International Journal on Artificial Intelligence Tools

Combining Audio and Video Surveillance with a Mobile Robot

Emanuele Menegatti, Manuel Cavasin, Enrico Pagello
Dept. of Information Engineering, University of Padua,
via G. Gradenigo 6/B , Padua, ITALY
{*emg,epv*}*@dei.unipd.it*

Enzo Mumolo, Massimiliano Nolich
Dept. of Information Engineering, University of Trieste,
via Valerio 10, Trieste, ITALY
{*mumolo,mnolich*}*@units.it*

This paper presents a Distributed Perception System for application of intelligent surveillance. The audio and video sensors distributed in the environment are used as a single sensor to reveal and track the presence of a person in the surveilled environment. The system prototype presented in this paper is composed of a static acoustic agent and a static vision agent cooperating with a mobile vision agent mounted on a mobile robot. The mobile robot extends the capabilities of the system by adding a mobile sensor, in this work an omnidirectional camera. The mobile omnidirectional camera can be used to have a closer look of the scene or to inspect portions of the environment not covered by the fix sensory agents. In this paper, the hardware and the software architecture of the system and of its sensors are presented. Experiments on the integration of the audio localization data and on the video localization data are reported.

Keywords: audio and video surveillance; sensor fusion; mobile robot; omnidirectional vision

1. Introduction

In this paper, we present our current project on the development of an intelligent surveillance system that uses both mobile and static surveillance agents. The scenario of application is the monitoring of a room or a multi-room environment with a dynamic structure, for instance 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 [4] [8] 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 send to inspect suspicious areas occluded by movable objects, as already introduced in [12].

Several work deal with the integration of the information gathered by a network of cameras (among the others [14,10,9]. 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.

In our approach, the sensors distributed in the environment cooperate in order to form a sort of "super-sensor" distributed among the agent 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 agents.

2. Related works

Many researchers focused on the integration of vision and acoustic senses, motivated by the fact that there usually exists a strong correlation between the motion of a sound source and the corresponding audio data. In [6], for example, this fact has been exploited for lip/speech-reading for improving speech recognition in adverse conditions. As far as the position of a sound source is concerned, two approaches have been considered. In the this approach, audio data and vision data are fused together with suitable information fusion methods. A system able to automatically detect the identity of the talker and the position of the talker's mouth is described in [5]. In this work, the speaker's head is first box-bounded in the video data and visual features from the image are extracted as a measure of change between two subsequent images. The audio features are mel-cepstrum coefficients, which are commonly used in speech recognition systems. A Time Delay Neural Network (TDNN) is then trained to learn the audio-visual correlations between audio and visual features. Another possibility is to process separately each channel to get the localization information of the two sources and the results are integrated only in the final step. One example is presented in [3]. In this work, the position of the sound source (a talking mouth) in a video scene is estimated by fusing auditory and visual information, based on skin-color and nonskin-color information, using a Bayesian network. A different approach is the system described in [17] 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.

3. A system overview

As we said in the introduction, the system is composed of several sensors. The sensors are shown in Fig. 1 and in Fig. 2. In Fig. 1 are depicted: one of the static Vision Agents composed of an omnidirectional camera with a hyperbolic mirror (on a tripode on the left of the image), and one the mobile robots (on the left bottom of the image). The robot is equipped with an omnidirectional camera with a different

miror profile. It mounts a multi-part omnidirectional mirror [11]. The vision system on board of the robot is called mobile Vision Agent. In Fig. 2 is imaged the audio sensor (Static Acoustic Agent) composed of a circular microphone array able to perform beamforming and to estimate the position of a person using its speech.

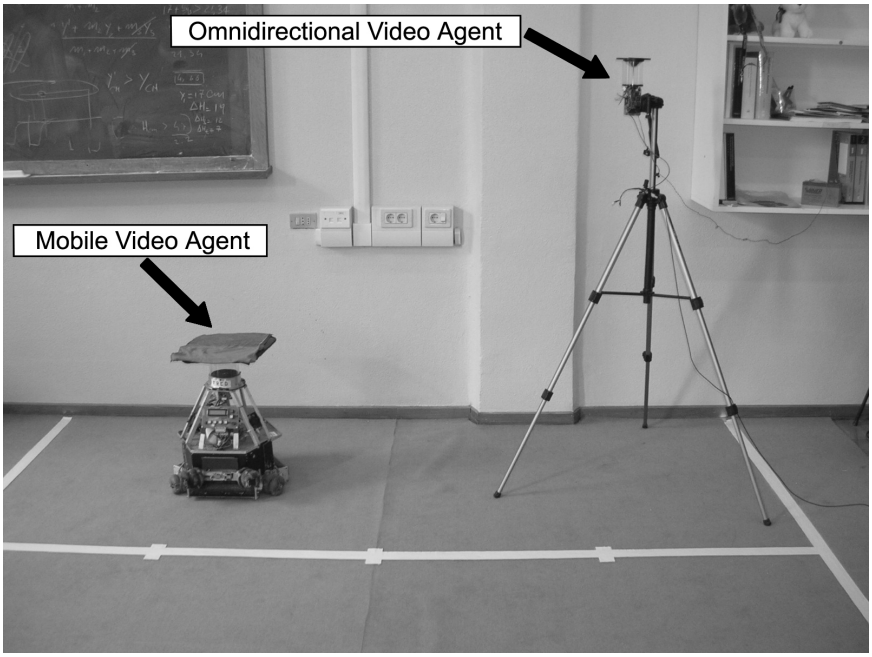

Fig. 1. The two vision agents the static one on the tripod and the mobile one on the mobile robot.

