A Surveillance System based on Audio and Video Sensory Agents cooperating with a Mobile Robot

Emanuele Menegatti [2], *Enzo Mumolo* [1], *Massimiliano Nolich* [1], *Enrico Pagello* [2,3]

[1] Smartlab Department of Electrotechnics, Electronics and Computer Science,
The University of Trieste, Italy
[2] Intelligent Autonomous Systems Laboratory (IAS-Lab)
Department of Information Engineering
The University of Padua, Italy
[3] *also with*: Institute ISIB of CNR, Padua, Italy

Abstract

We present a surveillance system that uses audio and video sensors to reveal and track the presence of an intruder in a off-limit area. The system is composed of a mobile agent and several static agents cooperating in the tracking task. The mobile agent is a vision agent composed of an omnidirectional vision system and a mobile robot. The mobile agents greatly increase the flexibility of the surveillance system with respect to traditional surveillance systems, e.g. the mobile robot can "go and seek" for intruders in areas occluded to the monitoring system or it can patrol areas outside of the field of view of the fix sensory agents. The static agents are acoustic agents composed of self-steerable microphone arrays and a vision agent implemented on a omnidirectional vision system. In this paper, the integration of pre-existing blocks is presented and preliminary tests on the system are reported

1 Introduction

In this paper, we present our current project on the development of an intelligent surveillance system that uses both mobile and fix surveillance agents. The scenario of application is the monitoring of a room or a multi room environment with a dynamic structure, e.g. the storage room of a shipping company where the position of piles of boxes can change day after day. In this case most of the traditional surveillance systems, e.g. [3] [5] based on static sensors will fail, because they will not be able to re-configure in order to avoid occlusions from objects piled-up in front of the sensors. In our system, one (or more) mobile robot can be sent to inspect suspicious areas occluded by moving objects. In this paper we focus on the integration of the visual and audio information provided by different "sensing agents". The concept of "sensing agent" is introduced to shade the lights on merely perceptual actions.

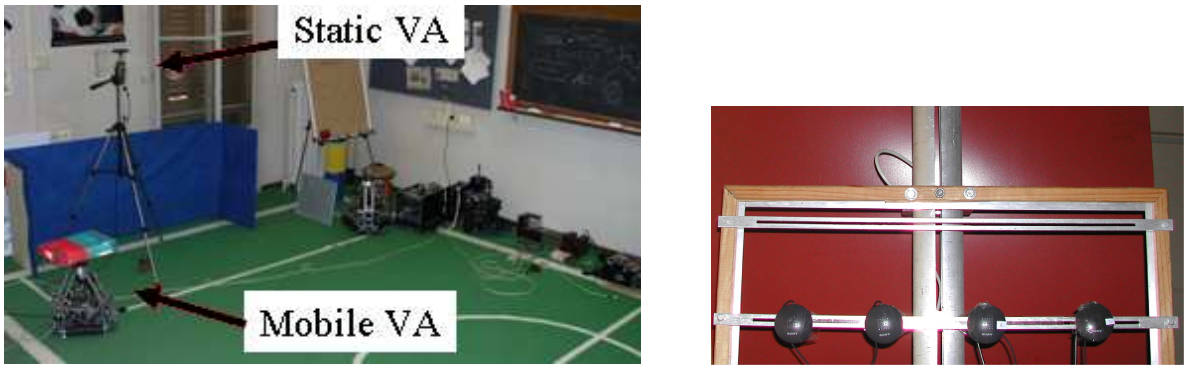

Figure 1: The sensory agent implemented: on the left, the two vision agents the static one on the tripod and the mobile one on the mobile robot; on the right, one of the acoustic agents present in the environment.

In our approach the sensors scattered in the environment cooperate in order to form a sort of "super-sensor" distributed among the robot team. This distributed sensor is used to provide the single mobile robot and the remote human supervisor of the system with richer information than the one coming from the single agent.

Many researchers have attempted to integrate vision and acoustic senses. Most of their implementations process each sense separately and integrate the overall results in the final step. The system described in [11] uses an array of eight microphones to initially locate a speaker and then to steer a camera towards the sound source. The camera does not participate in the localisation of objects. It is used simply to take images of the sound source after it has been localised. This system is well suited for video-conferences, but not for surveillance purposes. Our approach is more similar to the one described in [1], i.e. a multi-modal sound localisation system that uses two cameras and a 3-element microphone array. In this work, Aarabi and Zaki demonstrated that the localisation integrating audio and video information is more robust compared to localisation based on stand alone microphone arrays. Their approach seemed to be reliable only when using ad-hoc narrow band acoustic signals. In our work, we used generic broad band signals to locate the persons, like the person's step noise.

