International Journal on Artificial Intelligence Tools
© World Scientific Publishing Company

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,cavasinm,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 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 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 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

Several works deal with the integration of the information gathered by a network of cameras[18,12]. In this paper, we focus on the integration of the visual and audio information provided by different “sensing agents”.

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. Dupont *et al.* [7], for example, exploited this fact 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 first approach, audio data and vision data are fused together with suitable information fusion methods. Cutler et al. described a system able to automatically detect the identity of the talker and the position of

[*] also with ISIB-CNR corso Stati Uniti, Padua, ITALY

the talker's mouth [6]. In that 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 to integrate the results only in the final step. An example of this is presented by Chen *et al.*. [4] In that 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 by Rabinkin *et al.*[21] which 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 by Aarabi *et al.*[1], i.e. a multimodal sound localisation system that uses two cameras and a 3-element microphone array. Their approach seemed to be reliable only when using ad-hoc narrow band acoustic signals. In this work, we show that the integration of the data is effective even using the noise of the foot-steps of the intruder.

In this paper, we present 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 [5] [10] 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 movable objects. 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.

This paper extends a work already presented[16] by introducing a new acoustic sensor: an omnidirectional microphone array, by adopting a more standard communication middleware based on ACE/TAO [24] [25] in addition to the custom built called ADE [3], and by synchronizing all mobile and static clients and the servers existing in the system via the well-known Network Time Protocol (NTP)[a].

[a] URL: http://www.eecis.udel.edu/∼ntp

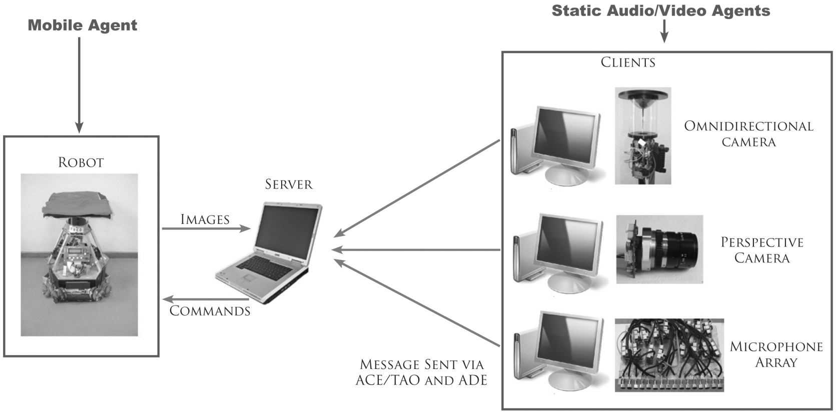


Fig. 1. A schematic representation of the elements of the surveillance Distributed Perception System.

2. A system overview

The Distributed Perception System (DPS) can be composed of several sensors, as shown in Fig. 1. Each sensor processes the data collected about the environment and sends messages containing the results of its processing to the central server via one of the different middleware ACE/TAO or ADE, depending on the type of message. The server is running a software able to integrate the different measurements of the different sensors. The server can reconstruct a high level model of the monitored environment and can control a mobile robot and use it as a mobile perceptual agent. The sensors used in this work are shown in Fig. 2 and in Fig. 3. In Fig. 2 we depict the static Vision Agent (SVA), composed of an omnidirectional camera with a hyperbolic mirror (on a tripode on the right of the image), and a mobile robot (on the bottom left of the image). The robot is equipped with an omnidirectional camera with a mirror profile different from the one used by the static Vision Agent. It mounts a multi-part omnidirectional mirror [15]. The vision system on board of the robot is called mobile Vision Agent. In Fig. 3 we depict the audio sensor (Static Acoustic Agent) which is composed of a circular microphone array able to perform beamforming and to estimate the position of a person using his/her speech.

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 two different middlewares. The first one was developed at the IAS-Lab for the RoboCup project, and is called ADE [3] (Thanks to ADE, message passing from one agent to the other is totally transparent, irrespectively if they reside on the same machine or on machines connected through a LAN or a wireless LAN).