Every sensory agent is realised with a sensor (microphone or camera) connected to a computer equipped with a IEEE 802.11b wireless LAN card. The computer 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 middlewar, we developed for the RoboCup project, called ADE [2]. 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 robots. The basic functioning of the system is:

- the static vision agent, i.e. the omnidirectional camera over the tripod, detects moving objects in the image and transmits their coordinates in the world frame of reference to the static acoustic agent;

Fig. 2. The audio sensory agent: on the left, the circular microphone array used by the audio agent; on the right, the acoustic agent present in the environment.

- the static acoustic agent performs beamforming in the direction of the detected motion, estimates the position of the noise produced by the intruder and start tracking it;
- the different measurements on the position of the intruder coming from the static vision agent and static acoustic agent are fused by the computer of the static acoustic agent in order to improve the position estimation, which is sent to the mobile robot and used for moving it toward the position of the localized intruder;
- once the intruder is detected by the mobile vision agent a close-up image is sent to the monitoring station, so an operator can check if the moving object represents a danger or if it is just a false alarm. Moreover the mobile robot might ask the intruder to present itself using speech and it is verified if the person is authorized or not on the basis of its speech.

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

4. The Static Vision Agent

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

To detect the intruder the image is segmented into a 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 obstacle might change configuration in time. For this reason we adopted a historical 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. 3 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 historical background is calculated according to Eq. 1, creating a grey-level image representing the fix luminance in the image.

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

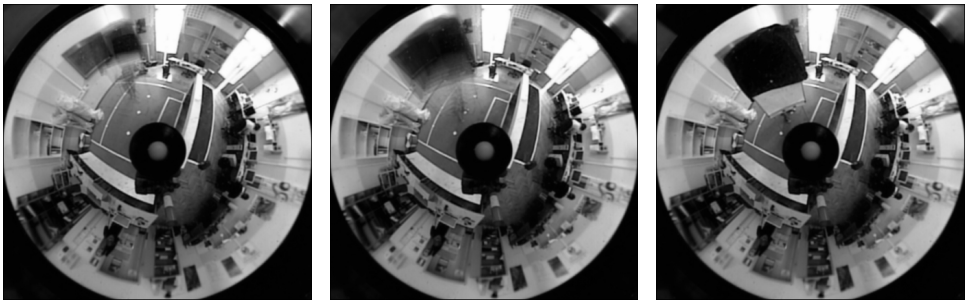

Fig. 3. An example of the evolution of the dynamic background. From left to right an object that was moved into a new position, and then stays stationary, is gradually merged into the static background.

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 historical image more than the standard deviation of these pixels, Eq. 2.

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

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

Once the foreground is calculated, it is divided in blobs of similar colors and the connected blobs are considered to belong to a single object. For every object are calculated the position in the world and the three principal colors. The position of the object are sent to the acoustic agents and to the robot to be used in the subsequent tracking steps (the three principal colors are sent to the mobile robot only).

5. The Static Acoustic Agent

The acoustic agent is composed of a microphone array (shown in Fig. 2), a DSP board for acoustic acquisition and processing and a host PC. The different tasks performed by the acoustic agent are discussed in details in the following.

5.1. *Circular microphone array based localization*

Microphone array technologies are commonly used for performing acoustic localization, both in 2D and in 3D, and several techniques can be adopted [16]. One class of algorithms can be derived directly from antenna array theory and are well suited for narrow-band signals. Another class of algorithms, well suited for wide band signals, are based on the Generalized Cross Correlation.

A 2D acoustic localization algorithm suited for wide band signals and circular arrays is presented. Circular arrays allow an omnidirectional localization around the acoustic agent. Only 2D localization is considered, which provides enough information to plan the movements of the robots.

In this work a circular array has been considered, which has a 30 cm diameter and 32 microphones equally spaced on the circumference. Out of the 32 microphones, the 16 microphones directed towards the acoustic source are selected on the basis of energetic considerations.

The localization of the source is determined from the knowledge of the time delay between microphone pairs. The estimation of the localization from the time delay is obviously a non linear problem. However, by introducing some approximations it is possible to derive simple geometrical methods to solve this problem.

- **Estimation of the time delay.** Popular approaches for the estimation of the time delay of arrival of an acoustic signal to a couple of microphones are based on the maximization of the cross-correlation between a couple of signals $s_i(t)$ and $s_j(t)$ received by microphones i and j: $R_{ik}(\tau) = E\{s_i(t)s_k(t+\tau)\}$. In fact, assuming that a reasonable model for the signal received by microphone i is $s_i(t) = \alpha_i r(t-\tau_i) + n_i(t)$, where τ_i is the time of flight from the source $r(t)$ to the microphone i and α_i is the propagation lossy factor, the cross-correlation becomes

$$R_{ik}(\tau) = \alpha_i \alpha_k R_{rr}(\tau - \delta_{ik}) + R_{n_i n_k}(\tau) \tag{3}$$

  where R_{rr} is the autocorrelation of the acoustic source $r(t)$. Sharp cross-correlation peaks can be obtained by filtering in the spectral domain. More

precisely, a spectral weighting filter $\psi(f)$ [?] can be introduced to whiten the input signal:

$$R_{ik}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_g(f) G_{ik}(f) e^{j2\pi f\tau} df \tag{4}$$

The function reported in (4) is called Generalized Cross Correlation (GCC). Various choices of the weighting function are possible. For instance, the $\psi(f)$ function can be derived with a Maximum Likelihood formulation leading to the TDOA (Time Delay Of Arrival) algorithm as described in [?].

