MULTIMODAL PERSON TRACKING IN A SMART-ROOM ENVIRONMENT

C. Segura, A. Abad, C. Nadeu, J. Hernando

Technical University of Catalonia, Barcelona, Spain
{csegura,alberto,climent,javier}@gps.tsc.upc.edu

ABSTRACT

Tracking speakers in multiparty conversations constitutes a fundamental task for automatic meeting analysis. In this work, we present the Person Tracking systems developed at UPC for audio, video and audio-video modalities. The proposed systems are designed to deal robustly in both single and multiperson localization task independently on the environmental conditions. Novelties proposed are aimed to enhance the accuracy of the system independently on the application scenario and to reduce the computational complexity. Besides the technology description, experimental results conducted for the CLEAR evaluation workshop are also reported.

1. INTRODUCTION

The automatic analysis of meetings recorded in multi-sensor rooms is an emerging research field. In this domain, localizing and tracking people and their speaking activity play fundamental roles in several applications, like microphone array beamforming or steering of pan-tilt-zoom cameras towards the active speaker. To locate persons with unobtrusive far-field sensors, either video or audio sources can be used, though eventually the most accurate and robust techniques will likely be based on multimodal.

The degree of reliable information provided by person localization systems on the basis of the audio and video signals collected in a smart-room environment with a distributed microphone and video network, depends on a number of factors such as environmental noise, room reverberation, person movements and camera occlusions. These factors, among others, demand an effort on the development of new robust systems capable of dealing with adverse environments.

In the present work, we get an insight on the development and design of robust Person Tracking systems based on audio, video and audio-video modalities. Results obtained in the CLEAR evaluation campaign and comparison among mono and multimodal systems are provided showing the performance of the proposed algorithms.

This work has been partially sponsored by the EC-funded project CHIL [1] (IST-2002-506909) and by the Spanish Government-funded project ACESCA (TIN2005-08852).

2. AUDIO PERSON TRACKING SYSTEM

Conventional acoustic person localization and tracking systems can be split into three basic stages. In the first stage, estimations of such information as Time Difference of Arrival or Direction of Arrival is usually obtained from the combination of the different microphones available. In general, in the second stage the set of relative delays or directions of arrival estimations are used to derive the source position that is in the best accordance with them and with the given geometry. In the third optional stage, a tracking of the possible movements of the sources according to a motion model can be employed.

The SRP-PHAT [2] algorithm (also known as Global Coherence Field [3]) performs and integrates the two first stages of localization in a robust and smart way. In general, the goal of localization techniques based on SRP (Steered Response Power) is to maximize the power of the received sound source signal using a delay-and-sum or a filter-and-sum beamformer. In the simplest case, the output of the delay-and-sum beamformer is the sum of the signals of each microphone with the adequate steering delays for the position that is explored. Thus, a simple localization strategy is to search for the energy peak through all the possible positions in 3D space. Concretely, SRP-PHAT algorithm searches for the maximum of the contribution of the cross-correlations between all the microphone pairs across the space. The main strength of this technique consists on the combination of the simplicity of the steered beamformer approach with the robustness offered by the PHAT weighting.

The proposed system for Audio Person Tracking is based on the SRP-PHAT algorithm with some additional robust modifications. The system design has been aimed to develop a robust system with independence on the acoustic and room conditions, such as the number of sources, their maneuvering modes or the number of microphones.

