



APIDIS

Autonomous Production of Images based on Distributed and Intelligent Sensing

STREP Project, 1st FP7-216023

D4.4 Audio feature extraction

Due date of deliverable: 31-12-2008

Actual submission date: 09-01-2009

Start date of project: 1st January, 2008

Duration: 36 months

Lead contractor for this deliverable: QMUL

[Revision Final v1]

D4.4	Audio feature extraction
Project Acronym :	APIDIS
Contract No :	FP7-216023
Due Date :	31-12-2008
Reply To:	Fahad Daniyal fahad.daniyal@elec.qmul.ac.uk
Actual date of delivery	09-01-2009

1. Executive Summary

This document describes the audio analysis carried out in APIDIS. This audio analysis aims to recognize audio events, in particular the whistle of the referee, and to estimate arrival angle of a sound using multiple microphones.

For audio event detection, we study and employ multiple features: Zero Crossing Rate (ZCR), Short Time Energy (STE) and Mel-Frequency Cepstral Coefficients (MFCC) features.

The position of the source emitting a sound is estimated by computing the Generalized Cross Correlation function-Phase Transform (GCCF-PHAT). Moreover use of precedence effect and multi-band frequency analysis is proposed to reduce the effect of reverberation.

.

Deliverable Identification Sheet

Project ref. no.	FP7-216023
Project acronym	APIDIS
Project full title	FP7-216023
Security (distribution level)	Public (PU)
Contractual date of delivery	Month 12, December 31, 2008
Actual date of delivery	Month 12, December 31, 2008
Deliverable number	D4.4
Deliverable name	Audio feature extraction
Type	Report
Status & version	Final
Number of pages	15
WP / Task responsible	WP4 / QMUL
Other contributors	
Author(s)	Fahad Daniyal
EC Project Officer	Albert Gauthier
Abstract	This document describes the audio analysis carried out in the APIDIS project.
Keywords	Audio feature analysis, audio source localization, direction of arrival estimation
Sent to peer reviewer	ACIC
Peer review completed	Yes
Circulated to partners	Yes
Read by partners	Yes
Mgt. Board approval	Pending

Table of contents

1. Executive Summary	2
2. Introduction	5
3. Related Work	6
4. Acquisition settings.....	7
5. Audio Event Detection.....	8
5.1. Zero Crossing Rate (ZCR).....	8
5.2. Short Time Energy (STE)	8
5.3. Mel Frequency Cepstral Coefficients (MFCC)	8
5.4. Audio Event Classification	9
6. Audio Source Localization.....	10
6.1. Audio detection.....	10
7. Conclusions.....	13
8. References.....	14

2. Introduction

The use of multiple sensors and modalities for feature extraction is of great interest for many applications, ranging from surveillance to sports scenarios. In audiovisual fusion, audio sensors can overcome some of the limitations of visual sensors, such as bad lighting and visual occlusions. Furthermore, audio can be used to detect several events such as crowd cheering and whistling. These limitations make a network of heterogeneous sensors consisting of cameras and microphones a desirable solution in the APIDIS framework. Each sensor in such a network can be a Stereo Audio and Cycloptic Vision (STAC) sensor ([28]) consisting of a camera mounted between a pair of microphones (see Figure 1).



Figure 1: Stereo Audio and Cycloptic Vision (STAC) sensor.

The audio-visual sensor networks (with camera and microphone arrays) have been used in a variety of sensor configurations. Figure 2 shows a summary of these configurations, which range from a single microphone-camera pair to single or stereo cameras with stereo, circular arrays or linear arrays of microphones.

Camera-microphone pairs are used for speaker detection in environments with limited reverberation under the assumption that the speaker is facing the microphone [14], [8]; single or stereo cameras with multiple microphones are used in meeting rooms and teleconferencing [9]. In particular, STAC sensors are suitable for wide area monitoring, are used to perform audio-visual tracking with a probabilistic graph model and fusion by linear mapping [1] or with particle filters (PF) [17].