Another approach is the Modified Cross-power Spectrum Phase (MCSP) estimator [?]:

$$\psi_{MCSP}(f) = \frac{1}{|G_{ik}(f)|^{\rho}} \tag{5}$$

where $0 < \rho \leq 1$.

- **Geometric consideration.** The localization of the source is determined

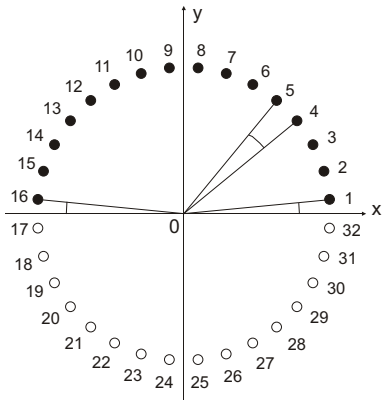

Fig. 4. Geometric location of the microphones in the circular array

from the knowledge of the time delay between microphone pairs. The estimation of the localization from the time delay is obviously a non linear problem. However, by introducing some approximations it is possible to derive simple geometrical methods to solve this problem.

In this work a circular array has been considered, which has a 30 cm diameter and 32 microphones equally spaced on the circumference as represented in Fig. 4. Out of the 32 microphones, the 16 microphones directed towards the acoustic source are selected on the basis of energetic considerations.

- **Estimation of the TDOA with Neural networks.** The neural network model adopted was a Multi-Layer Perceptron [?] with one hidden layer. Each

hidden node use the hyperbolic tangent as activation function. With reference to Fig. 4, the sixteen microphones towards the source are divided into eight couples as follows: $(1,5)$, $(2,6)$, $(3,7)$, $(4,8)$, $(9,13)$, $(10,14)$, $(11,15)$, $(12,16)$. For each couple the time delay is computed using MCSP. $\delta_1, \delta_2, \cdots, \delta_8$ are given as input to the neural network. Several optimization techniques [?], including in particular backpropagation with momentum, the Levemberg-Marquardt approach, and Newton-based approaches, have been tested for training the neural network, and the best results were obtained with Levemberg-Marquardt [?] and Rprop [?].

- **TDOA performances.** The localization is based on the estimation of the TDOA using the MCSP as described in eq. (5).

  Let us summarize now the procedure: first the signal is divided into frames and then a MCSP function is computed on the considered frame. The TDOA is then estimated by peak picking. Besides the usual approach to make an average estimation of the localized coordinate, which has a long algorithmic delay as it requires to localize each incoming frame, a faster approach was investigated: the localization was performed on the maximum energy frame only and on the first frame only. Both these approaches seemed reasonable, because the former implies a higher SNR while the latter is less affected by echoes and reverberations.

  The parameters to optimize are therefore: whether the best results are obtained using the first frame or the maximum energy one, the frame dimension and the value of the ρ used in the MCSP formulation. The optimization has been performed on the basis of the geometric TDOA described in eq. (6):

$$TDOA_{geometric} = \text{round}\left\{ \frac{d(\mathbf{p}, \mathbf{m_1}) - d(\mathbf{p}, \mathbf{m_2})}{V_{sound}} \cdot f_s \right\} \tag{6}$$

  where $\mathbf{p}$ is the source position, $(\mathbf{m_1}, \mathbf{m_2})$ is the microphone couple, $d()$ is the distance measure, V_{sound} is the sound velocity and f_s is the sampling frequency. The analysis has been carried out by computing the number of times that a set of parameters gave a TDOA equal to that obtained with eq. (6).

  The results are reported in 5, which shows that the best results are obtained for the first detected frame of the vocal signal with a frame length equal to 1024 and $\rho = .5$ while for DTMF signals the best results are obtained for the first detected frame but with a frame dimension equal to 128 and a $\rho = 0$.

  It was the considered the possibility to average several TDOA results instead than a single frame. The results are that both for the vocal signal and the DTMF signal the TDOA improvements obtained averaging several frames are not significant.

