective hang-over scheme

which considers the previous observations by a first-order Markov process modeling of speech occurrences. This serves as an improvement to typical VAD algorithms which normally operate on heuristics.

We then fuse the VAD decision with the tracking system in the same manner authors did in [12]. This is done by noting that the probability  $1 - p_0$  corresponds to the likelihood of the acoustic source being active i.e. the result returned by the VAD. Thus, we can choose  $p_0$  to vary in time in accordance to the result of the VAD as  $p_0^{[t]} = 1 - a^{[t]}$  with  $a^{[t]} \in [0,1]$  the return value of the VAD with 1 denoting presence and 0 absence of speech in time frame  $t$ .

The summary of the proposed algorithm can be itemized as follows:

- 1) Start with a set of particles  $\mathbf{a}_0^{[i]}, i = 1 \dots N$  with uniform weights  $w_0^{[i]}, i = 1 \dots N$ . For every new frame of data perform steps 2-7.
- 2) Resample the particles from state  $\mathbf{a}_{t-1}^{[i]}$  using some resampling method [13] (we used the *residual resampling* algorithm) and form the resampled set of particles  $\tilde{\mathbf{a}}_{t-1}^{[i]}, i = 1 \dots N$ .
- 3) Using the Langevin model, propagate  $\tilde{\mathbf{a}}_{t-1}^{[i]}$  to predict the current set of particles  $\mathbf{a}_t^{[i]}$ .
- 4) Take a set of frames of  $L$  samples from each microphone and convert them into the frequency domain to get  $X_{mp}(\omega), m = 1, 2, p = 1 \dots P$ .
- 5) For every pair of microphones calculate the MI only at the time delays corresponding to  $\mathbf{a}_t^{[i]}$ , using the delays found by (8).
- 6) Weight the particles using the likelihood function i.e.  $w_t^{[i]} = p(\mathbf{x}_t | \mathbf{a}_t^{[i]}), i = 1 \dots N$  and normalize the weights so that they add up to unity.
- 7) The source location for the current frame is then given as the weighted average of the particles:  $\mathbf{s}_t = \sum_{i=1}^N w_t^{[i]} \mathbf{s}_a^{[i]}$ . If the external PF  $\mathbf{e}_t^{[i]}$  source estimate remains at a distance greater than  $d_e$  for more than  $T_e$  sec then set  $\mathbf{a}_t^{[i]} = \mathbf{e}_t^{[i]}, i = 1 \dots N$ .

### 3 Video 3D Tracker

The video 3D tracker employs multiple instances of the AIT 2D face tracker [4], each operating on the synchronized video streams of multiple calibrated cameras [3]. Possible associations of the different views of a face are constructed by projecting a grid of 3D points onto the different image planes and collecting face evidence. A stochastic tracker then selects the best association. Additionally, a panoramic camera view is also processed by the AIT body tracker [14] used in the CLEAR 2006 evaluations. The 3D locations of the people estimated by associating the faces are projected onto the panoramic camera plane and are validated by the tracked bodies in that camera view.

### 3.1 Face Tracker

The 2D face localization system is detailed in [4]. Face localization is constrained in the body areas provided by the AIT body tracker [14]. Three face detectors for frontal and left/right profile faces provide candidate face regions in the body areas. They are of the cascades of simple features type, trained with AdaBoost. The face candidates are validated using the probability scores from a Gaussian Mixture Model trained from texture and color properties of faces. The surviving candidates are checked for possible merging, as both the profile detectors and the frontal one can detect different portions of the same face if the view is half-profile. The resulting face candidates are associated with faces existing in the previous frame and also with tracks that currently have no supporting evidence and are pending to either get an association, or be eliminated. Any faces of the previous frame that do not get associated with candidate faces at the current frame have a CAM-Shift tracker initiated to attempt to track similarly colored regions in the current frame. If CAM-Shift also fails to track, then these past faces have their track in pending status for  $F_p$  frames. Finally, all active face tracks are checked for duplicates, i.e. high spatial similarity. In the following subsections, the various modules of the system are detailed.