2.1. Brief Description of the SRP-PHAT Algorithm

As already mentioned above, the SRP-PHAT algorithm searches for the maximum of the contribution of the cross-correlations between all the microphone pairs across the space. The process can be summarized into four basic steps:

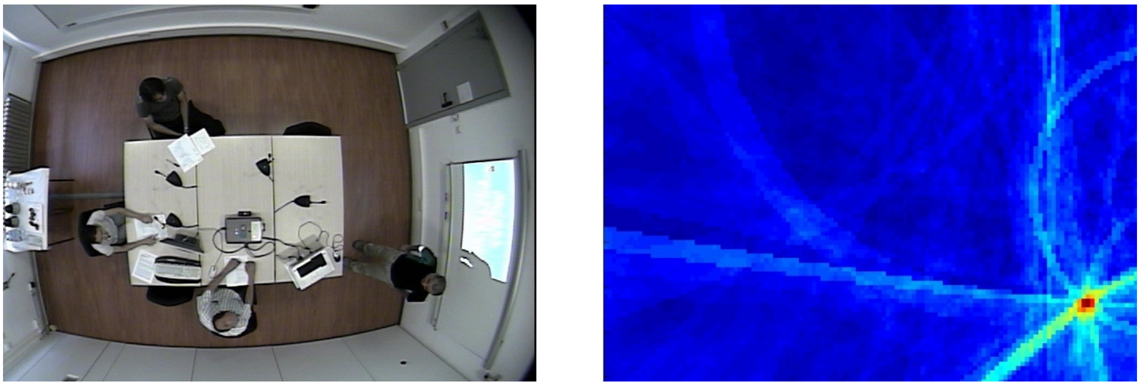

Figure 1. On the left, zenithal camera snapshot. On the right, example of the *Sound Map* obtained with the SRP-PHAT process.

Step 1 The exploration space is firstly split into small regions (typically of 5-10 cm). Then, theoretical delays from each possible exploration region to each microphone pair is pre-computed and stored.

Step 2 Cross-correlations of each microphone pair are estimated for each analysis frame. Concretely, the Generalized Cross Correlation with PHAT weighting [4] is considered. It can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density ($\widehat{G}_{x_1x_2}(f)$) as follows,

$$\widehat{R}_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \frac{\widehat{G}_{x_1x_2}(f)}{|\widehat{G}_{x_1x_2}(f)|} e^{j2\pi f\tau} df \quad (1)$$

Step 3 The contribution of the cross-correlations is accumulated for each exploration region using the delays pre-computed in *Step 1*. In this way, it is obtained a kind of *Sound Map* as the one shown in Figure 1.

Step 4 Finally, the position with the maximum score is selected as the estimated position.

2.2. The Implementation of the Robust Audio Person Tracker

On the basis of the conventional SRP-PHAT, a robust system for Audio Person Tracking is developed. The main novelties introduced and some aspects related to other implementation details are introduced in the following.

2.2.1. Implementation Details

The analysis frame consists of Hanning windowed blocks of 4096 samples, 50% overlapped, obtained at a sample rate of 44.1 kHz. The FFT computation dimension is fixed to 4096 samples.

2.2.2. Adaptive Smoothing Factor for the Cross-Power Spectrum (CPS) Estimations

Smoothing over time of the GCC-PHAT estimations is a simple and efficient way of adding robustness to the system. This smoothing can be done in the time domain (GCC-PHAT) or in the frequency domain (CPS). Considering the smoothed cross-power spectrum $\widehat{G}_{x_1x_2}(k, f)$ in time instant k and the instantaneous estimation $G_{x_1x_2}(k, f)$ our system performs the smoothing in the frequency domain as follows,

$$\widehat{G}_{x_1x_2}(k, f) = \beta\widehat{G}_{x_1x_2}(k-1, f) + (1-\beta)G_{x_1x_2}(k, f) \quad (2)$$

From experimental observation it can be seen that the right selection of this β factor is crucial in the system design. A high smoothing value can greatly enhance the results obtained in an almost static scenario, while it can be dramatically inconvenient in a scenario with many moving speakers.

Hence, an adaptive smoothing factor has been designed. This adaptive factor is obtained based on the velocity estimation provided by a Kalman filter.