  The TDOA estimation described so far is obtained from a couple of

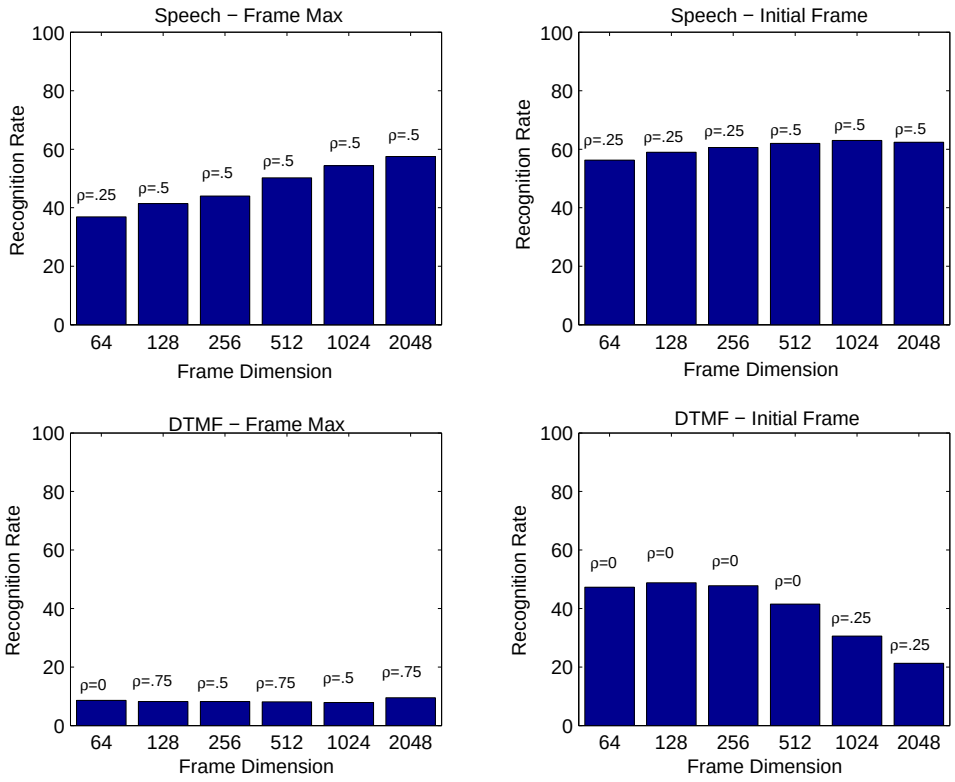

Fig. 5. TDOA results.

microphones. Coming back to Fig.4 we see that, out of the 32 microphones of the array, several definitions of the microphone couples are possible. We considered 8 couples in each semi-circle, according to the description reported in Table 1.

Conf.1	1-9	2-10	3-11	4-12	5-13	6-14	7-15	8-16
Conf.2	1-8	2-7	3-6	4-5	9-16	10-15	11-14	12-13
Conf.3	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-16
Conf.4	1-3	2-4	5-7	6-8	9-11	10-12	13-15	14-16
Conf.5	1-5	2-6	3-7	4-8	9-13	10-14	11-15	12-16
Conf.6	1-4	2-3	5-8	6-7	9-12	10-11	13-16	12-15
Conf.7	1-16	2-15	3-14	4-13	5-12	6-11	7-10	8-9
Conf.8	1-8	2-5	3-6	4-7	9-16	10-13	11-14	12-15

In Fig. 6 the average absolute localization errors obtained using geo-

metrical localization are reported. The configuration that provides better results is the nr. 5.

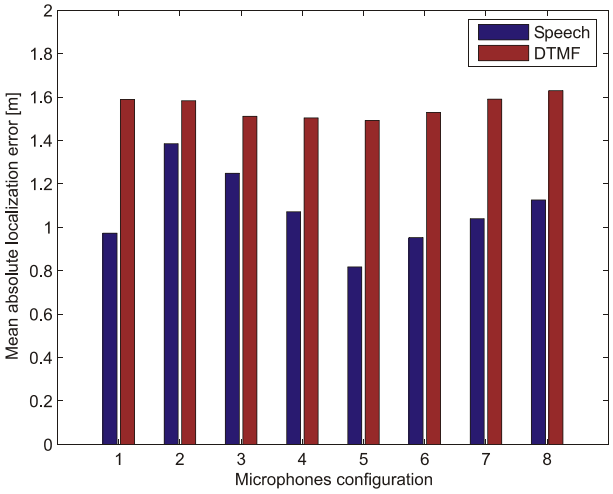

Fig. 6. Localization performance using different microphone configurations.

- **From TDOA to source coordinates: acoustic localization.** We tested two approaches for acoustic localization. The first approach is based on a classical triangulation as described in [?].

  The second approach is based on Neural Networks (NNs). The training of the NN has been performed by dividing a $8m \cdot 8m$ area around the omnidirectional device into a grid, as shown in Fig. 7, and playing in the points of such grid a signal. Half a grid is used for training the NN while the remaining half is used for testing. The network has 8 inputs, coming from 8 microphone couple, and two outputs, that is the X, Y coordinates of the sound source. For increasing the effectiveness of the training, other artificially shifted signals has been added to the signal played in the points of the grid. Two classical techniques for training the network are used, namely he Rprop and the Levenberg-Marquardt. The former is less computational expensive but it requires a higher number of iterations to converge towards a good local minimum while the latter has a greater computational cost but it requires a lower number of iterations. Average localization errors in meters for the two algorithms are shown in Fig. 8 for speech and DTMF signals respectively.

Two kinds of acoustic signals were tested: speech and DTMF tones. The speech used for testing is composed by three Italian phrases typical of human-robot inter-

action:

(1) Vieni qui (Come here);
(2) Vai al sito A (Go to A site);
(3) Prendi l'oggetto B (Take the object B).

The DTMF tones used for testing are three dial tones used in telephony.

