# VIDEO TRACKING THROUGH OCCLUSIONS BY FAST AUDIO SOURCE LOCALISATION

*Eleonora D'Arca, Ashley Hughes, Neil M. Robertson and James Hopgood*

Joint Research Institute for Signal and Image Processing,
Heriot-Watt University & University of Edinburgh, UK
visionlab.eps.hw.ac.uk

## ABSTRACT

In this paper we present a novel audio-visual speaker detection and localisation algorithm. Audio source position estimates are computed by a novel stochastic region contraction (SRC) audio search algorithm for accurate speaker localisation. This audio search algorithm is aided by available video information (stochastic region contraction with height estimation (SRC-HE)) which estimates head heights over the whole scene and gives a speed improvement of $56\%$ over SRC. We finally combine audio and video data in a Kalman filter (KF) which fuses person-position likelihoods and tracks the speaker. Our system is composed of a single video camera and 16 microphones. We validate the approach on the problem of video occlusion i.e. two people having a conversation have to be detected and localised at a distance (as in surveillance scenarios vs. enclosed meeting rooms). We show video occlusion can be resolved and speakers can be correctly detected/localised in real data. Moreover, SRC-HE based joint audio-video (AV) speaker tracking outperforms the one based on the original SRC by $16\%$ and $4\%$ in terms of multi object tracking precision (MOTP) and multi object tracking accuracy (MOTA). Speaker change detection improves by $11\%$ over SRC.

***Index Terms*—** Video Tracking, Speaker Tracking, Multimodal tracking, Optimization methods, Sampling Methods

## 1. INTRODUCTION

Solving visual tracking occlusion is inherently challenging when only video information is available. Many existing papers solve the problem by using sophisticated multi-camera 3-dimensional (3D) systems [1] which are still prone to occlusions when the camera fields-of-view do not overlap. Moreover, they are computationally expensive, often requiring GPU/FPGA implementations to function at frame-rate. Thus, supporting tracking with non-visual information, i.e. audio, may compensate for noisy, missing and erroneous video data via speaker detection info, reducing the number of cameras and the computational resources required at the expense of a few microphones. Video and audio "fusion" (or combination) can be achieved in several ways mostly using variations of sampling techniques [2–4]. Existing system architectures work well in very sanitised scenarios e.g. meeting analysis and diarisation [5–9]. They use large sensor networks composed at least of 4 cameras and 16 microphones [3, 4, 6, 8]. Little attention has been focussed on uncontrolled (and larger) areas of interest using smaller and less "invasive" sensor networks. Attention in the literature is principally focussed on general event detection [10–12], rather than on people interactions and behaviour analysis [13, 14]. The novel system we present can localise and recognise a speaker among two people in an ample, reverberant and noisy environment when large video occlusion occur using a small sensor network. To the best of our knowledge this work is similar to the ones from [4, 15]. In contrast, we improve on the state-of-the-art via: *a*) new, high accuracy, fast audio localisation algorithm; *b*) real-time video localisation and tracking using particle filter (PF) [1]; *c*) improved precision and accuracy metrics for multi object tracking (2006 and 2007 CLEAR dataset [16]).

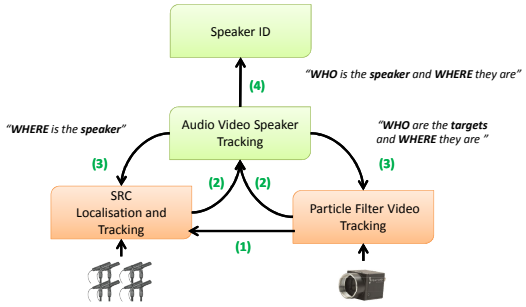


**Fig. 1**: A schematic of the system presented in this paper. Constituent parts of this diagram are referred to explicitly in the text (e.g. "arrow 1").

## 2. THEORY

A schematic diagram of our system is shown in Figure 1. In the following sections we describe it in detail.

NR and JH are supported by EC FP7 LOCOBOT (Grant EC/260101). AH is supported by scholarships from the James Clerk Maxwell Foundation and the Maxwell Advanced Technology Fund.