2 The system setup

As we said in the introduction, the system is composed of several sensors. The sensors are shown in Fig. 1. In the left picture are depicted: the static Vision Agent composed of an omnidirectional camera with an hyperbolic mirror and the mobile robot called mobile Vision Agent. In the right picture is imaged one of the audio sensors (Acoustic Agent) composed of an array of four microphones able to perform beam forming and to track a person by the noise of its steps.

Every sensory agent is implemented as a sensor (microphone or camera) connected to a computer fitted with a IEEE 802.11b wireless LAN card. The PC provides the agent the computational power necessary to process the raw sensory data and to transmit the results of this processing via the wireless LAN to a remote console, where an human operator can monitor the situation. The communications are managed by a middle-ware called ADE [2], we developed for the RoboCup project. Thanks to ADE, message passing from one agent to the other is totally transparent; independently if they reside on the same machine or on machines connected through a LAN or a wireless LAN.

The system is able to detect and track intruders in a indoor dynamic environment grabbing close-up images of the intruder with the mobile robot. The basic functioning of the system is:

- the static vision agent, i.e. the omnidirectional camera over the tripod, detects moving object in the image;
- the static vision agent communicates the position of the detected moving object to the mobile robot and to the acoustic agents;
- the acoustic agents perform beamforming in the direction of the detected motion, determine the characteristics of the foot steps of the intruder, locate him and start tracking it;
- the different measurement on the position of the intruder coming from the vision agent and the acoustic agent are fused in order to improve the position estimation;
- once the acoustic tracking started, the mobile robot move toward the position of the intruder obtained fusing all acoustic and visual information;
- once the intruder is detected by the mobile vision agent a close-up image is sent to the monitoring station to check if the moving object represents a danger or if it is just a false alarm;

Let us discuss the implementation of the single sections of the system.

3 The Static Vision Agent

As mentioned before, the static vision agent is composed of a catadioptric omnidirectional camera composed of a standard perspective camera and a hyperbolic mirror[1].

To detect the intruder the image is segmented into the moving foreground and into the stationary background. As we said, our system is conceived to work in a dynamic environment in which the objects and the obstacles might change configuration in time. For this reason we adopted an incremental background subtraction algorithm. In this technique the background image is not a static image, but it is updated frame after frame slowly incorporating changes in the scene.

In Fig. 2 is depicted a sequence in which the history image is changing to incorporate a person that entered the scene and was stationary for a long time.On the left image, the person is just a ghost in the image on the left of the door, in the middle image, the ghost of the person become more tangible, on the right image the person is merged into the background.

The incremental background is calculated according to Eq. 1. This is a grey-level image representing the fix luminance in the image.

$$\text{background}_t(i,j) = \text{background}_{t-1}(i,j) \cdot (1-\alpha) + \text{luminance}_t(i,j) \cdot \alpha \qquad (1)$$

The parameter α describe how fast the changes in luminance of the single pixels are incorporated in the image. The foreground, i.e. the moving objects in the scene, is obtained as the set of pixels that differ from the corresponding value stored in the background image more than the standard deviation of these pixels, Eq. 2.

[1]The camera and the hyperbolic mirror are kindly lent by prof. H. Ishiguro of Osaka University.

Figure 2: An example of the evolution of the dynamic background. From left to right a stationary person is gradually merged into the static background.

$$|\text{luminance}_t(i,j) - \text{background}_{t-1}(i,j)| > c \cdot \text{stdDev}_{t-1}(i,j) \tag{2}$$

Once the foreground is calculated, it is divided in blobs of similar colours and the connected blobs are considered to belong to a single object. The vision system calculates the position in the world and the three principal colours for every object in the image. The positions of the objects are sent both to the acoustic agents and to the robot, the three principal colours are sent to the mobile robot only.

4 The Acoustic Agent

Each acoustic agent is an embedded device (Fig. 1) composed of a microphone array, a DSP board for acoustic acquisition, and a PC.