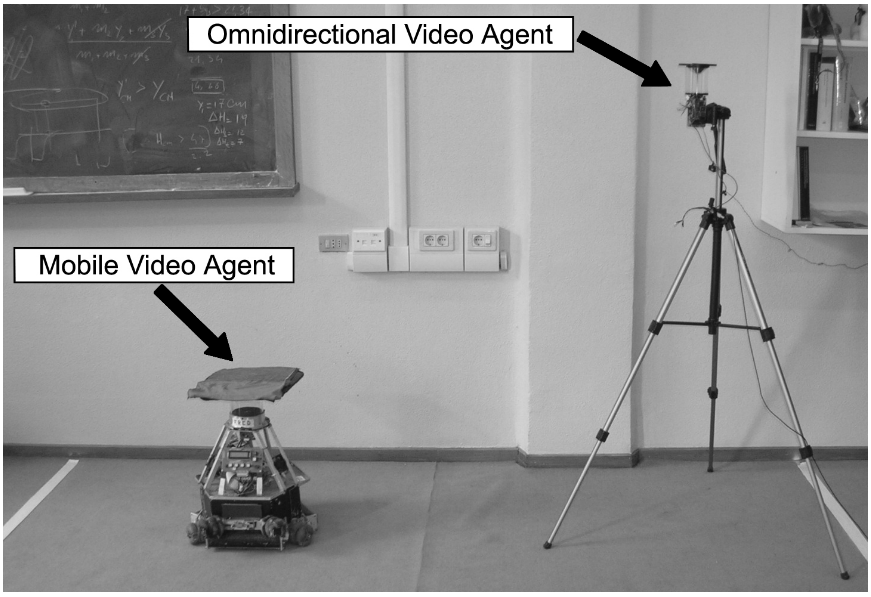


Fig. 2. The two vision agents the static one on the tripod and the mobile one on the mobile robot.

The second one is ACE/TAO and offers higher flexibility and performances thanks to the standardization.

The system is able to detect and track intruders in an 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;
- 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 estimate, which is then 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 verify if the person is authorized or not with a speaker recognition system.

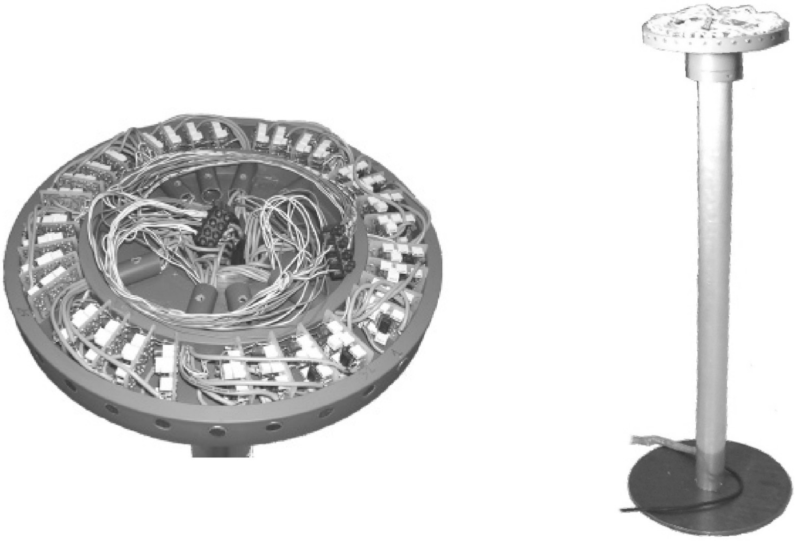


Fig. 3. The audio sensory agent: on the left, a close-up view of the circular microphone array used by the audio agent; on the right, the acoustic agent in the final setup, mounted on a pole 1.2 m hight.

In the next sections we discuss the implementation of the individual parts of the system: the Static Vision Agent, the Static Acoustic Agent, the Mobile Vision Agent, and the sensor fusion module.

3. The Static Vision Agent

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

To detect the intruder, the image is segmented into a moving foreground and into a stationary background. As we said, our system is designed to work in a dynamic environment in which the objects and the obstacles might change configuration over 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. 4 is depicted a sequence in which the history image is changing to incorporate a black object moved close to the omnidirectional camera staying there for a long time. On the left image, the object is just a ghost on the top left of the image, in the centre image, the ghost of the object becomes more perceptible, on the right image the object is merged into the historical background. The historical background is calculated according to