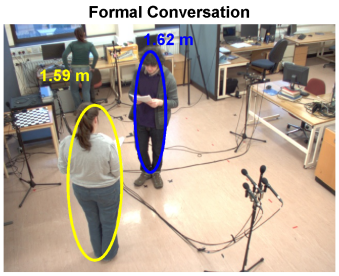


**Fig. 2**: Video detected height data are novelly used to reduce the search of space for audio source localisation SRC.

### 2.1. Height detection and video tracking

Full details of the video tracker based on a GPU-accelerated particle filter with ellipsoid models for people can be found in [1]. It is worth noting that we hereby use the video data coming from only 1 camera view. Height measurement is also extracted (Figure 2) to cue the audio localisation algorithm, since it directly corresponds to a good estimate of the speaker's head position.

### 2.2. Audio source localisation

A popular method of audio source tracking is extracting maximal time difference of arrival (TDOA) values from the generalised cross correlation with phase transform (GCC-PHAT) [17] of signals from a pair of microphones in the frequency domain, given by Equation (1), which is an inverse Fourier transform where $\hat{G}_{x_m x_n}$ is the product of the signals $x_m$ and $x_n$ in the frequency domain.

$$\hat{R}_{x_m x_n}(\tau) = \int_{-\infty}^{\infty} \frac{\hat{G}_{x_m x_n}(f)}{|G_{x_m x_n}(f)|} e^{2\pi f \imath \tau}\, df \qquad (1)$$

A method more robust to reverberation, the steered response power (SRP), makes use of the GCC-PHAT to build an energy map using Equation (2) in a system with $M$ microphones. This is the sum over all pairs $(m,\, n)$ of microphones of the corresponding value of the GCC-PHAT for the TDOA $\tau$.

$$S(x, y, z) = \sum_{n=1}^{M} \sum_{m=n+1}^{M} \hat{R}_{x_n\, x_m}[\tau_{n\, m}(x, y, z)] \qquad (2)$$

The TDOA is defined by Equation (3), where $\mathbf{p}$ is the vector $x,\, y,\, z$ of the point under investigation, $c$ is the speed of sound, and $\mathbf{m}$ and $\mathbf{n}$ are the positions of microphones $m$ and $n$ respectively.

$$\tau_{n\, m}(\mathbf{p}) = \left(|\mathbf{m} - \mathbf{p}| - |\mathbf{n} - \mathbf{p}|\right) / c \qquad (3)$$

Evaluating the SRP across an entire room is computationally costly. In this work we use an enhanced version of the SRC [18] algorithm to localise quicker and better an audio source. This works by sampling the SRP randomly and choosing a subset of the largest samples to form a new region to sample within. This is repeated until the process has discovered a maximum. In order to further improve upon the SRC, instead of sampling uniformly over height, a different sampling distribution is used, centred around a head height. To choose head height, existing knowledge of the current positions and heights of people in a room which is obtained from the camera (Figure 2), is novelly used (SRC-HE). In particular, the height data is updated on each iteration to the height of the last SRP peak found. This reduction of the search space decreases its effective dimensionality, thereby decreasing the computational complexity of SRC.

From a sparse set of people, the head height at every $x$-$y$ co-ordinate in the SRP map needs to be defined. This is achieved using interpolation and extrapolation. When doing the interpolation, there is a trade-off between the smoothness of the curve produced and the size of ripples produced. The interpolation should not contain severe ripples as they would lead to large errors in the head height estimation across the room. Ideally, it should be monotonic and one way to achieve this is to use Delaunay triangulation [19] on the set of speakers, which creates a surface which can be evaluated at any 2-dimensional (2D) point.

The height $h_{sub}$ to use at each time step for every point $\mathbf{p} = (x,\, y)$ is then drawn from 4, which mixes a Gaussian with a Uniform distribution across $h_r$, the entire height of the room.

$$\begin{aligned} p\,(z \mid \mathbf{p}) &= \alpha_0 \mathcal{N}\left(\mu_h,\, \sigma_h^2\right) + (1 - \alpha_0)\, \mathcal{U}\,(0,\, h_r) \\ \mu_h &= \mathbf{H}[\mathbf{p}] \\ \sigma_h^2 &= \hat{q}(\mathbf{p},\, \mathbb{T}) \end{aligned} \qquad (4)$$

Around each person, we can be relatively confident of their height. Further away from them, the decreasing confidence is modelled by increasing the variance of the sampling probability density function (PDF). The variance at a distance $l$ metres from a speaker is chosen to be modelled by a sigmoid function, q, such as Equation (5), which is a scaled error function. This is 0 at the origin and asymptotically approaches a constant as its argument tends towards infinity.