When the acoustic agent receives the position of the intruder from the static agent, a beamforming algorithm is used to direct the microphone array toward the acoustic source, i.e. the intruder. The beamforming algorithm in frequency domain is performed using a microphone linear array, obtaining a main lobe in the reception diagram as presented in [9]. In other words, the inputs of the microphone array are combined in order to obtain a directional microphone. In Fig. 3 a reception diagram is reported; in this case the array is steered towards a -30 degree direction and the interfering noise coming from the broadside direction (0 degree) is de-emphasised. The beamforming algorithm is schematically depicted in Fig. 3. Besides emphasizing the signal, the beamforming aims at reducing noise and reverberation.

The adaptive algorithms for beamforming apply a vector of weights W to the vector of observations, i.e. the signals coming from the microphones in the frequency domain, in order to minimise the mean square value of the weighted observations,

$$E[(w'y)^2] \tag{3}$$

subjected to some given constraint, described by

$$Cw = c, \tag{4}$$

where C is a 'constraints matrix' and c is a vector of constraint values. If the quantity

$$R = E[yy'] \tag{5}$$

is defined as *observations correlation*, using the method of the Lagrange multipliers the general solution of the minimization problem is described by

$$w_{opt} = R^{-1}C' \left(CR^{-1}C'\right)^{-1} c. \tag{6}$$

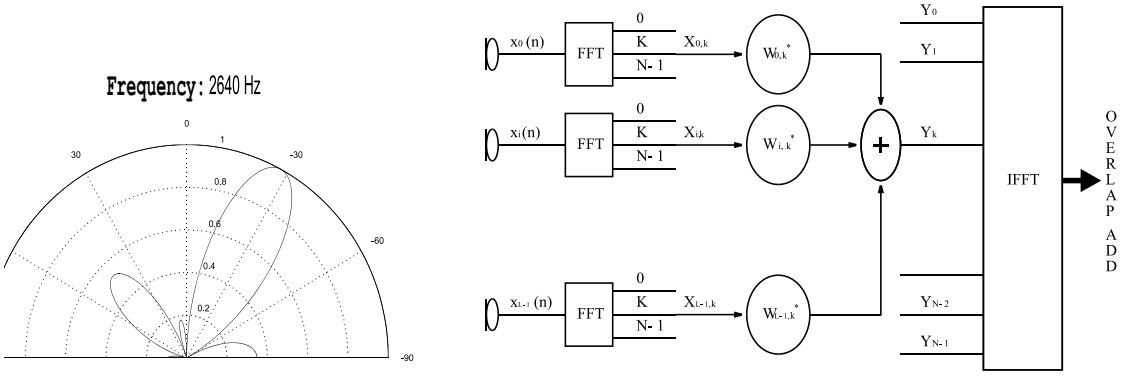

Figure 3: (Left) The reception diagram obtained for the array of microphones once beamforming is performed. (Right) A schematic representation of the beamforming algorithm.

where C' is the transpose of C.

Our frequency domain constraint, represented by equation (4) consists in a signal emphasis on a given direction (represented by the steering vector s), $s'w = 1$ and in a signal reduction on another direction (represented by the steering vector t), $t'w = 0$.

The beamforming algorithm is applied to frames derived from an incoming signal. As a sequence of frame is obtained, the signal can be reconstructed using the overlap-add method to the result of the IFFT block.

The acoustic signal obtained by beamforming is therefore cleaned up by most of other noises and it is used to train an HMM (Hidden Markov Model) of the steps of the intruder, with the technique described in [8]. When the person moves, the learnt HHM can be used to distinguish a person moving in the environment from another one and so allowing a audio tracking of a walking person, if the steps of the walker are known. Otherwise, if the walker is unknown, a new HMM is trained using the next 5 acquisitions of the acoustic agent.

The acoustic agent is also able to calculate the position of the intruder with respect to itself. The localisation algorithm is implemented using a neural network based algorithm as described in [10, 8], that takes as input the generalised cross correlation of signals acquired and gives as output the (x, y) coordinates of the acoustic source detected.

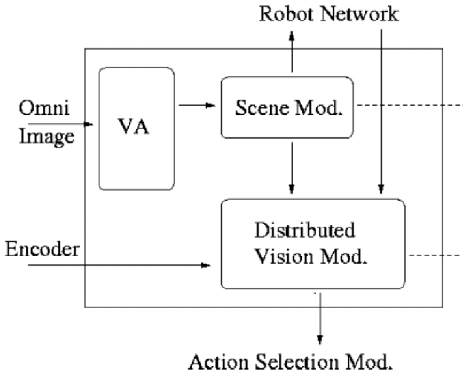

Figure 4: The architecture of the sensor fusion module.