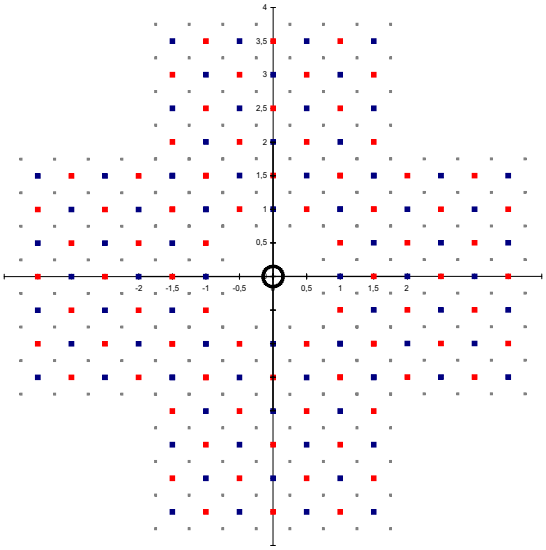

Fig. 7. Training grid

Real signals were acquired in the big points on the grid of Fig. 7 in our laboratory. In each point 10 replicas of the same 6 signals were acquired: 5 replicas have been used for training and the other 5 for testing. Other synthetic signals were created shifting the original signals as it were emitted in the small points of Fig. 7. In Fig. 8 results concerning speech and DTMF tones localization are reported. It is depicted the absolute mean localization error of acoustic signal considering two different neural network training algorithms: Rprop and Levemberg-Marquard. Better results have been obtained using the Rprop learning algorithm, obtaining an absolute mean localization error of about 45 cm.

The training is performed offline and the system operates really fast only using the pre-learned neural network. Using such approach we can obtain better results than using linear intersection algorithm of Rabinkin [?].

In Fig. 9 a comparison between geometric linear intersection localization and

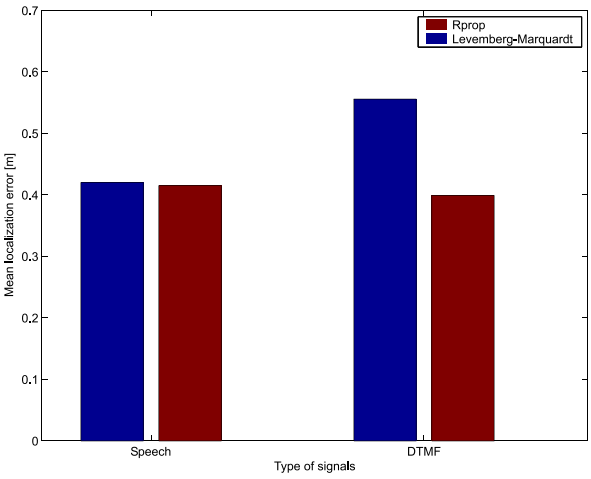

Fig. 8. Average performances of sound localization using Speech and DTMF tones

neural network localization (trained using Rprop) is presented: the histograms report the mean absolute localization error (in meter) for two types of signal used, namely Speech and DTMF tones. It is evident that the neural network approach gives better performances.

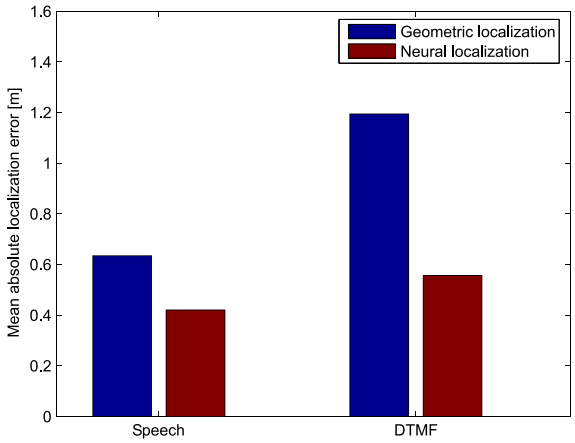

Fig. 9. Comparison between geometric linear intersection and neural localization.

5.2. *Microphone array and beamforming*

5.2.1. *Preliminaries*

A sensor can be viewed as a window, called *aperture*, through which a field of certain physical quantities is measured. The aperture is described by its aperture function, which contains information on dimension and shape of the window, and describes how the measure depends on the direction of arrival of the variable physical quantities. If we consider a situation where there is a source generating a field which propagates in the space, identified by $f(x,t)$, and a finite number of apertures, we have a signal which is the result of a spatial sampling of the field, that is the signal $y_m(t) = f(m \cdot d, t)$ where d is the spatial distance between the apertures, or the sampling interval in space. As in temporal sampling, the original signal can be reconstructed from its spatial samples using the sampling function, where the spatial frequency instead of oscillation frequency is used. Each signal $y_m(t)$ measured at the m-th aperture can be modified by multiplying the signal itself by a weight w_m. Let us consider the weighted signal $z(t) = \sum_{m=0}^{M-1} w_m y_m(t - \tau_m)$. This is the simplest form of beamforming, called *delay and sum*, since if the delays τ_m are chosen equal to the time delay of arrival (TDOA) of the second to the M-th microphone relative to the first microphone, the signal coming from a certain direction is incremented while the signal coming from other directions is decremented. The delay and sum beamforming operation can thus be described, in the spectral domain, as $Z(\omega) = \sum_{m=0}^{M-1} w_m Y_m(\omega) e^{j\omega\tau_m}$. Defining the steering vector $s_M(\omega)$ as the set of elements which cancel the plane-wave signal's propagation related phase, more precisely $s_M(\omega) = [1, e^{-j\omega\tau_2}, e^{-j\omega\tau_3}, \ldots, e^{-j\omega\tau_M}]$, the beamforming operation is described as $Z(\omega) = \sum_{m=0}^{M-1} Y_m(\omega) w_m s_m^*$.