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

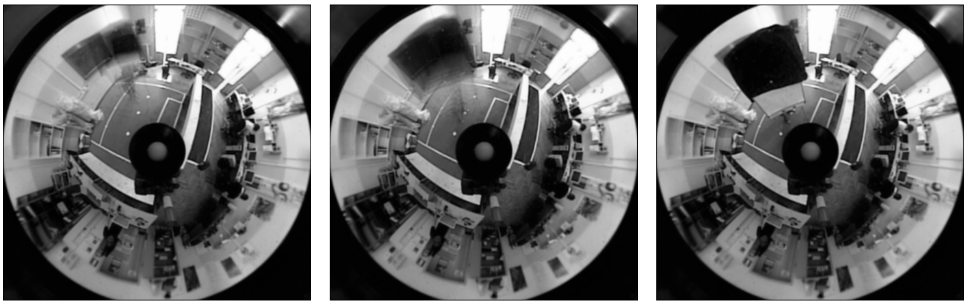


Fig. 4. An example of the evolution of the dynamic background. From left to right an object that was moved into a new position, and then stays stationary, is gradually merged into the static background.

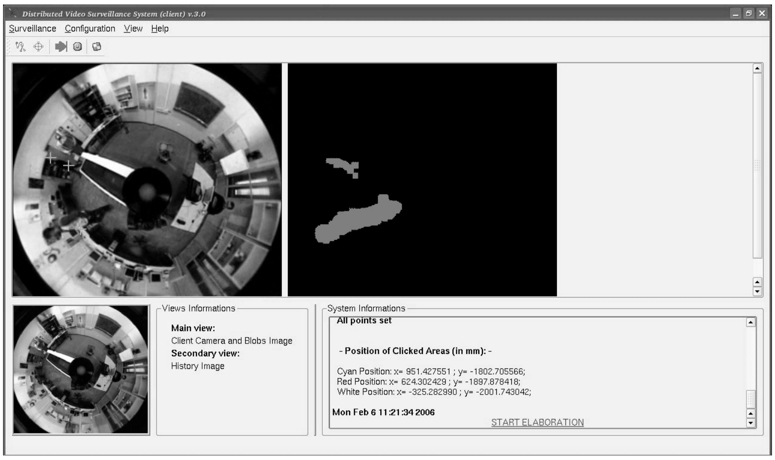
Eq. 1, by creating a grey-level image representing the fix luminance in the image. The luminance is obtained by the channel Y of the image representation in the YUV color space.

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

The parameter α describes how fast the changes in luminance of the individual 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 a certain percentage of the standard deviation of these pixels. The constant c in Eq. 2 is controlling this percentage. Thus, the standard deviation of each pixel is used as an adaptive threshold to determine if the pixel belongs to the foreground. This takes into account situations in which some pixels can change quite a lot in time, but they should not be considered as part of the foreground, as in the classical example of the leaves of a waving tree. They are moving, so the corresponding pixels change in time, but this change does not correspond to a moving object in the scene. The standard deviation of each of these pixels captures this variation. A pixels is considered to belong to a moving object only if it changed more than its usual variation. The image processing software of the SVA is running at 15 frame per second. This ensures that for typical speed of a walking person the ghosting effect typical of background subtraction algorithms is not present or is very limited.

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

Once the foreground is calculated on the Y component of the image, the colors existing in the foreground are taken into account, in order to divide it into blobs of similar colors. A two-sweep connected component algorithm is used to cluster the pixels into differentblobs. The connected blobs are considered to belong to a single object. For every object in the foreground its position in the world coordinate system and its three principal colors are calculated and sent to the Distributed Perception System. The world coordinates of the object in the foreground are calculated as



Main view:
Client Camera and Blobs Image
Secondary view:
History Image

All points set

- Position of Clicked Areas (in mm): -

Cyan Position: x= 951.427551 ; y= -1802.705566;
Red Position: x= 624.302429 ; y= -1897.878418;
White Position: x= -325.282990 ; y= -2001.743042;

Mon Feb 6 11:21:34 2006

Fig. 5. A screenshot of the graphical interface of the client of the Static Vision Agent (SVA) client. On the left, the omnidirectional image grabbed by the SVA. On the right, the foreground calculated. Bottom left, the current historical background. In this image two persons are moving.

the world coordinates of the object's pixel closest to the centre of the image. This assumes the camera calibration is known and the objects lay on the floor (sensible assumptions for the system in use).

4. The Static Acoustic Agent

The acoustic agent is composed of a microphone array (shown in Fig. 3), 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.