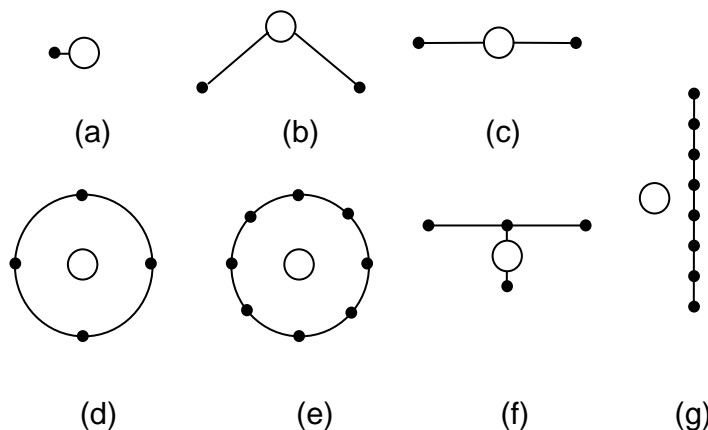


Figure 2: Examples of sensor configurations for audio-visual analysis (filled circles indicate microphones; empty circles indicate cameras – single or stereo): (a) single microphone-camera pair; (b-c) STAC sensors; (d-e) circular microphone array with single camera; (f) triangular microphone array with single camera; (g) linear microphone array with single camera

3. Related Work

The detection of sport audio events has recently gained much attention from the research community.

In [31] short-time energy, MFCCs and signal entropy are employed for speech detection and ball-hit sounds detection are used in a baseball scenario.

In [32] a Support Vector Machine was employed to train sound recognizers (applause, speech and whistles). It was assumed that those sounds are closely related to some events under specific sport game rules.

After detection of a target the next step is to estimate its direction. This estimation of the direction of arrival can be done by using audio only [15], [23], [25], [26] or by using audio and video simultaneously [3], [4], [5], [6], [7], [11], [14], [24], [29]. A summary of state-of-the-art algorithms is presented in Table 1.

Table 1: Multimodal tracking algorithms. (Key: PF=Particle Filter, KF=Kalman Filter, DKF=Decentralized KF, LDA=Linear Discriminant Analysis, TDNN=Time Delay Neural Networks, GM= Graphical Models, MFA=Multi-Feature Analysis, HCI=Human Computer Interaction).

Sensor types	Algo.	Application	Ref.
Stereo camera and circular microphone array	PF	Multi-modal user interface systems	[9]
2 cameras and 4 microphone arrays	PF	Indoor multiple person tracking	[6]
Camera and 10 element uniform circular microphone array	PF	Outdoor surveillance	[5]
Panoramic camera and 4 omni-microphones	MFA	Face detection	[14]
Wide-angle camera and a microphone array	I-PF	Meeting rooms	[10]
PTZ camera and 2 microphones	PF	Teleconferencing	[18]
Camera and a microphone	TDNN	Lip reading, HCI	[8]
Camera and 2 microphones	PF	Surveillance and teleconferencing	[3], [27], [17], [24], [2]
	GM	Indoor environment	[1]
	TDNN	Surveillance	[30]
Multiple camera and microphone arrays	KF, DKF	Smart rooms	[13], [20],
	LDA	Smart rooms	[19]
	PF	Multimodal meeting room	[12], [11], [4], [29], [16]

4. Acquisition settings

As the first multi-camera video data acquisition campaign was done for video only, the first audio data acquisition campaign was performed in a lab environment (see section 3 of D3.1). The purpose of this acquisition was to generate sample scenarios that are similar to the one in a basketball court in order to gain an understanding of the type of audio data and the challenges it may produce.

This acquisition campaign provides us with the data containing desired events to be analysed and on which the algorithms can be developed for later testing on real data.

The main event that is of interest in this data set is the detection of the referee's whistle. For this purpose we extract multiple features from the audio signals, as described in the following section.