5.2.2. *Minimum variance beamforming*

When the acoustic agent receives the position of the intruder from the static vision 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 the circular microphone array, obtaining a directional main lobe in the reception diagram. In other words, the inputs of the microphone array are combined in order to obtain a directional microphone. In Fig. 10 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. 11.

The adaptive algorithms for beamforming apply a vector of weights $W_i = w_i e^{-j\omega\tau_m}$ 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, such that $w_i = argminE[|z(t)^2|]$. Minimizing power presumably reduces the effect of noise and unwanted signals. Using the method of the Lagrange multipliers the general solution of the minimization problem is described

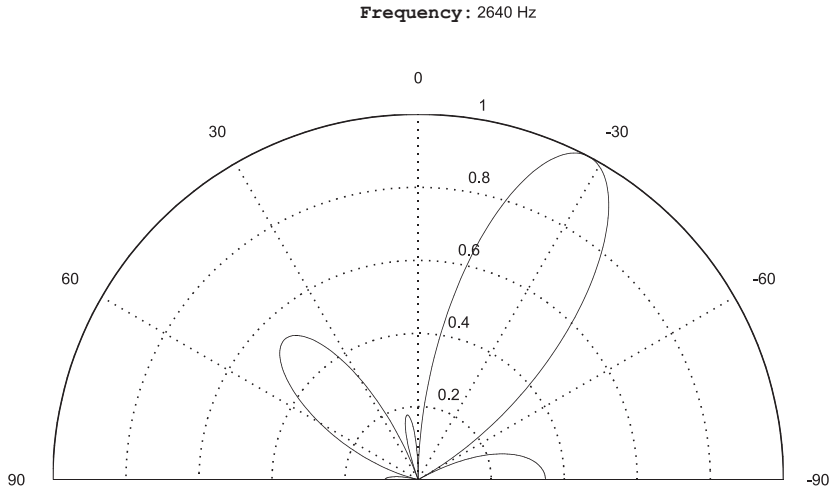

Fig. 10. The reception diagram obtained for the array of microphones once beamforming is performed.

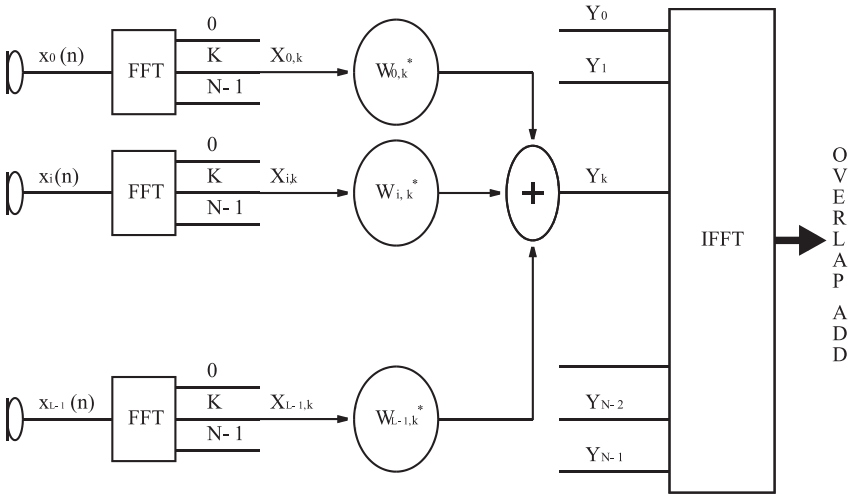

Fig. 11. A schematical representation of the beamforming algorithm.

by

$$w_{opt} = \frac{R^{-1}d}{d^{*}R^{-1}d}. \tag{7}$$

where R is the normalized cross power spectral density.

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.

5.3. *Speaker classification*

The acoustic signal obtained by beamforming is therefore cleaned up by most of other noises and can be used to train an HMM (Hidden Markov Model) of the speech of the human person, with the technique described in [15]. When the person moves, the learnt HHM can be used to identify a person moving in the environment by his/her voice from another person and so allowing a audio tracking of a walking person. Otherwise, if the voice is unknown, a new HMM can be trained using the next five acquisitions of the acoustic agent.

6. The fusion of the observations

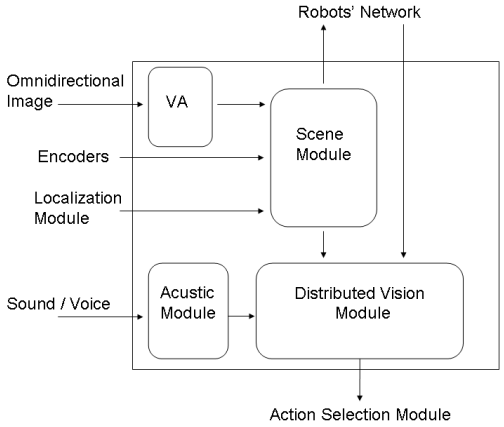

Fig. 12. The architecture of the sensor fusion module.

To improve the localization results, 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 [14]. This technique was developed to fuse 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. 12 is sketched the architecture of the module performing the data fusion.

7. The Mobile Vision Agent

The mobile vision agent is implemented on board of a Golem platform developed by the Golem Team [7] 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 customly designed omnidirectional mirror [13]. The processing power is assured by a PC-104 with a AMD K6 400MHz CPU. As one can notice in Fig. 13, 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. 15.