### 3.2 Mapping of faces in 2D into heads in 3D

Our approach for 3D tracking utilizes the 2D face localization system presented in the previous section, applied on multiple calibrated [3] and synchronized cameras. To solve the problem of associating the views of the face of the same person from the different cameras, a 3D space to 2D image planes approach is utilized. The space is spanned by a 3D grid. Each point of the grid is projected onto the different image planes. Faces whose centers are close to the projected points are associated to the particular 3D point. 3D points that have more than one face associated to them are used to form possible associations of views of the face of the same person from the different cameras. If in each camera view  $c$  there are  $n_c$  faces then the  $k$ -th association (of the total  $K$  ones) that span the 3D space is of the form  $a^{(k)} = \{i_1^{(k)}, \dots, i_c^{(k)}\}$ , where  $C$  is the number of available cameras and  $i_c^{(k)} \in \{0, 1, \dots, n_c\}$ . A value  $i_c^{(k)} = 0$  corresponds to no face from the  $c$ -th camera in the  $k$ -th association, while any other value corresponds to the membership of a face from those in the  $c$ -th camera in the  $k$ -th association. Obviously  $\forall c \in \{1, \dots, C\}, i_c^{(k_i)} > 0$  and  $k_1 \neq k_2$ , it is  $i_c^{(k_1)} \neq i_c^{(k_2)}$ , i.e. the same face in a camera view cannot be a member of different valid associations. This condition renders some of the associations mutually exclusive. After eliminating duplicate associations, the remaining ones are grouped into possible sets of mutually exclusive associations and sorted according to a weight that depends on the distance of each association from the face center and on the number of other associations that contradict it.

All the  $M$  mutually exclusive sets of possible associations  $a^{(k)}$  are validated using a Kalman filter in the 3D space. For each new frame, all possible solutions are

compared to the state established on the previous frame, penalizing solutions which fail to detect previously existing targets, or in which there are detections of new targets in the scene. While this strategy reduces the misses and false positives, it does not prevent new targets from appearing, as in the case of new people entering the room, all solution pairs will include that new target and thus will be equally penalized.

#### 4 Audiovisual 3D Tracker

The audiovisual tracker uses the synchronized outputs of both Audio 3D and Video 3D person trackers to generate the audiovisual output. It constantly keeps track of all provided video tracker positions for the current frame. However, no output is provided, unless the voice activity detector of the audio tracker indicates that only one person is talking at the time period under consideration. On the other hand, whenever the audio tracker indicates that there is a person talking, the system tries to find the closest match between the provided speaker position and any of the recently updated video states. If there is such an association, the person is tracked by the video tracker, keeping consistent track id over time, otherwise the system outputs the position given by the audio localizer. A flowchart of the system is shown in Figure 1.