$$\mathrm{q}(l) = \alpha_1 \operatorname{erf}(\alpha_2 l) \qquad (5)$$

These need to be combined to form a global variance. At any point $\mathbf{p}$ in space, the appropriate variance $\hat{q}$ to use will be the sigmoid function q of the minimum of the set of all 2D Euclidian distances $\overline{\mathbf{pq}}$ to known sources, where the set of known source locations is denoted as $\mathbb{T}$ and an element from the set of know sources is denoted as $\mathbf{q}$. This is expressed in Equation (6). The minimum is chosen to ensure that the change in variance remains smooth even for overlapping sigmoids from multiple sources.

$$\mathbb{L}_{\mathbf{p},\mathbb{T}} = \{l : (\exists \mathbf{q} \in \mathbb{T})(l = \overline{\mathbf{pq}})\}$$
$$\hat{q}_{\mathbf{p},\mathbb{T}} = \min_{l \in \mathbb{L}_{\mathbf{p},\mathbb{T}}} q(l) \tag{6}$$

### 2.3. Joint Audio-Video Speaker Tracking

SRC-HE algorithm allows for direct speaker position calculation, $\mathbf{x}$. Nevertheless, speaker position estimations are characterised by missing and false detections. This is mostly due to speech pauses and room reverberation respectively. Thus, we filter SRC estimated positions $\mathbf{x_a}$ by a KF. We said already that, to speed up SRC searching time, speaker's height computed by the video PF, is input into the audio unit to drive height sampling (arrow 1, Figure 1). Then, after the audio and video data have been aligned, the posteriors of the KF audio tracker and of the PF $\mathbf{x}_a$ and $\mathbf{x}_v$ are fused in a common KF node (arrow 2, Figure 1). As data are gathered simultaneously and used all at once in a centralised fashion, we assume the audio and video *pdfs* to be independent of one another thus, on the basis of the *a priori* local estimates for the state $\mathbf{x}_a(t|t-1)$ and $\mathbf{x}_v(t|t-1)$ predicted by the single-modality trackers at each time step $t$, we evaluate the joint state estimate $\mathbf{x}_{av}$ as follows (where time dependency has been omitted for clarity):

$$\mathbf{p}(\mathbf{z}_{av} \mid \mathbf{x}) = \mathbf{p}(\mathbf{z}_a \mid \mathbf{x})\mathbf{p}(\mathbf{z}_v \mid \mathbf{x}); \tag{7}$$

this means the joint likelihood is still a Gaussian probability, although no longer normalised, and the *a posteriori* state estimate is given by:

$$\mathbf{x}_{av} = \quad \mathbf{P}_{av}\left\{\mathbf{P}_a^{-1}\mathbf{x}_a + \mathbf{P}_v^{-1}\mathbf{x}_v\right\}, \tag{8}$$

where

$$\mathbf{P}_{av} = (\mathbf{P}_a^{-1} + \mathbf{P}_v^{-1})^{-1}. \tag{9}$$

$\mathbf{P}_a^{-1}$ and $\mathbf{P}_v^{-1}$ are the inverse of the audio and video *a posteriori* covariance estimation matrices. $\mathbf{P}_{av}$ is the joint *a posteriori* covariance estimation matrix. Finally, the last joint AV output $\mathbf{x}_{av} = \mathbf{P}_a\mathbf{x}_a + \mathbf{P}_v\mathbf{x}_v$ is fed back into the individual audio and video trackers as the best estimate of the previous time step to improve the single modality estimation (arrow 3, Figure 1). It is important to notice that, as we make the assumption that people speak alternatively, like in a normal conversational mode, to a single audio signal $\mathbf{z}_a$, correspond several video measurements $\mathbf{z}_{v_i}$ at a time, one for each of the $N$ detected targets. By basing the audio-to-video data association step on spatial proximity, i.e. nearest neighbour (NN), speaker segmentation and recognition can also be obtained as long as people are resolved by the AV tracker and its measurements can be considered robust with respect to the speaker motion model. In particular, the speaker identity inferred by the joint tracker is equal to the one of the $i$-th target if $\mathbf{S}_{av} = \arg\max_i \left\{\mathbf{p}(\mathbf{z}_a, \mathbf{z}_{v_i} \mid \mathbf{x})\right\}, i = 1, ..., N$ (arrow 4, Figure 1). Saying that, once an identity $i$ has been assigned to every target in an image frame, the speaker change detection output by the audio unit is used in order to recover identity (ID) tracking when occlusions occur. In particular, in case audio and video inference about the detected number of targets in the scene is conflicting, or when audio and video data do not both fall within a certain region ($\|\mathbf{x}_a - \mathbf{x}_v\| \leq A$), audio source position is considered to be correct and it is also sent back to the video tracking unit to indirectly re-assign the correct appearance models to the targets, successfully resolving occlusions (arrow 3, Figure 1).