4.1. *Circular microphone array based localization*

Microphone array technologies are commonly used for performing acoustic localization, both in 2D and in 3D. Several techniques can be adopted [20]. One class of algorithms can be derived directly from antenna array theory. They are well suited for narrow-band signals. Another class of algorithms, well suited for wide band signals, is based on the Generalized Cross Correlation. A 2D acoustic localization algorithm suited for wide band signals and circular arrays are presented. Circular arrays allow for omnidirectional localization around the acoustic agent. We use only 2D localization, 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 its 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

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

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 $\psi(f)$ [14] can be introduced to whiten the input signal:

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

The function reported in Eq. 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 [2].

Another approach is the Modified Cross-power Spectrum Phase (MCSP) estimator [22]:

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

where $0 < \rho \leq 1$.

- **Geometric consideration.**
- **Estimation of the TDOA with Neural networks.** The neural network model adopted was a Multi-Layer Perceptron [11] with one hidden layer. Each hidden node use the hyperbolic tangent as activation function. With reference to Fig. 6, the sixteen microphones towards the source are divided into eight pair as follows: $(1,5)$, $(2,6)$, $(3,7)$, $(4,8)$, $(9,13)$, $(10,14)$, $(11,15)$, $(12,16)$. For each pair the time delay is again computed using the MCSP. $\delta_1, \delta_2, \cdots, \delta_8$ are given as input to the neural network. Several optimization techniques [11], including in particular backpropagation with momentum, the Levemberg-Marquardt approach, and Newton-based approaches, have been tested for training the neural network. The best results were obtained with Levemberg-Marquardt and Rprop [23].
- **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

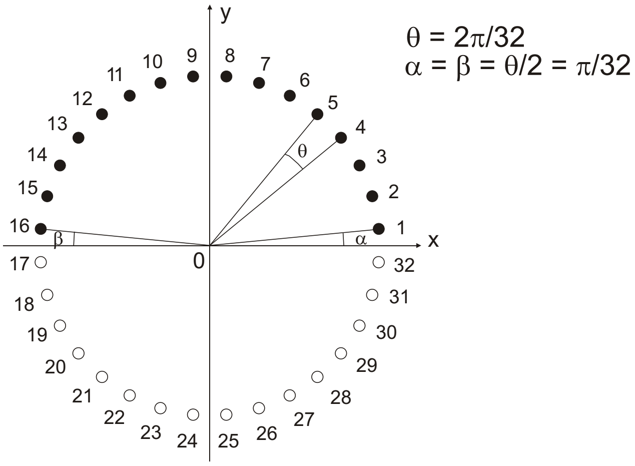


Fig. 6. Geometric location of the microphones in the circular array

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 7, 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 $\rho = 0$.

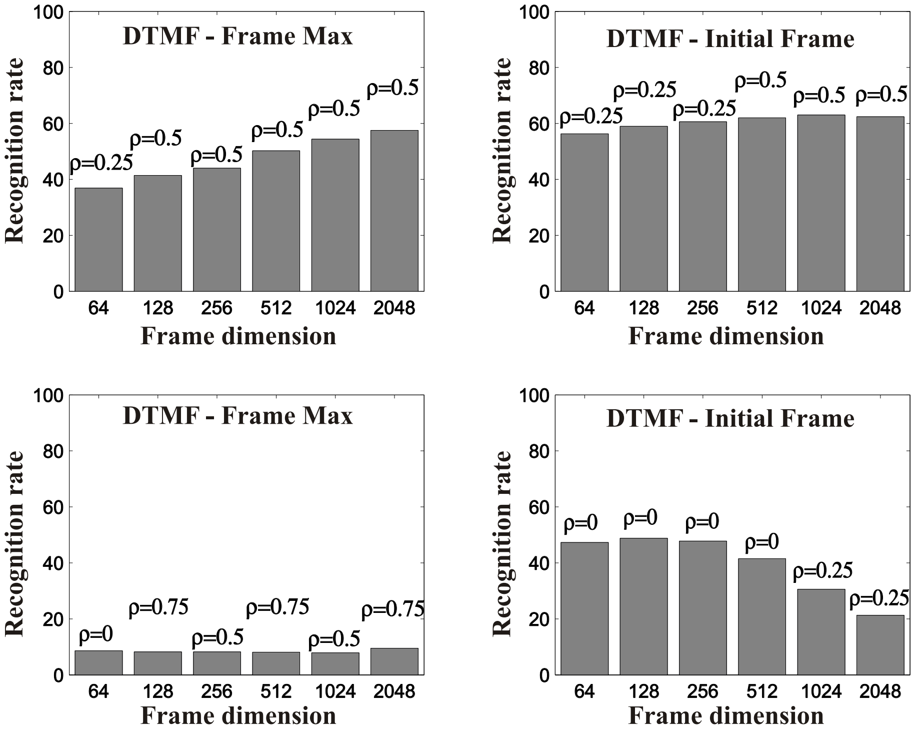


Fig. 7. TDOA results.