2.2.3. Two-Pass SRP Search

It can be seen from experimental observations that most of the information for a rough localization is concentrated in the low-frequency bins of the GCC-PHAT, while high frequency bins are useful in order to obtain a finest estimation given a first coarse estimation. Taking into account this observation a two-pass SRP search has been designed:

Coarse Search This search procedure is performed only in the *x*-*y* axis (z is assumed to be 1.5 m), with a searching cell dimension of 16 cm and only using the low frequency information of the cross-correlations ($f < 9kHz$). A first coarse estimation is obtained from this search, say $(x_1, y_1, 150)$ cm.

Fine Search A new limited search area around the obtained *coarse* estimation is defined $(x_1 - 50 : x_1 + 50, y_1 - 50 : y_1 + 50, 110 : 190)$ cm. In this new fine search, dimension of the cell search is fixed to 4 cm for the *x*-*y* axis and to 8 cm for the *z*-axis. In the *fine search* all the frequency information of the cross-correlations is used and a more accurate estimation is obtained.

Moreover, the double SRP searching procedure is adequate to reduce computational load, since the *fine* exploration is only performed across a very limited area.

2.2.4. *Confidence Threshold*

In SRP-PHAT algorithm the position with the maximum value obtained from the accumulated contributions of all the correlations is selected (*Step 4*). This value is assumed to be well-correlated with the likelihood of the given estimation. Hence, this value is compared to a fixed threshold (depending on the number of microphone-pairs used) to reject/accept the estimation. The threshold has been experimentally fixed to 0.5 for each 6 microphone pairs.

Finally, it is worth noting that although a Kalman filter is used for the estimation of the adaptive CPS smoothing factor, it is not considered for tracking purposes. The reason is that the Kalman filter design and the data association strategies adopted showed a different impact depending on the scenario. In other words, it showed to be too much dependent on the number and the velocities of sources to perform correctly.

3. VIDEO PERSON TRACKING SYSTEM

For this task we propose creating a 3D representation of the room combining the views from a calibrated [5] set of cameras. The scene is discretized into box-shaped regions, voxels, and then each voxel is classified as foreground or background. Indeed, the foreground voxels provide enough information for precisely object detection and tracking. The image at each camera view is segmented forming foreground regions. This foreground regions are then projected and combined in the 3D space forming volumes, where the 3D objects must lie, using the Shape from silhouette technique. The main drawback of the method is that it doesn't always capture the true shape of the object, as concave shape regions are not expressed in the silhouettes. However, this is not a severe problem in a tracking application as the aim is not to reconstruct photorealistic scenes.

After the voxelization process (see figure 2), a connected component analysis *CCA* follows to cluster and label the voxels into meaningful 3D-blobs [6, 7, 8] , from which some representative features are extracted. Finally, there is a template-based matching process aiming to find persistent blob correspondences between consecutive frames.

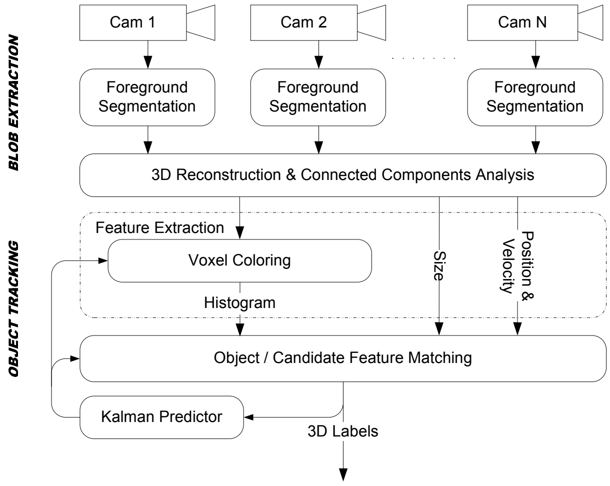

Figure 2. The system block diagram showing the chain of functional modules

4. MULTIMODAL PERSON TRACKING SYSTEM

Multimodal Person Tracking is done based on the audio and video person tracking technologies described in the previous sections. These two technologies may have different nature, for example different frame rate, the video tracking system is able to track several persons, but usually only one person estimate is given by the audio tracking system and only when actively speaking, etc. A multi-modal system aiming on the fusion of information proceeding from these two technologies has to take into account these differences.