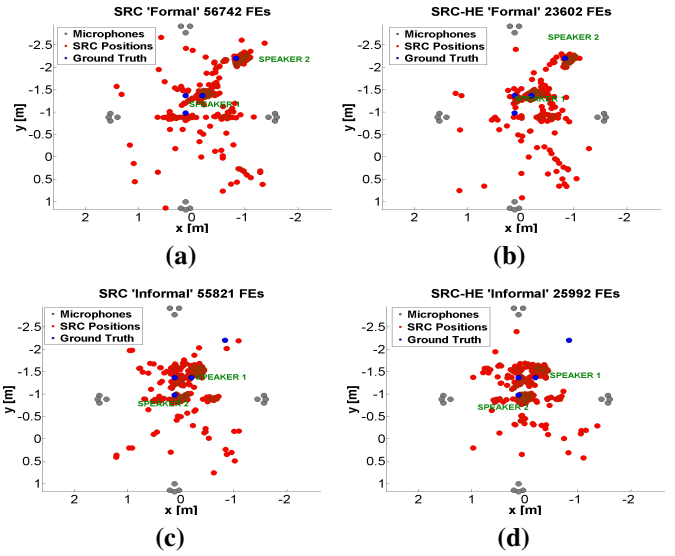


**Fig. 3**: SRC and SRC-HE raw speaker position detections. Interesting is the number of FEs which on average is reduced by $56\%$ (FEs $56,281$ vs $24,797$) for the SRC-HE implementation. **(a)** and **(b)** show respectively audio source SRC and SRC-HE detections for the *'Formal'* experiment. While **(c)** and **(d)** show them for the *'Informal'* one.

| Experiment | System | SSL Accuracy (%) | FEs |
|---|---|---|---|
| *'Formal'* | **SRC** | 62.50 | 56742 |
| | **SRC-HE** | 69.07 | 23601 |
| *'Informal'* | **SRC** | 47.30 | 55821 |
| | **SRC-HE** | 51.22 | 25992 |

**Table 1**: SRC vs SRC-HE performance comparison for the two set of data (*'Formal'* and *'Informal'*). Results are shown for 2 off-line runnings of the two algorithms. SSL accuracy changes by $4\%$ when adding up extracted video height info. More interesting is the $56\%$ change in the number of FEs which has to be calculated, meaning that narrowing down the space of search effectively results in speeding up the localisation task.

## 3. EXPERIMENTATION AND RESULTS

In this section we show that SRC-HE outperforms original SRC using video data and that our global AV system can maintain and recover speaker ID. We used 1 camera and $4-by-4$ T-shape microphone arrays to record AV data in a typical open office room, whose size is $111.44\ m^2$, where the area considered of interest is $12\ m^2$ (as seen in Figure 4**(a)**). Ground-truth data was hand labelled to $5\ cm$ of accuracy, on a ground plane common to camera and microphones. Audio

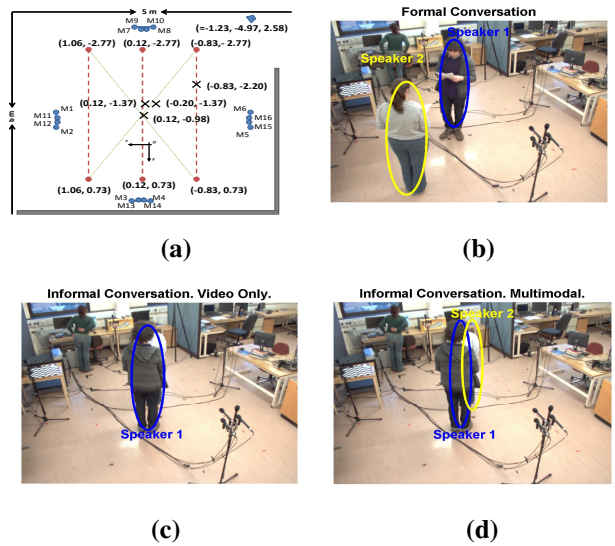


**Fig. 4**: Real experiments layout **(a)** and *'Formal'* and *'Informal'* visual results. In **(b)** a formal conversation between two people is shown. Video tracker, as well as multimodal tracker, can detect and recognise there are two targets speaking alternatively and their output is the same. **(c)** shows an informal conversation between two people. They are so close the video tracker on its own cannot detect there are two different targets. In **(d)** instead, the AV multimodal tracker is shown to detect the two speakers and successfully recognise their identity.