Figure 5: (Left) The mobile robot on which is mounted the mobile vision agent. This is an holonomous robot with an omnidirectional vision system where the mirror has a custom profile. (Right) A close-up view of the omnidirectional camera of the static vision agent with the hyperbolic mirror. Note the two omnidirectional sensors have very different mirrors, so they produce very different images.

5 The fusion of the observations

The measurements on the position of the intruder coming from the static vision agent and the static acoustic agents are fused using the technique described in [7]. This technique was developed to fuse arbitrary position data coming from heterogeneous sensors. The only assumption on the measurements were that they could be described as a Gaussian probability distribution and that they are labelled with a time stamp indicating the time in which they were acquired. This system used a modified Kalman filter to fuse the measurement coming from different sensors and the information on the position of the tracked objects where stored in tracks. The peculiarity of this system is that it can accept measurements coming from heterogeneous sources with different errors associated to every estimation and that the measurements can arrive also in the wrong time order and they will be reordered thanks to the time-stamp associated to every measure. In Fig. 4 is sketched the architecture of the module performing the data fusion.

6 The Mobile Vision Agent

The mobile vision agent is implemented on board of a Golem platform developed by the Golem Team [4] bought a couple of years ago by the IAS-Lab. The Golem platform is an holonomic robot driven by three motors with omnidirectional wheels. It mounts an omnidirectional vision system realised with a Hitachi camera and a custom designed omnidirectional mirror [6]. The processing power is assured by a PC-104 with a AMD K6 400MHz CPU. As you can notice in Fig. 5, the omnidirectional camera of the mobile robot is very different from the omnidirectional camera mounted on the tripod (the static vision agents). An example of how different are the two images grabbed by these cameras is depicted in the screenshot of Fig. 8.

The mobile robot receives from the static vision agent its own position and the position of the intruder. From these data, it calculates the relative position of the intruder with respect to itself and it moves toward this position driven by the odometric data. An update on its position and the position of the intruder is received five times per second and on this short

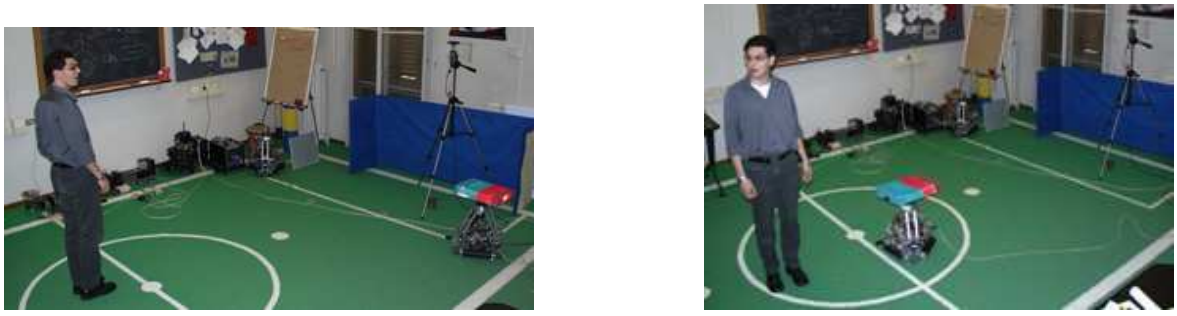

Figure 6: Two pictures taken during the preliminary experiments: (Left) the intruder enters the room; (Right) the robot approaches the intruder and recognise it in the omnidirectional image.