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 microphones. Coming back to Fig.6 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.

In Fig. 8 the average absolute localization errors obtained using geometrical localization are reported. The configuration that provides better results is the nr. 5.

- **From TDOA to source coordinates: acoustic localization.** We tested two approaches for acoustic localization. The first approach is based on a classical triangulation [21].

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

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

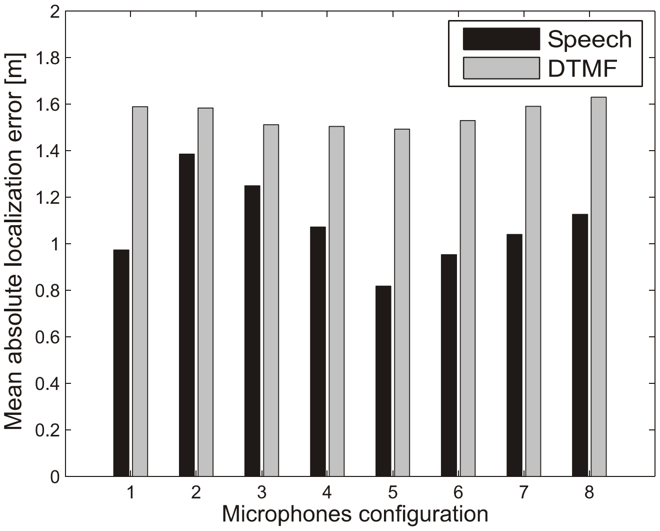


Fig. 8. Localization performance using different microphone configurations.

omnidirectional device into a grid, as shown in Fig. 9, 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. 10 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 interaction:

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

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

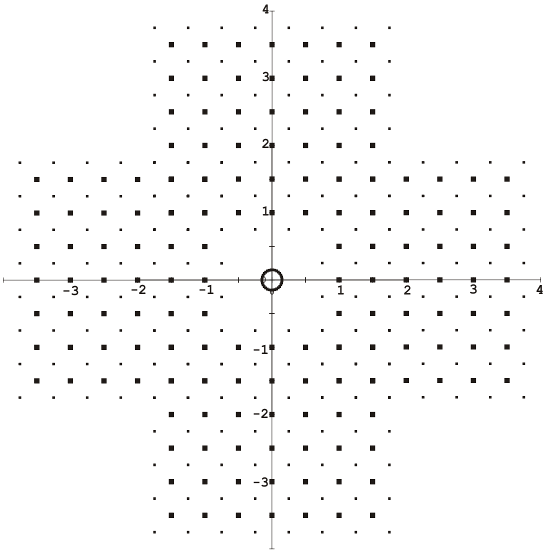


Fig. 9. Training grid plotting the position of the emitter in the training phase of the acoustic sensor. The big points correspond to real position of the emitter. Small points correspond to virtual position of the emitter.

Real signals were acquired in the big points on the grid of Fig. 9 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 if they were emitted in the small points of Fig. 9. In Fig. 10 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.