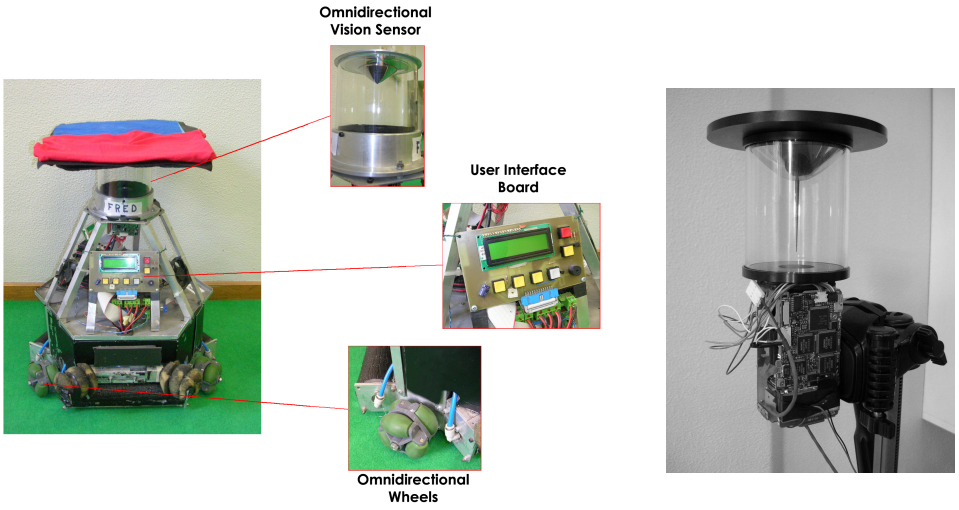

Fig. 13. (Left) The mobile robot on which is mounted the mobile vision agent. This is an holonomous robot with an omnidirectional vision system where the mirror has a custom profile. (Center) A close-up view of the principal robot's carachteristics. (Right) A close-up view of the omnidirectional camera of the static VA.

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 ten times per second and on this short 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 omnidirectional 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. 15.

8. Experimental results

For testing the data fusion and tracking system some simple experiments were performed. In the first one, an intruder enters the surveilled room from the left in Fig. 14. Once the position of the intruder is acquired, the mobile robot moves toward the intruder, as shown in the right panel of Fig. 14, and a close-up image of the intruder is grabbed and sent back to the monitoring station that displays it to the remote operator with the graphical interface depicted in Fig. 15.

In the graphical interface it is displayed 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.In the second experiment, a talker enters the room and follows the walls. Its position is tracked by the system as shown in Fig. 14. 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, but the system is conceived to allow the tracking of multiple intruders.

9. 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 data gathered by the heterogeneous sensory agents are fused to obtain a global estimation of the position of the intruder. The system uses a static vision agent, a mobile vision agent and five steerable acoustic agents, but it has been designed in order to connect any number of sensory agents.

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.

10. 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 this experiments. We wish to thank also Prof. Hiroshi Ishiguro of Osaka University (Japan) for lending us the omnidirectional camera.

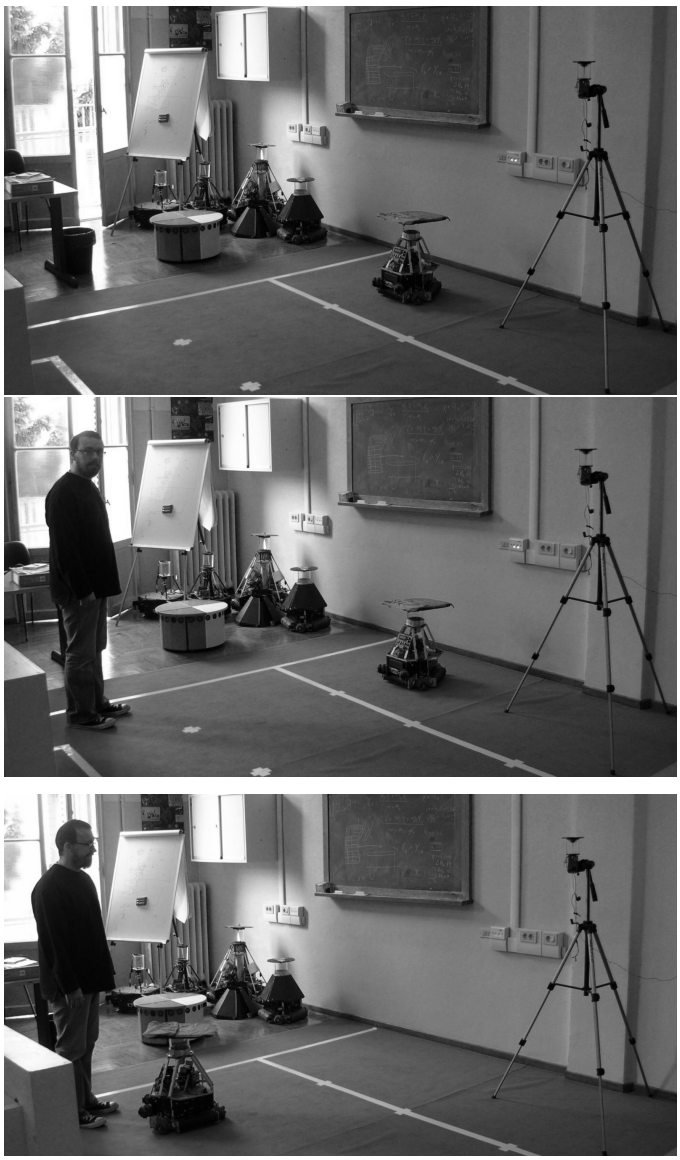

Fig. 14. Three pictures taken during the preliminary experiments: (Top) the robot is patrolling; (Middle) an intruder enters in the surveilled room; (Bottom) the robot approaches the intruder directed by the Static VA on the right of the picture and recognize it in its omnidirectional image.