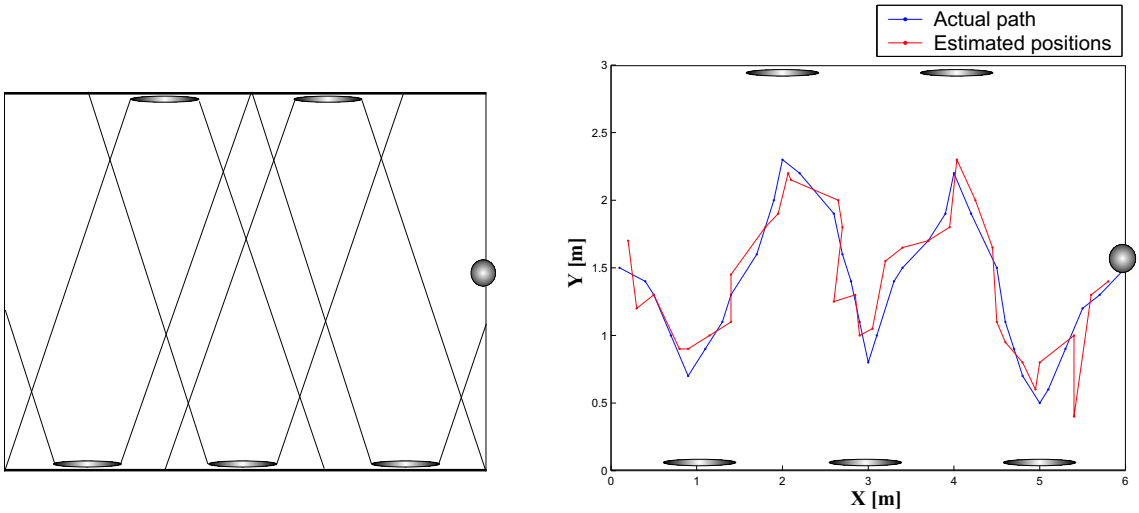

Figure 7: (Left) A sketch of the test environment with the areas monitored by every acoustic sensor. Note that in some areas of the environment there is an overlapping allowing for a simultaneous and more precise acoustic localisation of the intruder. (Right) A plot of a tracking experiment of the intruder with the ground-truth of the followed by the intruder and the estimated path.

time interval the odometric data can be considered reliable.

Once the robot reach the position communicated by the static vision agent, it analyses the current images to identify the intruder in the image. Because the two mirrors of the omni-directional cameras are different the appearance of the intruder in the two vision sensor will be very different. So the robot identifies the intruder by locating in the image the three blob of the colours transmitted by the static agent. If the intruder is identified in the image the grabbed image is sent to the monitoring station, where a graphical interface display it to the operator, see Fig. 8.

7 Experimental results

For testing the data fusion and tracking system some simple experiments were performed.

An intruder entered the room under surveillance from the left in Fig. 6 (Left). The person is detected as an intruder, the acoustic sensors perform beamforming by steering toward the intruder and its steps are learnt with the HMM as explained in Section 4. Once the position of the intruder is acquired the mobile robot is set toward the intruder, Fig. 6 (Right) and a

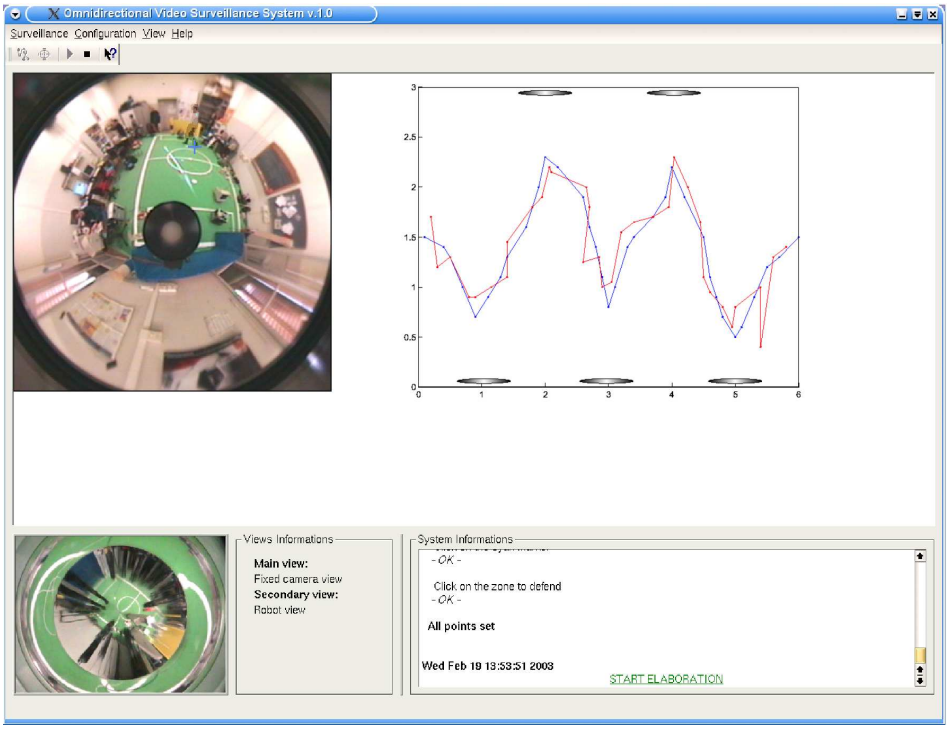

Figure 8: A screenshot of the graphical interface on the monitoring station. The remote operator is presented with the current image acquired by the static vision agent (up left), the status of the system (bottom right) and the omnidirectional image sent by the robot once the intruder is recognised by the robot (bottom left). The operator may also display the tracked path of the intruder (up right).

close-up image of the intruder is grabbed and sent back to the monitoring station that display it to the remote operator with the graphical interface depicted in Fig. 8.