We expect to have far more position estimates from the video system than from the audio system since persons in the smart room are visible by the cameras during most of the video frames; in contrary, the audio system can estimate the person's position only if she/he is speaking (so called active speaker). Thus, the presented multimodal approach relies more on the video tracking system and it is extended to incorporate the audio estimates to the corresponding video tracks. This is achieved by first synchronizing the audio and video estimates and then using data association techniques. After that a decentralized Kalman filter is used to provide a global estimate of person's position. The frame rate of the multimodal tracking is the same as that of the video system.

4.1. Audiovisual Fusion

The Kalman filter algorithm provides an efficient computational solution for recursively estimating the position, in situations where the system dynamics can be described by a state-space model. A detailed description of the Kalman filter for tracking can be found in [10, 11].

The decentralized Kalman filter [12] is used for the fusion of audio and video position estimates. As shown in Figure 3, the system can be divided in two modules associated with the audio and video systems. Each modality computes a local a-posteriori estimate $\hat{\mathbf{x}}_i[k|k]$, $i = 1, 2$ of the person position using a local Kalman filter (KF1 and KF2, respectively), based on the corresponding observations $\mathbf{y}_1[k]$, $\mathbf{y}_2[k]$. These partial estimates are then combined to provide a global state estimate $\hat{\mathbf{x}}[k|k]$ at the fusion center such as:

$$\hat{\mathbf{x}}[x|x] = \mathbf{P}[k|k]\Big(\mathbf{P}^{-1}[k|k-1]\hat{\mathbf{x}}[k|k-1] + \sum_{i=1}^{2}\Big[\mathbf{P}_i^{-1}[k|k]\hat{\mathbf{x}}_i[k|k] - \mathbf{P}_i^{-1}[k|k-1]\hat{\mathbf{x}}_i[k|k-1]\Big]\Big) \tag{3}$$

$$\mathbf{P}^{-1}[k|k] = \mathbf{P}^{-1}[k|k-1] + \sum_{i=1}^{2}\Big[\mathbf{P}_i^{-1}[k|k] - \mathbf{P}_i^{-1}[k|k-1]\Big] \tag{4}$$

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

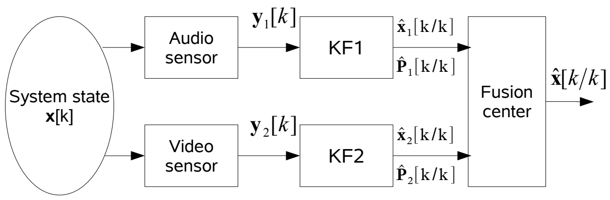

Figure 3. Structure of the decentralized Kalman filter. The fusion center combines the local estimates to compute a global estimate of the system state.

5. EVALUATION

Person Tracking evaluation is run on the data collected by the CHIL consortium for the CLEAR 06 evaluation. Two tasks are considered: single and multiperson tracking, based on non-interactive seminar (collected by ITC and UKA) and highly interactive seminar (collected by IBM, RESIT and UPC) recordings, respectively. Complete description of the data and the evaluation can be found in [13].

5.1. Summary of the Experimental Set-Up

5.1.1. Data description

Room set-ups of the contributing sites present two basic common groups of devices: the *audio* and the *video* sensors.

Audio sensors set-up is composed by 1 (or more) NIST Mark III 64-channel microphone array, 3 (or more) T-shaped 4-channel microphone cluster and various table-top and close-talk microphones.

Video sensors set-up is basically composed by 4 (or more) fixed cameras. In addition to the fixed cameras, some sites are equipped with 1 (or more) PTZ camera.

5.1.2. Evaluation metrics

Three metrics are considered for evaluation and comparison purposes:

Multiple Object Tracking Precision (MOTP) [mm] This is the precision of the tracker when it comes to determining the exact position of a tracked person in the room. It is the total Euclidian distance error for matched *ground truth-hypothesis* pairs over all frames, averaged by the total number of matches made.