| Experiment | System | MOTP ($m$) | MOTA (%) | DER (%) |
|---|---|---|---|---|
| *'Formal'* | **SRC** | 0.35 | 85 | 21 |
| | **SRC-HE** | 0.34 | 90 | 7 |
| *'Informal'* | **SRC** | 0.20 | 97 | 20 |
| | **SRC-HE** | 0.12 | 100 | 11.80 |

**Table 2**: Experiment results. SRC AV tracker does not incorporate prior video height information while SRC-HE does.

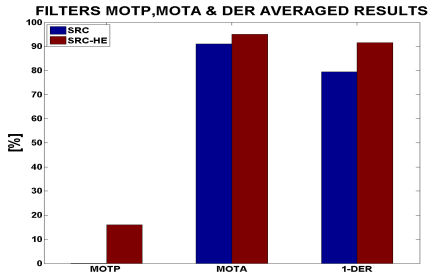


**Fig. 5**: SRC vs SRC-HE AV tracking averaged over both the experiments and 100 montecarlo runs performance comparison. SRC-HE detection accuracy improvement results in an AV tracker which outperforms SRC based AV tracker precision (MOTP) by 16% and accuracy by 4%. Of interest here, is that DER is also improved by 11%, which make this solution 11% better than SRC in handling large video occlusions. Note that the video tracker on its own instead can not resolve occlusion at all.

signals were sampled by the audio interface with a 24-bit precision resolution at $44.1\ kHz$, whereas the camera recorded the $640 \times 480$ RGB video frames at a $7.5\ Hz$ rate. Moreover, each audio signal was filtered using a $\approx 20\ ms$ long Gaussian window to ensure signal stationarity [20]. We made no attempt to reduce normal background noise (desk fans, footsteps, talking etc.) and a large reverberation time ($T_{60} \approx 0.5\ s$) was measured. Synchrony of data was insured by processing audio and video streams accordingly to the camera frame rate. Filters were initialised using the video detected position of their correspondent targets and static matrices Q and R [21], whose values were chosen on the basis of an optimisation step. We describe the results in terms of MOTP and MOTA [16]. We also calculate the diarisation error rate (DER), which measures the ability of detecting a change in speaker ID, expressing the speaker error only [22].

Experiments meant to simulate a personal (formal) and intimate (informal) conversation between two people, resulting in an occlusion in the case of informal conversation. Specifically:

***'Formal Conversation'***, considers two people having a $60\ s$ conversation. Throughout all the experiment they are separated by a distance of approximately $104\ cm$. Results as presented in Figure 4 **(b)**.

***'Informal Conversation'***, considers two people having a $56\ s$ conversation. Throughout all the experiment they are separated by a distance of approximately $40\ cm$. Results are shown in Figure 4 **(c)** and **(d)** .

Figure 3 demonstrates SRC vs SRC-HE raw speaker position detections for the two set of data (*'Formal'* and *'Informal'*). In Table 1 we enumerate their performance comparison. Results are shown in terms of SSL accuracy and number of FE calculations. In both cases, the results show a significant decrease in the number of FEs as well as an improvement in accuracy. Moreover, video only and SRC-HE based AV tracker outputs are shown in Figures 4 **(c)** and **(d)** for a comparison. Furthermore, in Table 2 we present MOTP, MOTA and detection error rate (DER) of the joint AV trackers based on SRC only and on SRC-HE. At last, their performance comparison is shown in Figure 5. Please note that, when we talk about SRC results we refer to an AV system as in Figure 1 where arrow 1 does not exist (no video cueing).