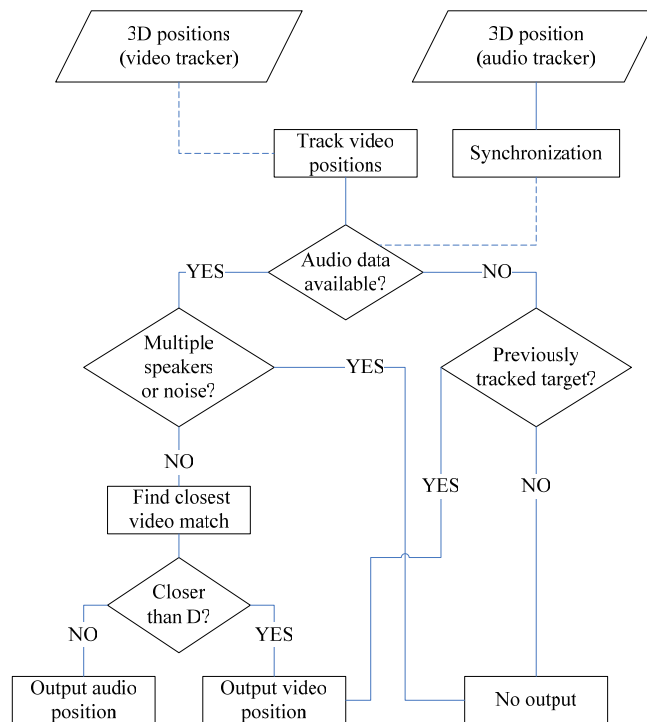


Fig. 1. Audiovisual tracker flowchart.

## 5 CLEAR 2007 Evaluation Results

The video, audio and audiovisual trackers presented in sections 2, 3 and 4 are evaluated using the CLEAR 2007 multimodal recordings. Typical results are shown in Figure 2.



**Fig. 2.** Operation of the individual 2D face trackers on the four corner cameras and association of the 2D evidence into 3D tracks. The detected faces are marked by bounding boxes. The IDs of the tracks are of the form AIT\_XXX shown at the projection of the tracked head centroids on the floor. The tracks are also projected to a panoramic camera (not used by the system) for better visualization.

The quantitative evaluation of the proposed system follows the CLEAR2007 evaluation protocol [15]. According to it, the tracking system outputs (hypotheses) are mapped to annotated ground truths based on centroid distance and using the Hungarian algorithm [16]. The metrics for face tracking are five [15]. The Multiple Object Tracking Precision (MOTP) is the position error for all correctly tracked persons over all frames. It is a measure of how well the system performs when it actually finds the face. There are three kinds of errors for the tracker, false positives, misses and track identity mismatches. They are reported independently and also jointly in an accuracy metric, the Multiple Object Tracking Accuracy (MOTA). The MOTA is the residual of the sum of these three error rates from unity. For the audio system, there is no penalty for identity switches, hence the MOTA only have two

components: The misses and the false positives. The quantitative performance of the video, audio and audiovisual systems is summarized in Tables 1-4.

**Table 1.** Video-based person tracking performance of the system on all 40 seminars, averaged per site.

Site	MOTP (mm)	MOTA (%)	Misses (%)	False positives (%)	Mismatches (%)
AIT	81.22	60.15	33.67	3.57	2.61
IBM	93.56	60.39	29.39	8.47	1.75
ITC	97.73	59.66	30.45	6.83	3.06
UKA	93.03	46.67	45.66	5.11	2.56
UPC	88.70	65.13	23.77	8.47	2.63
<b>Overall</b>	<b>91.29</b>	<b>58.20</b>	<b>32.56</b>	<b>6.76</b>	<b>2.48</b>

**Table 2.** Video-based person tracking performance of the system on all 40 seminars, averaged per site, including panoramic camera.

Site	MOTP (mm)	MOTA (%)	Misses (%)	False positives (%)	Mismatches (%)
AIT	80.11	58.65	36.29	2.50	2.56
IBM	92.44	61.19	30.50	6.56	1.75
ITC	97.73	59.66	30.45	6.83	3.06
UKA	93.04	43.36	52.09	2.23	2.32
UPC	88.74	67.69	23.28	6.46	2.57
<b>Overall</b>	<b>90.88</b>	<b>58.03</b>	<b>34.51</b>	<b>5.05</b>	<b>2.41</b>