Multiple Object Tracking Accuracy (MOTA) [%] This is the accuracy of the tracker when it comes to keeping correct correspondences over time, estimating the number of people, recovering tracks, etc. It is the sum of all errors made by the tracker, false positives, misses, mismatches, over all frames, divided by the total number of ground truth points.

Acoustic Multiple Object Tracking Accuracy (A-MOTA) [%] This is like the *original* MOTA metric in which all mismatch errors are ignored and it is used to measure tracker performance only for the active speaker at each point in time for better comparison with the acoustic person tracking results (where identity mismatches are not evaluated).

5.2. Audio Person Tracking Results

We have decided to use all the *T-clusters* available in the different seminars and only to use the *MarkIII* data of those sites where the *MarkIII* is located in a wall without a *T-cluster* (IBM, RESIT and UPC). In general, only microphone pairs of the same *T-cluster* or *MarkIII* array are considered by the algorithm.

In the experiments where the *MarkIII* is used, 6 microphone pairs are selected for GCC-PHAT computation The pairs selected out of the 64 microphones of *MarkIII* are *1-11*, *11-21*, *21-31*, *31-41*, *41-51* and *51-61*. Hence, an inter-microphone separation of 20 cm for each microphone-pair is considered.

In Table 1 individual results for each data set and average results for both tasks are shown. Notice that task results are not directly the mean of the individual results, since the scores are recomputed jointly. The evaluating system in both tasks is the same and the multi-person task is only evaluated when only one speaker is active. In this way mean performances obtained, as it could be expected, are quite similar. In fact, there is a fail in the multi-person task, but it is more related with the particular characteristics of each data set, that with the task indeed. For instance, UPC data is particularly noisy and present some challenging situations such as *coffee breaks*. Hence, we can conclude that acoustic tracking performs reasonably well in controlled scenarios with one or few alternative and non-overlapping speakers, while it shows a considerable decrease in difficult noisy scenarios with many moving and overlapping speakers.

Table 1. Audio results for both single and multi-person tracking.

Task	MOTP	Misses	False Positives	A-MOTA
ITC data	108mm	8.56%	1.46%	89.98%
UKA data	148mm	15.09%	10.19%	74.72%
Single Person	145mm	14.53%	9.43%	76.04%
IBM data	180mm	17.85%	10.54%	71.61%
RESIT data	150mm	12.96%	6.23%	80.80%
UPC data	139mm	32.34%	28.76%	38.89%
Multi Person	157mm	20.95%	15.05%	64.00%

5.3. Video Person Tracking Results

Seminar sequences from UPC and RESIT have been evaluated and results are reported in Table 2. Since our algorithm required empty room information, we were constrained to only evaluate UPC and RESIT. By analyzing the results in detail we reached the following conclusions.

The high number of False Positives (FP) is mainly due to the fact our algorithm detected many foreground objects after the 3D reconstruction due to shadows and other lighting artifacts. Moreover, MOTA is related with the FP score thus dropping as FP increases. Further research to avoid such problems include an improvement of the Kalman filtering and association rules. Since our tracking strategy relies on the 3D reconstruction, rooms with a reduced common volume seen by a number of cameras (typically less N-1 cameras) produce less accurate results. Other reconstruction schemes more accommodated to different camera placement scenarios are under research to generate reliable volumes even if a reduced number of cameras is viewing a given part of the room.

Table 2. Video results for the multiperson tracking.

Task	MOTP	Misses	False Pos.	Mism.	MOTA
RESIT data	205mm	26.67%	74.62%	2.18%	-3.47%
UPC data	188mm	16.92%	23.56%	5.85%	53.67%
Multi Person	195mm	21.24%	46.16%	4.22%	28.35%

5.4. Multimodal Person Tracking Results

Only seminar sequences from RESIT and UPC have been evaluated due to the constrains of the Video tracking system mentioned above. For the Multimodal Person Tracking task, two different scorings under two different conditions are defined. For the condition A, the scoring shows the ability to track the active speaker at the time segments that he is speaking, while under the condition B the scoring measures the ability to track all the persons in the room during all the seminar.