5. Audio Event Detection

To detect the referee's whistle, we have extracted and analyzed multiple features from the audio signal being received by each microphone. Out of these features we selected the three most discriminative: Zero Crossing Rate (ZCR), Short Time Energy (STE) and Mel-Frequency Cepstral Coefficients (MFCC) features. In the following section we describe the selected features.

5.1. Zero Crossing Rate (ZCR)

The Zero Crossing Rate is the rate of sign-changes of a signal. The rate at which zero crossings occur is a simple measure of the frequency content of a signal and it is calculated as

$$ZCR = \frac{1}{T} \sum_{t=1}^{T-1} \text{sign}(y(t) \cdot y(t-1)) ,$$

where y is a signal of length T and the indicator function $\text{sign}(A)$ is the algebraic sign of its argument A .

5.2. Short Time Energy (STE)

The short time energy is the mean square of samples in each frame which is weighted with a Hamming window $h(n)$ and it is calculated as

$$STE = \frac{1}{T} \sum_{t=0}^{T-1} \|y(t)h(T-t)\|^2 ,$$

where y is a signal of length T .

Figure 3, shows the extracted short time energy, for both the whistle signal and a non-whistle signal. The non-whistle signal in this case consist of speech, shouting and silent segments.

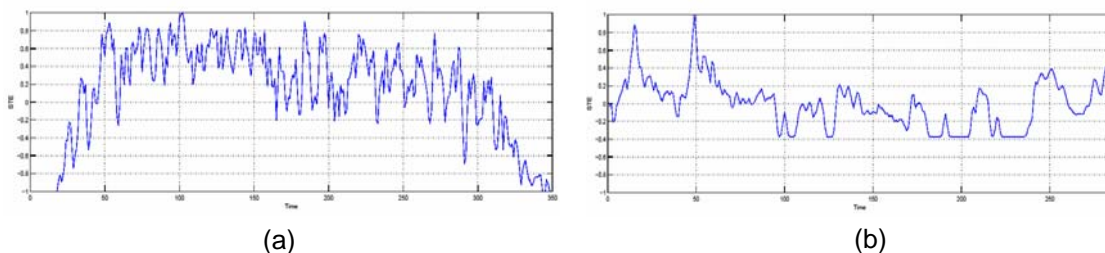


Figure 3- Comparison of short time energies between (a) whistle signal and (b) non-whistle signal

5.3. Mel Frequency Cepstral Coefficients (MFCC)

The Mel-Frequency Cepstral Coefficients are proved to be effective in speech recognition and modelling the subjective pitch and frequency content of audio signals. The frequency bands are positioned logarithmically (on the Mel scale)

and approximate the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT.

The MFCCs are computed from the FFT power coefficients that are filtered by a triangular pass filter bank as follows:

$$C_n = \sqrt{\frac{2}{k} \sum_{k=1}^K (\log Y_k) \cos \left(n (k - 0.5) \frac{\pi}{k} \right)},$$

where Y_k is the output of the k^{th} filter bank. and $n=1, \dots, N$ with N is the number of MFCCs dimensions. The temporal variation and acceleration of MFCCs are also used in our experiments.

5.4. Audio Event Classification

We segment the original audio signals into 40ms per frame as the basic unit (25 frames per second at 44100 HZ). Each frame is described by its observation of the low-level features extracted. The features of one frame are first normalized and combined into a vector. Then we use Subspace LDA to classify the samples into three classes: whistle; crowd uproar; and normal segments. In the dataset that was acquired the crowd uproar was classified when there was speech from more than one target. Normal segments were labelled when no audio events of interest occurred.

To evaluate the accuracy of the proposed algorithm we use precision and recall. Precision is regarded as a measure of exactness, whereas recall is a measure of completeness and are given as

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}.$$

The precision and recall for the event class (whistle) on sample test sequences is given in Table 2. The results for whistle detection performed on the in-house data set are encouraging. However the algorithm failed to recognize the audio signal when the whistle event was of a very short duration. This limitation is due to the fact that the energy ratio within the whistle frequency range might not be the peak due to the other dominant sounds or noises. A similar consideration is valid when a whistle event occurs while another whistle event is in progress.