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

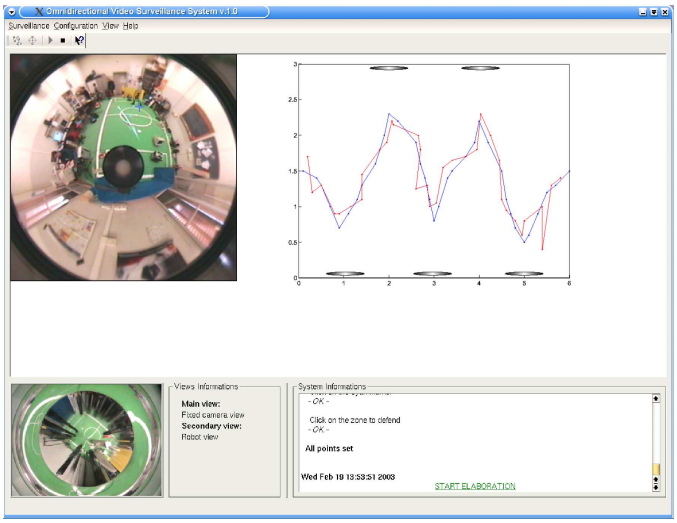

Fig. 15. A screenshot of the graphical interface on the monitoring station. The remote operator is presented with the current image acquired by the static vision agent (up left), the status of the system (bottom right) and the omnidirectional image sent by the robot once the intruder is recognised by the robot (bottom left). The operator may also display the tracked path of the intruder (up right).

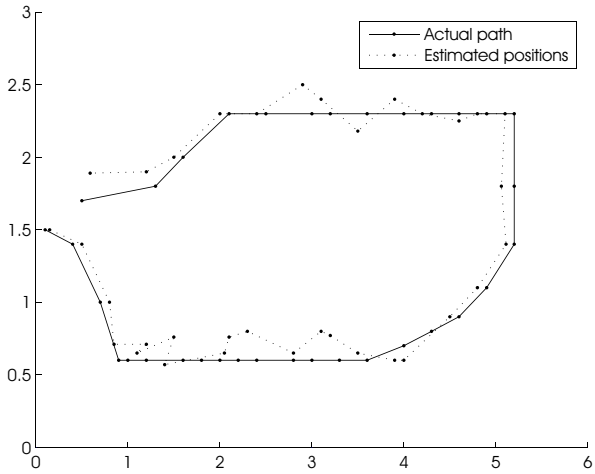

Fig. 16. The actual and the estimated path followed by a speaker walking in the environment. The static acoustic agent is placed in the middle of the room.

Laboratory, Department of Information Engineering, University of Padova, ITALY, 2002.

3. B. Chen, M. Meguro, and M. Kaneko. Probabilistic integration of audiovisual infor-

mation to localize sound source in human-robot interaction. *Proceedings of the 2003 International Workshop on Robot and Human Interactive Communication*, 2003.
4. R. Collins, A. Lipton, and T. Kanade. A system for video surveillance and monitoring. Technical report, Robotics Institute at Carnagie Mellon University, 2000.
5. R. Cutler and L. Davis. Look who's talking: speaker detection using video and audio correlation. *IEEE International Conference on Multimedia and Expo, 2000.*, 2000.
6. S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE IEEE Transactions on Multimedia*, 2(3):14–151, 2000.
7. 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.
8. D. Gutchess, A. K. Jain, and Sei-Wang. Automatic surveillance using omnidirectional and active cameras. In *Asian Conference on Computer Vision (ACCV)*, January 2000.
9. H. Ishiguro. Distributed vision system: A perceptual information infrastructure for robot navigation. In *Proceedings of the Int. Joint Conf. on Artificial Intelligence (IJCAI97)*, pages 36–43, 1997.
10. E. Kruse and F. Wahl. Camera-based monitoring system for mobile robot guidance. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS98), pages=pp. 1248-1249, year = 1998.*
11. F. Marchese and D. G. Sorrenti. Omni-directional vision with a multi-part mirror. In P. Stone, T. Balch, and G. Kraetzschmar, editors, *RoboCup 2000: Robot Soccer World Cup IV*, LNCS. Springer, 2001.
12. E. Menegatti, E. Mumolo, M. Nolich, and E. Pagello. A surveillance system based on audio and video sensory agents cooperating with a mobile robot. In *Proc. of 8th International Conference on Intelligent Autonomous Systems (IAS-8)*, pages 335–343, Amsterdam - The Netherlands, Month 2004.
13. 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.
14. 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.
15. 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.
16. E. Mumolo, M. Nolich, and G. Vercelli. Algorithms for acoustic localization based on microphone array in service robotics. *Robotic and Autonomous Systems*, 1024:1–20, 2002.
17. 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.