The results are reported in Tables 3 and 4 for each condition. It can be seen that the results are very similar to those of the Video Person tracking task. This observation suggests that the multimodal algorithm is mainly influenced by the performance of the video tracking system.

Table 3. Multimodal results for Condition A.

Task	MOTP	Misses	False Pos.	Mism.	A-MOTA
RESIT data	143mm	52.66%	7.14%	3.92%	40.20%
UPC data	101mm	29.48%	25.28%	6.35%	45.24%
Cond. A	118mm	41.18%	16.13%	5.12%	42.70%

Table 4. Multimodal results for Condition B.

Task	MOTP	Misses	False Pos.	Mism.	MOTA
RESIT data	201mm	26.43%	74.47%	2.20%	-3.10%
UPC data	190mm	17.95%	24.61%	5.98%	51.46%
Cond. B	195mm	21.71%	46.71%	4.31%	27.28%

6. CONCLUSIONS

In this paper we have presented the audio, video and audio-video Person Tracking systems developed by the UPC.Novelties proposed in the three systems have been specially designed to add robustness to scenario and environment variabilities. Results obtained in the CLEAR evaluation campaign show that the audio tracker performs reasonably well in situations with few non-overlapping speakers, while it shows a considerable loss of performance in some challenging and noisy situations that must be addressed. Improvement of the Kalman filtering and association rules are also expected to enhance the video system. Finally, the multimodal audio-video system shows a high dependence on the video results caused by the fusion procedure. Thus, future efforts will be devoted to develop new fusion strategies at a higher level.

7. REFERENCES

[1] CHIL Computers In the Human Interaction Loop. Integrated Project of the 6th European Framework Programme (506909). http://chil.server.de/, 2004-2007

[2] DiBiase, J., Silverman, H., Brandstein, M.: Microphone Arrays. Robust Localization in Reverberant Rooms, Chapter 8, Springer, January 2001

[3] 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. Proceedings of Interspeech 2005, Lisboa September 2005

[4] Knapp, C.H. and Carter, G.C.: The Generalized Correlation Method for Estimation of Time Delay Rooms. IEEE Trans. on Acoustics, Speech, and Signal Processing August 1976

[5] Zhang, Z.: A flexible new technique for camera calibration. Technical report, Microsoft Research, August 2002

[6] Landabaso, J.L., Xu, L-Q., Pardàs, M.: Robust Tracking and Object Classification Towards Automated Video Surveillance. Proceedings of ICIAR **2** 2004 463–470

[7] Xu, L-Q., Landabaso, J.L. Pardàs, M.: Shadow removal with blob-based morphological reconstruction for error correction. Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on, **Vol.2**, Iss., March 18-23, 2005 729–732

[8] Landabaso, J.L., Pardàs, M., Xu, L-Q.: Hierarchical representation of scenes using activity information. Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on, **Vol.2**, Iss., March 18-23, 2005 677–680

[9] Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE trans. on Pattern Analysis and Machine Intelligence, **22(8)**, August 2000

[10] Bar-Shalom, Y., Fortman,T.E.: Tracking and Data association. Academic Press, 1988

[11] Sturim, D. E., Brandstein, M. S., Silverman, H. F.: Tracking Multiple Talkers Using Microphone-Array Measurements. Proceedings of ICASSP 1997, Munich, April 1997

[12] Hashemipour, H. R., Roy, S., Laub, J.: Decentralized structures for parallel Kalman filterin. IEEE Transactions on Automatic Control, 33(1):88-93, 1988

[13] The Spring 2006 CLEAR Evaluation and Workshop. http://www.clear-evaluation.org/