Table 2 Whistle detection using multiple features

	True Positive	False Positive	False Negative	Precision (%)	Recall (%)
Test Sequence 1	33	5	4	86.84	89.18
Test Sequence 2	23	8	4	74.19	85.18

6. Audio Source Localization

We have studied and developed algorithms for audiovisual localization and tracking [3], [28], [22]. Audio measurements are derived from a multi-band generalized cross correlation that is used for audio source localization. To improve the localization accuracy, we employ reverberation filtering based on onset detection. The proposed algorithm is capable of detecting and localizing an active audio source.

6.1. Audio detection

Let the site be monitored by a set $\mathbf{M} = \{\mathbf{M}_1, \dots, \mathbf{M}_N\}$ of N microphone pairs, where $\mathbf{M}_i = (M_{i1}, M_{i2})$. We assume that the sound field of multiple microphone pairs \mathbf{M}_i overlap each other. Let each target generate a sound which is received at the microphones after a certain attenuation and delay. Let $y(t)$ be a sound wave generated by the source containing f_s / n_v samples, where f_s is the audio sampling frequency and n_v is the number of video frames per second.

This signal reaches the Stereo Audio Cycloptic Vision (STAC) sensor (Figure 1), at a certain arrival angle θ . It is assumed that the source is at a distance from the STAC sensor, thus by the time the sound waves reach the microphones they can be considered parallel.

Let the audio signals received at two microphones be defined as

$$\begin{aligned}\hat{y}_1(t) &= \gamma_1 y(t+n) + W_1 \\ \hat{y}_2(t) &= \gamma_2 y(t+n+\tau) + W_2,\end{aligned}$$

where γ_1 and γ_2 are the attenuation factors; n is the delay, in samples, occurred for the signal to reach the first microphone M_{i1} ; τ is the extra delay, in samples, for the signal to reach the second microphone M_{i2} and W_1, W_2 are the process noise added to the signal which is assumed to be zero mean Gaussian with unit variance.

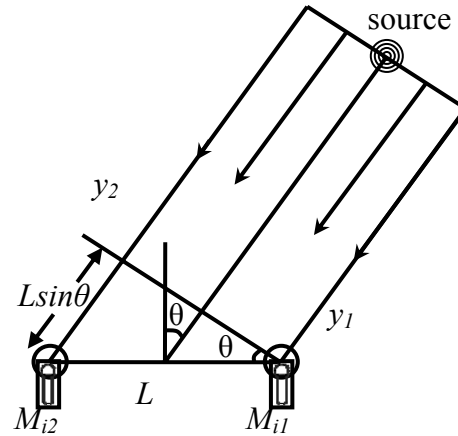


Figure 4: Source-receiver geometry for a STAC sensor in the far field. The distance between the microphones M_{i1} and M_{i2} is denoted by L and the arrival angle by θ . The sound wave has to travel an additional distance of $L \sin \theta$ to reach microphone M_{i2} .

The audio signals received at a microphone couple (M_{i1}, M_{i2}) at each time step (Figure 5) are used to compute the time difference of arrival (TDOA) τ of the audio signal for estimation of arrival angle θ for target localization.