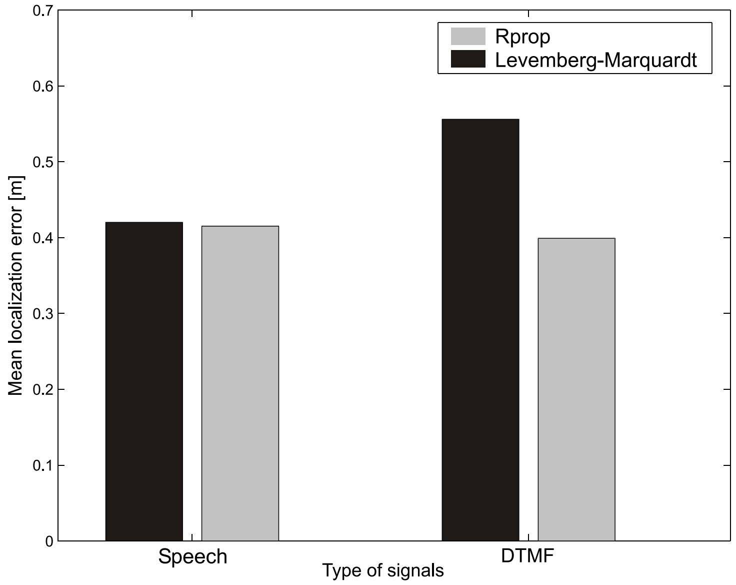


Fig. 10. Average performances of sound localization using Speech and DTMF tones

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 [21].

In Fig. 11 a comparison between geometric linear intersection localization and 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.

4.2. *Microphone array and beamforming*

4.2.1. *Preliminaries*

A sensor can be viewed as a window, called *aperture*, through which a field of certain physical quantities is measured. [13] 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

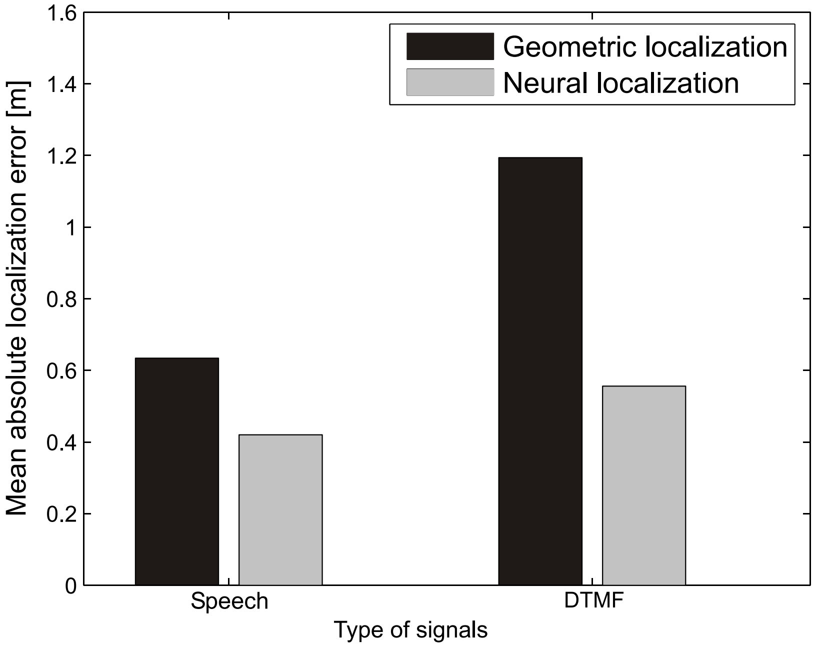


Fig. 11. Comparison between geometric linear intersection and neural localization.

spatial frequency, instead of the 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 $Z(\omega) = \sum_{m=0}^{M-1} Y_m(\omega) w_m s_M^*$.

4.2.2. *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

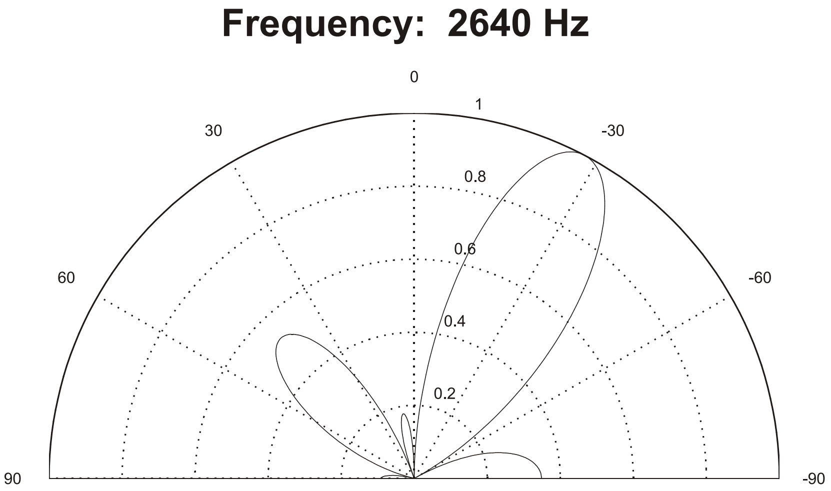


Fig. 12. The reception diagram obtained for the array of microphones once beamforming is performed.