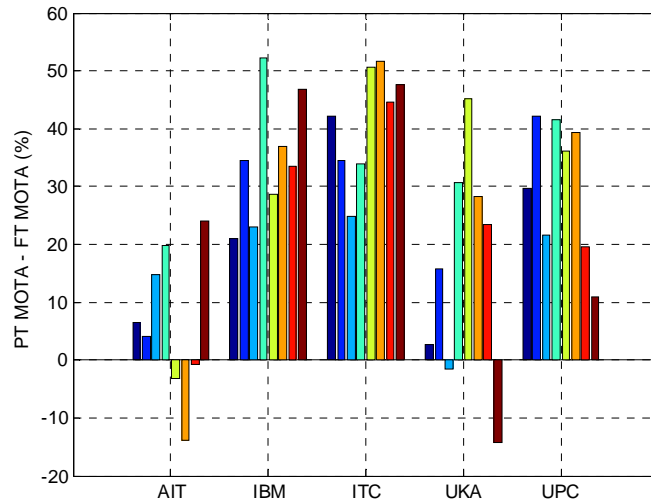
**Table 3.** Audio-based person tracking performance of the system on all 40 seminars, averaged per site.

Site	MOTP (mm)	A-MOTA (%)	Misses (%)	False positives (%)
AIT	229.00	25.85	51.45	22.70
IBM	266.76	10.16	66.10	23.74
ITC	269.99	-12.46	56.97	55.49
UKA	316.70	-27.65	66.05	61.60
UPC	214.19	11.60	64.32	24.08
<b>Overall</b>	<b>256.89</b>	<b>1.46</b>	<b>61.39</b>	<b>37.15</b>

**Table 4.** Audiovisual person tracking performance of the system on all 40 seminars, averaged per site.

Site	MOTP (mm)	MOTA (%)	Misses (%)	FPs (%)	Mismatches (%)
AIT	93.32	29.02	60.57	6.67	3.74
IBM	144.62	5.34	83.49	8.83	2.34
ITC	122.21	2.29	64.19	28.80	4.72
UKA	91.15	-5.22	70.02	32.42	2.78
UPC	98.37	21.68	66.13	9.25	2.94
<b>Overall</b>	<b>107.70</b>	<b>10.01</b>	<b>69.33</b>	<b>17.36</b>	<b>3.30</b>

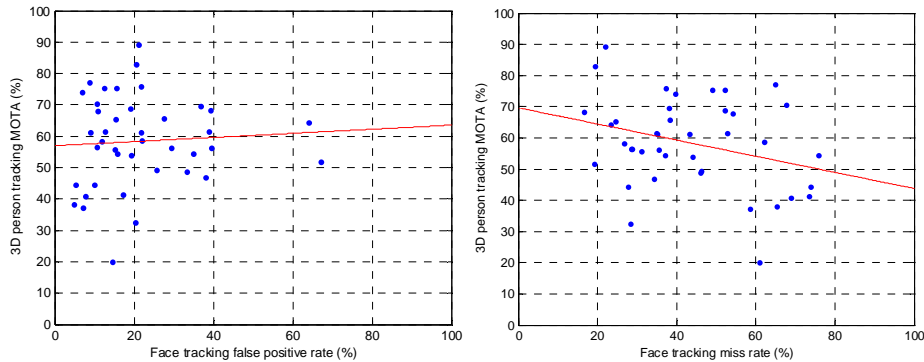
It is instructive to compare the MOTA obtained by the 2D face localization system, to that of the 3D person tracking system. This way we understand the benefit of combining the 2D information into 3D. This is done in Figure 5 for all seminars. Obviously, for seminars where the 2D MOTA has been low, there is a lot to gain. Only rarely the 3D MOTA is lower than the 2D, since misses, false positives and identity switches can be constrained by combining all four cameras. It is also important to understand which component of the 2D face tracking MOTA is more correlated with the 3D person tracking MOTA, affecting it the most: the misses or the false positives. The scatter plots of Figure 6 address this question. It is evident from the scatter and the slope of the linear regression that the 3D MOTA is not strongly correlated to the 2D false positives. On the other hand there is a strong correlation of the 3D MOTA to the 2D misses.



**Fig. 3.** Improvement of the 3D visual person tracking MOTA over the 2D face tracking MOTA. There is significant gain from combining the 2D face locations into the 3D locations.