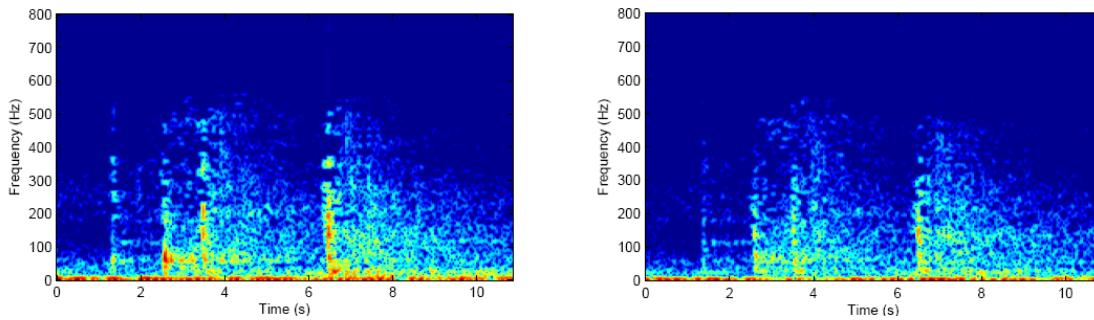


Figure 5: Sample spectrograms from (a) microphone 1 (M_{i1}) and (b) microphone 2 (M_{i2}). The signal at M_{i2} is delayed and attenuated.

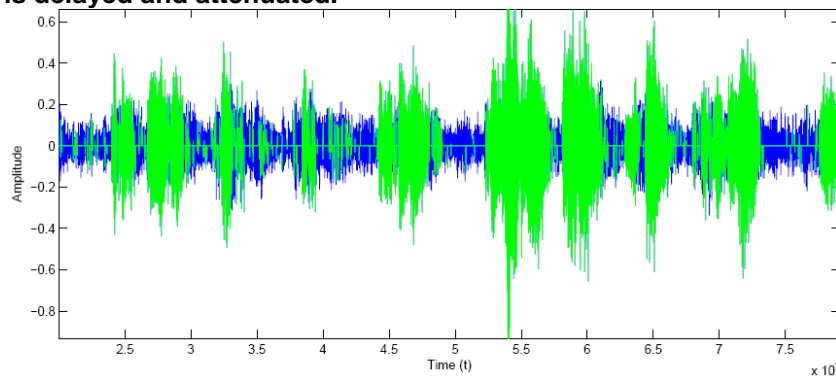


Figure 6: Filtering results on an audio signal. The original signal is shown in blue and the filtered signal is shown in green.

The position x of the sound source can be estimated by computing the cross-correlation $\bar{R}(y_1 y_2)$ of y_1 and y_2 using the Generalized Cross Correlation function-Phase Transform (GCCF-PHAT). To reduce the effect of reverberation in the source localization process, we exploit the precedence effect and Multi-Band

Frequency Analysis. The GCCF-PHAT $\bar{R}(y_1, y_2)$ is estimated only on ensemble of frames that are classified as onset $F_o(t)$ using the *precedence effect*.

Onset frames $F_o(t)$ are frames containing a significant signal component and a limited or absent reverberation component caused by the signal itself. These onsets are located at the beginning of a signal audio block (the audio segment between two salient segments of the audio signal). A frame $F(t)$ is considered a signal frame if the SNR at both microphones is larger than a threshold. Assuming that the frame under analysis is the first frame of an onset $F_o(1)$, the subsequent T frames are processed if identified as signal frames; whereas the signal frames from $F(t+T)$ to the first *null* frame are considered reverberant frames and therefore discarded.

The *multi-band frequency analysis* is based on the observation that low frequencies are less subject to reverberation than high frequencies and that the effects of correlated noise, located in a single frequency band, can be reduced by evaluating the signal in different frequency bands [21]. The two audio signals y_1 and y_2 are divided into three different frequency bands: a low frequency band (B_1), a middle frequency band (B_2), and a high frequency band (B_3). The frequency band division is computed using three different 36-coefficient band-pass linear phase FIR filters, frame-by-frame, for onset frames. The cross-correlation function is then estimated for each frequency band.

The final estimation of the GCC is obtained by a weighted combination of the three sub-band cross-correlations. The weights are chosen such that higher frequency components contribute less than the low frequency ones.