In the graphical interface is possible to display also the path followed by the tracked intruder. In this experiment the intruder moves slowly (mean velocity of about 0.5 km/h) but continuously from the entrance on the left to the exit of the environment on the right.

At the time of writing, we are performing more intense tests to have a statistical analysis of the reliability of the tracking system in determining the intruder position. Up to now the system is limited to track an intruder a time, by the system is conceived to allow the tracking of multiple intruders.

8 Conclusion and future works

In this work, we presented an intelligent surveillance system able to autonomously monitor a room and to locate a track an intruder entering the room. The system uses a static vision agent, a mobile vision agent and five steerable acoustic agents. The data gathered by the heterogeneous sensors are fused to obtain a global estimation of the position of the intruder.

Future developments concern the fusion of the sensorial data provided by several mobile robot in order to have a team of surveillance robots that can “go and seek” for several intruders. At the time of writing we are further testing the system.

9 Acknowledgements

The authors wish to thanks: the students of the IAS-Lab, especially Nicola Milani and Alberto Scarpa, for writing part of the software used in these experiments. We wish to thank also Prof. Hiroshi Ishiguro of Osaka University (Japan) for lending us the omnidirectional camera.

This research has been partially supported by: the Italian Ministry for the Education, the University, and Research (MIUR), the Italian National Council of Research (CNR), The University of Padua, and the University of Trieste.

References

[1] P. Aarabi and S. Zaky. Robust Sound Localization using Multi-Source Audio-Visual Information Fusion. *Information Fusion*, 2:209–223, 2001.

[2] L. Burrelli, S. Carpin, F. Garelli, E. Menegatti, and E. Pagello. Ade: a software suite for multi-threading and networking. Technical report, Intelligent Autonomous Systems Laboratory, Department of Information Engineering, University of Padova, ITALY, 2002.

[3] R. Collins, A. Lipton, and T. Kanade. A system for video surveillance and monitoring. Technical report, Robotics Institute at Carnagie Mellon University, 2000.

[4] M. Ferraresso, M. Lorenzetti, A. Modolo, P. de Pascalis, M. Peluso, R. Polesel, R. Rosati, N. Scattolin, A. Speranzon, and W. Zanette. Golem team in middle-sized robots league. In P. Stone, T. Balch, and G. Kraetzschmar, editors, *RoboCup 2000: Robot Soccer World Cup IV*, LNCS. Springer, 2001.

[5] D. Gutchess, A. K. Jain, and Sei-Wang. Automatic surveillance using omnidirectional and active cameras. In *Asian Conference on Computer Vision (ACCV)*, January 2000.

[6] E. Menegatti, F. Nori, E. Pagello, C. Pellizzari, and D. Spagnoli. Designing an omnidirectional vision system for a goalkeeper robot. In A. Birk, S. Coradeschi, and S. Tadokoro, editors, *RoboCup-2001: Robot Soccer World Cup V.*, pages pp. 78–87. Springer, 2002.

[7] E. Menegatti, A. Scarpa, D. Massarin, E. Ros, and E. Pagello. Omnidirectional distributed vision system for a team of heterogeneous robots. In *Proc. of IEEE Workshop on Omnidirectional Vision (Omnivis'03), in the CD-ROM of Computer Vision and Pattern Recognition (CVPR 2003)*, pages On CD–ROM only, June 2003.

[8] E. Mumolo and M. Nolich. A Neural Network Algorithm for Talker Localization in Noisy and Reverberant Environments. In *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, June 8-11 2003.

[9] E. Mumolo and M. Nolich. Distant Talker Identification by Nonlinear Programming and Beamforming in Service Robotics. In *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, June 8-11 2003.

[10] E. Mumolo, M. Nolich, and G. Vercelli. Algorithms and Architectures for Acoustic Localization based on Microphone Array in Service Robotics. In *ICRA2000*, volume 3, pages 2966–2971, 2000.

[11] D. Rabinkin, R. Renomeron, A. Dahl, J. French, J. Flanagan, and M. Bianchi. A DSP Implementation of Source Location Using Microphone Arrays. *J. Acous. Soc. Am.*, 99(4), April 1996.