## 4. CONCLUSION AND FUTURE WORK

In this paper integrating height information coming from a video PF with a SRC SSL algorithm (SRC-HE), has been proved to speed up by 56% speaker detection based on the original SRC algorithm. Moreover, it has been shown that augmenting video tracking with audio data does solve large occlusion which otherwise would not be solved by the video tracker only. Furthermore, using audio data detected with SRC-HE improves by 16% and 4% AV speaker MOTP and MOTA tracking and by 11% AV speaker change detection, if compared to an AV tracker which uses the original SRC implementation. In future, we would like to carry out a tighter integration between audio and video using updated height information from every frame to investigate further improvements on SRC-HE. Furthermore, we would like to record datasets similar to other existing works to carry out a thorough comparison against state-of-the art joint AV systems in non-meeting rooms.

## 5. REFERENCES


[1] Wasit Limprasert, Andrew M. Wallace, and Greg Michaelson, "Accelerated people tracking using texture in a camera network," in *VISAPP (2)'12*, 2012, pp. 225–234.

[2] N. Checka, K.W. Wilson, M.R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," vol. 5, pp. V–881–4 vol.5, May 2004.

[3] Yeongseon Lee and R. Mersereau, "Data association for people tracking using multiple cameras," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 312008-april4 2008, pp. 2585 –2588.

[4] Huiyu Zhou, M. Taj, and A. Cavallaro, "Target detection and tracking with heterogeneous sensors," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 2, no. 4, pp. 503–513, Aug. 2008.

[5] Kai Nickel, Tobias Gehrig, Hazim Kemal Ekenel, John W. McDonough, and Rainer Stiefelhagen, "An audio-visual particle filter for speaker tracking on the clear'06 evaluation dataset," in *CLEAR*, 2006, pp. 69–80.

[6] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 601–616, Feb. 2007.

[7] Gerald Friedland, Chuohao Yeo, and Hayley Hung, "Visual speaker localization aided by acoustic models," in *Proceedings of the 17th ACM international conference on Multimedia*, New York, NY, USA, 2009, MM '09, pp. 195–202, ACM.

[8] Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *J. Sel. Topics Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[9] Xavier Alameda-Pineda, Vasil Khalidov, Radu Horaud, and Florence Forbes, "Finding audio-visual events in informal social gatherings," in *Proceedings of the 13th international conference on multimodal interfaces*, New York, NY, USA, 2011, ICMI '11, pp. 247–254, ACM.

[10] E. Kidron, Y.Y. Schechner, and M. Elad, "Pixels that sound," vol. 1, pp. 88–95 vol. 1, June 2005.

[11] Marco Cristani, Manuele Bicego, and Vittorio Murino, "Audio-visual event recognition in surveillance video sequences," *Multimedia, IEEE Transactions on*, vol. 9, no. 2, pp. 257–267, Feb. 2007.

[12] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *Multimedia, IEEE Transactions on*, vol. 15, no. 2, pp. 378 –390, feb. 2013.

[13] Maria Andersson, Stavros Ntalampiras, Todor Ganchev, Jrgen Ahlberg Rydell, and Nikos Fakotakis, "Fusion of acoustic and optical sensor data for automatic fight detection in urban environments," in *FUSION 2010*, 2010.

[14] M. Andersson and R. Johansson, "Multiple sensor fusion for effective abnormal behaviour detection in counter-piracy operations," in *Waterside Security Conference (WSS), 2010 International*, nov. 2010, pp. 1 –7.

[15] M. Bregonzio, M. Taj, and A. Cavallaro, "Multi-modal particle filtering tracking using appearance, motion and audio likelihoods," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 16 2007-oct. 19 2007, vol. 5, pp. V –33 –V –36.

[16] Keni Bernardin and Rainer Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *J. Image Video Process.*, vol. 2008, pp. 1:1–1:10, January 2008.

[17] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, January 2003.

[18] Hoang Do, H.F. Silverman, and Ying Yu, "A real-time srp-phat source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, april 2007, vol. 1, pp. I–121 –I–124.

[19] D. T. Lee and B. J. Schachter, "Two algorithms for constructing a delaunay triangulation," *International Journal of Parallel Programming*, vol. 9, no. 3, pp. 219–242, June 1980.

[20] Maurice Fallon, *Acoustic Source Tracking Using Sequential Monte Carlo*, Ph.D. thesis, Darwin College, University of Cambridge, September 2008.

[21] T. Gehrig, K. Nickel, H.K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," pp. 118–121, Oct. 2005.

[22] Jonathan G. Fiscus, Jerome Ajot, Martial Michel, and John S. Garofolo, *The Rich Transcription 2006 Spring Meeting Recognition Evaluation*, NIST, 2006.