A peak is retained if it is simultaneously located in the same position in the three GCCs. Peaks that appear in a single band only are reduced proportional to the weight associated. The resulting improvements compared to the plain GCCF-PHAT can be seen in Figure 7. The green line shows the distance between the ground truth and the results obtained with the proposed approach. The blue line shows the distance between the ground truth and the GCC-PHAT result. It can be seen that error for the proposed system (green line) is much smaller than that of the GCC-PHAT (blue line).

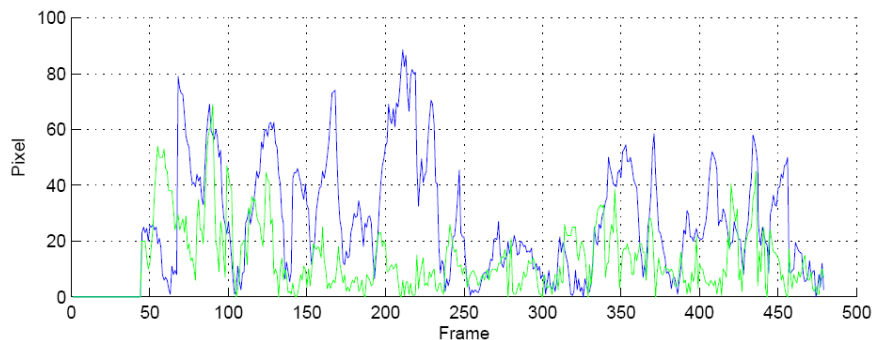


Figure 7: Deviation from the ground truth for source localization using the GCC-PHAT transform (blue) and the proposed method (green)

7. Conclusions

The sports scenario contains several interesting situations which can either be identified by an action from the referee such as a whistle or by cheering of the crowd. Such changes in audio activity can be detected and localized to assist in identifying the best view. To this end, based on a in-lab data acquisition campaign as a proof of concept, we demonstrate audio event detection for sport events and direction of arrival estimation on this dataset.

8. References

- [1]. M. J. Beal, H. Attias, and N. Jovic. Audio-video sensor fusion with probabilistic graphical models. In Proc. of the European Conf. on Computer Vision, Copenhagen, Denmark, June 2002.
 - [2]. A. Blake, M. Gangnet, P. Perez, and J. Vermaak. Integrated tracking with vision and sound. In Proc. of IEEE Int. Conf. on Image Analysis and Processing, volume 1, September 2001.
 - [3]. M. Bregonzio, M. Taj, and A. Cavallaro. Multi-modal particle filtering tracking using appearance, motion and audio likelihoods. In Proc. of IEEE Int. Conf. on Image Processing, San Antonio, TX (USA), September 2007.
 - [4]. R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia. A generative approach to audio-visual person tracking. In CLEAR, Southampton, UK, April 2006.
 - [5]. V. Cevher, A. C. Sankaranarayanan, J. H. McClellan, and R. Chellappa. Target tracking using a joint acoustic video system. IEEE Trans. on Multimedia, 9(4):715–727, June 2007.
 - [6]. N. Checka, K.W. Wilson, and M.R. Siracusa T. Darrell. Multiple person and speaker activity tracking with a particle filter. In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, volume 5, Cambridge, MA, USA, May 2004.
 - [7]. Y. Chen and Y. Rui. Speaker tracking using particle filter sensor fusion. Proceedings of the IEEE, 92(3):485–494, 2004.
 - [8]. R. Cutler and L. S. Davis. Look who's talking: Speaker detection using video and audio correlation. In Proc. of IEEE Int. Conf. on Multimedia and Expo (III), New York, NY USA, July-August 2000.
 - [9]. H. Asoh et al. An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion. In Proc. of the Seventh Int. Conf. on Information Fusion, Stockholm, Sweden, June 2004.
 - [10]. D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez, and D. Moore. Audio-visual speaker tracking with importance particle filters. In Proc. of IEEE Int. Conf. on Image Processing, 2003.
 - [11]. D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. IEEE Trans. on Audio, Speech, and Language Processing, 15:601–616, 2007.
 - [12]. D. Gatica-Perez, G. Lathoud, J.M. Odobez, and I. McCowan. Audio-visual probabilistic tracking of multiple speakers in meetings. IEEE Trans. Audio, Speech and Language Processing, March 2006.
 - [13]. T. Gehrig, K. Nicel, H. K. Ekenel, U. Klee, and J. McDonough. Kalman filters for audio-video source localization. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New York, NY, USA, October 2005.
 - [14]. B. Kapralos, M. Jenkin, and E. Milios. Audio-visual localization of multiple speakers in a video teleconferencing setting. Int. Journal of Imaging Systems and Technology, 13(1):95–105, June 2003.
 - [15]. E.A. Lehmann, D.B. Ward, and R.C. Williamson. Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room. In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pages 177–180, Hong Kong, 6–10 April 2003.
 - [16]. K. Nickel, T. Gehrig, H.K. Ekenel, J. McDonough, and R. Stiefelhagen. An audio-visual particle filter for speaker tracking on the CLEAR06 evaluation dataset. In CLEAR, Southampton, UK, April 2006.
 - [17]. P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. Proc. of IEEE, 92:495–513, March 2004.
 - [18]. Y. Rui and Y. Chen. Better proposal distributions: object tracking using unscented particle filter. In Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pages 786–793, HI, USA, December 2001.
 - [19]. A. Stergiou, A. Pnevmatikakis, and L. Polymenakos. A decision fusion system across time and classifiers for audio-visual person identification. In CLEAR, Southampton, UK, April 2006.
-