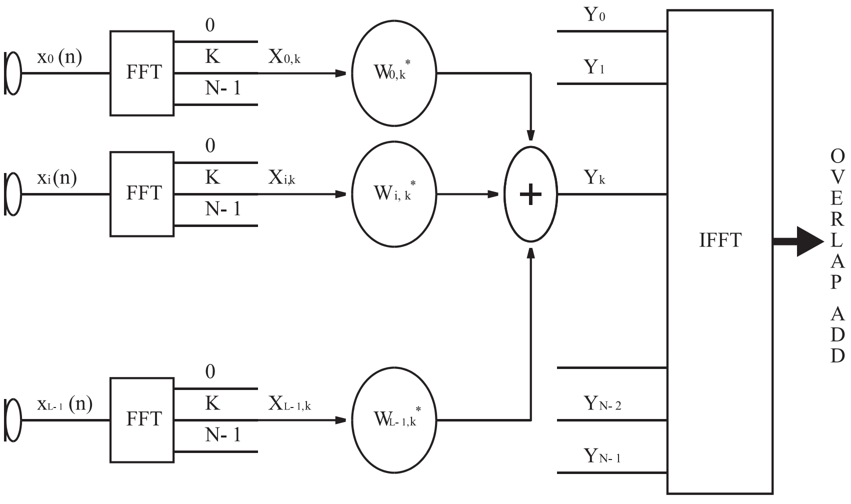


Fig. 13. A schematical representation of the beamforming algorithm.

combined in order to obtain a directional microphone. In Fig. 12 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. 13.

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 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.

4.3. *Speaker classification*

The acoustic signal obtained by beamforming is therefore cleaned up by most of the noise and can now be used to train an HMM (Hidden Markov Model) to recognize the speech of the talking person [19]. The learnt HHM can be used to identify the person while moving in the environment by his/her voice from another person, so allowing audio tracking of a walking person. A HMM can be trained to recognize an unknown voice in five acquisitions of the acoustic agent.

5. The sensor fusion module

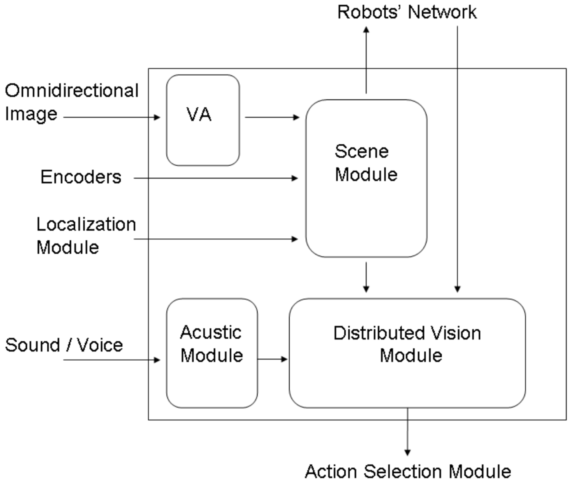


Fig. 14. The architecture of the sensor fusion module.

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 by Menegatti *et al.*[18,8]. This technique was developed to fuse position data coming from heterogeneous sensors. The only assumption on the measurements is that each measurement could be described as a Gaussian probability distribution and that each measurement is labeled with a time stamp indicating the time at which they were acquired. This system uses a modified Kalman filter to fuse the measurements 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 in the wrong time order, since they can be reordered thanks to the time-stamp associated to every measure. The architecture of the module performing the data fusion is sketched in Fig. 14.