## 6 Conclusions

In this paper we have presented and evaluated a 3D audiovisual tracking system that employs multiple audio and video sensors. The video subsystem exhibits good performance, consistent among the different test sites. On the other hand, the audio tracker does not perform as expected, especially on 2 out of the 5 test sites. Preliminary analysis of the results shows that a main cause of errors is the voice activity detection module. The audiovisual module improves the performance of the audio tracker, showing that further development of the latter will bring promising results.



**Fig. 4.** Correlation of the 3D person tracking MOTA to the 2D face tracking false positives and misses. As the slope of the linear regression shows, the most important factor affecting the 3D MOTA is the 2D face misses.

## Acknowledgements

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909. The authors wish to thank the people involved in data collection, annotation and overall organization of the CLEAR 2007 evaluations for providing such a rich test-bed for the presented algorithms.

## References

- [1] A. Waibel, H. Steusloff, R. Stiefelhagen, et. al: CHIL: Computers in the Human Interaction Loop, *5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Lisbon, Portugal, (Apr. 2004).
- [2] A. Pnevmatikakis, F. Talantzis, J. Soldatos and L. Polymenakos: Robust Multimodal Audio-Visual Processing for Advanced Context Awareness in Smart Spaces, *Artificial Intelligence Applications and Innovations*, Peania, Greece, (June 2006).
- [3] Z. Zhang: A Flexible New Technique for Camera Calibration, Technical Report MSR-TR-98-71, Microsoft Research, (Aug. 2002).
- [4] A. Stergiou, G. Karame, A. Pnevmatikakis and L. Polymenakos: The AIT 2D face detection and tracking system for CLEAR2007, *CLEAR 2007*, (May 2007).
- [5] F. Talantzis, A. G. Constantinides, and L. Polymenakos: Estimation of Direction of Arrival Using Information Theory, *IEEE Signal Processing*, 12, 8 (Aug. 2005), 561-564.
- [6] F. Talantzis, A. G. Constantinides, and L. Polymenakos: Real-Time Audio Source Localization Using Information Theory, *Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006)*, (May 2006).
- [7] M. S. Brandstein, J. E. Adcock and H. Silverman: A Closed-Form Location Estimator for Use with Room Environment Microphone Arrays, *IEEE Trans. on Acoust. Speech and Sig. Proc.*, 5 (1997), 45-50.

12 **Nikos Katsarakis, Fotios Talantzis,  
Aristodemos Pnevmatikakis and Lazaros Polymenakos**

- [8] N.J. Gordon, D.J. Salmond, and A.F.M. Smith: Novel approach to nonlinear/nongaussian bayesian state estimation, *IEE Proceedings-F (Radar and Signal Processing)*, pp. 140(2):107113, 1993.
- [9] J. Vermaak and A. Blake: Nonlinear filtering for speaker tracking in noisy and reverberant environments, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, USA, vol. 5, pp. 30213024, May 2001.
- [10] C. H. Knapp and G. C. Carter: The generalized correlation method for estimation of time delay, *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-24, 4 (Aug. 1976), 320–327.
- [11] J. Sohn, N.S. Kim and W. Sung: A statistical model-based voice activity detection, *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.
- [12] E.A. Lehmann and A.M. Johansson: Particle Filter with Integrated Voice Activity Detection for Acoustic Source Tracking, *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 50870, 2007.
- [13] M. Bolic, P.M. Djuric, S. Hong: New Resampling Algorithms for Particle Filters, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, vol. 2, pp. 589-592, 2003.
- [14] A. Pnevmatikakis and L. Polymenakos: 2D Person Tracking Using Kalman Filtering and Adaptive Background Learning in a Feedback Loop, *CLEAR 2006*, (Apr. 2006).
- [15] D. Mostefa et. al: CLEAR Evaluation Plan, document CHIL-CLEAR-V1.1-2006-02-21, (Feb 2006).
- [16] S. Blackman: *Multiple-Target Tracking with Radar Applications*, Artech House, Dedham, MA (1986), chapter 14.