- [20]. N. Strobel, S. Spors, and R. Rabenstein. Joint audio-video object localization and tracking. *IEEE Signal Processing Magazine*, 18(1):22–31, January 2001.
 - [21]. T. Sullivan and R. Stern. Multi-microphone correlation-based processing for robust speech recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 91–94, 1993.
 - [22]. M. Taj and A. Cavallaro. Audio-assisted trajectory estimation in non-overlapping multi-camera networks. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 19–24 April 2009.
 - [23]. J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 3021–3024, Salt Lake City, Utah, May 2001.
 - [24]. J. Vermaak, M. Gangnet, A. Blake, and P. Pérez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Proc. of IEEE Int. Conf. on Computer Vision*, pages 741–746, Vancouver, Canada, July 2001.
 - [25]. D. B. Ward, E. A. Lehmann, and R. C. Williamson. Particle filtering algorithms for acoustic source localization. *IEEE Trans. on Speech and Audio Processing*, pages 826–836, November 2003.
 - [26]. D. B. Ward and R. C. Williamson. Particle filter beamforming for acoustic source localisation in a reverberant environment. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 1777–1780, Orlando, FL, USA, May 2002.
 - [27]. H. Zhou, M. Taj, and A. Cavallaro. Audiovisual tracking using STAC sensors. In *ACM/IEEE Int. Conf. on Distributed Smart Cameras*, Vienna, Austria, September 25-28 2007.
 - [28]. H. Zhou, M. Taj, and A. Cavallaro. Target detection and tracking with heterogeneous sensors. *IEEE Journal of Selected Topics In Signal Processing (J-STSP)*, 2(4), 2008.
 - [29]. D. N. Zotkin, R. Duraiswami, and L. S. Davis. Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, 2002:1154–1164, 2001.
 - [30]. X. Zou and B. Bhanu. Tracking humans using multi-modal fusion. In *IEEE Int. Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum (OTCBVS)*, San Diego, CA, USA, June 2005.
 - [31]. Y. Rui, A. Gupta, and A. Acero, “Automatically extracting highlights for TV baseball programs,” In *Proc. of ACM, Multimedia*, Los Angeles, CA, (2000) 105-115
 - [32]. M. Xu, N. C. Maddage, C. S. Xu, M. Kankanhalli, and Q. Tian, “Creating audio keywords for event detection in soccer video,” in *Proc. of International Conference on Multimedia and Expo*. (2003) 6-9
-