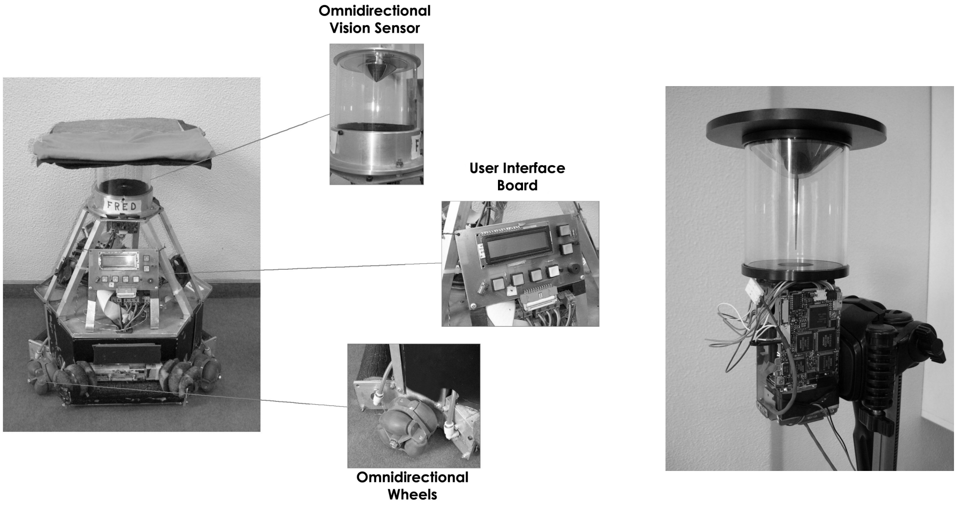


Fig. 15. (Left) The mobile robot on which the mobile vision agent is mounted. This is an holonomic robot with an omnidirectional vision system where the mirror has a custom profile. (Center) A close-up view of the principal robot's carachteristics. (Right) A close-up view of the omnidirectional camera of the static VA with the hyperbolic mirror. Note the two omnidirectional cameras have very different mirrors, so they produce very different images.

6. The Mobile Vision Agent

The mobile vision agent is implemented on board of a Golem platform developed by the Golem Team [9]. 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 [17].

The processing power is assured by a PC-104 with a AMD K6 400MHz CPU. As one can notice in Fig. 15, the omnidirectional camera on the mobile robot is very different from the omnidirectional camera mounted on the tripod (the SVA).

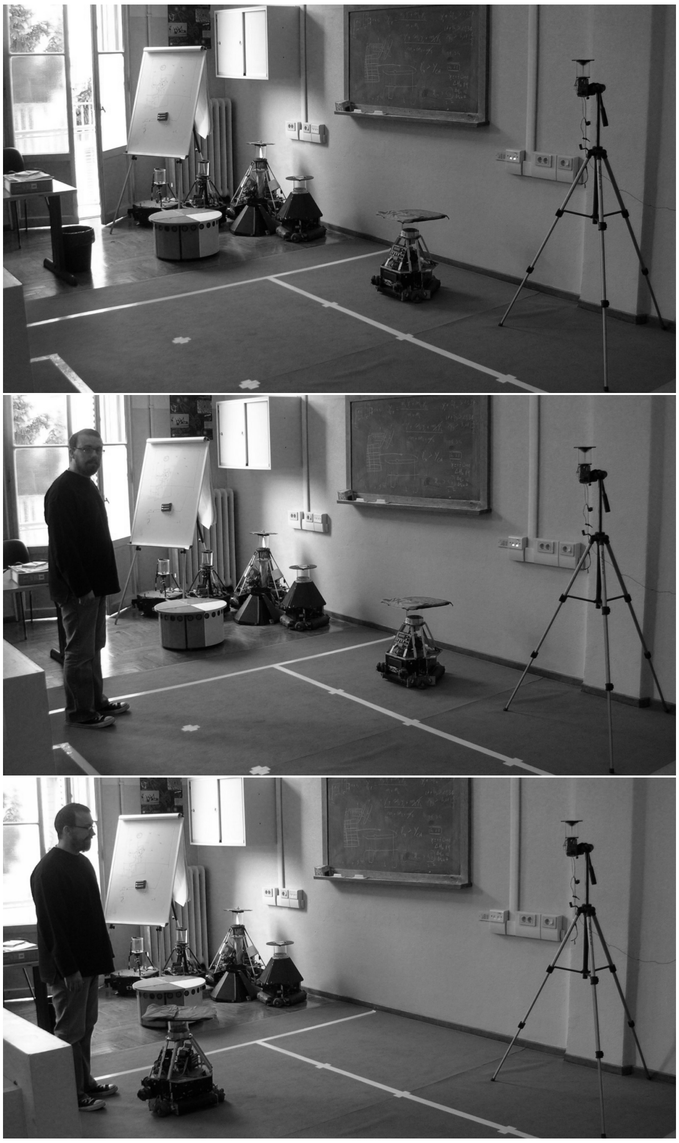
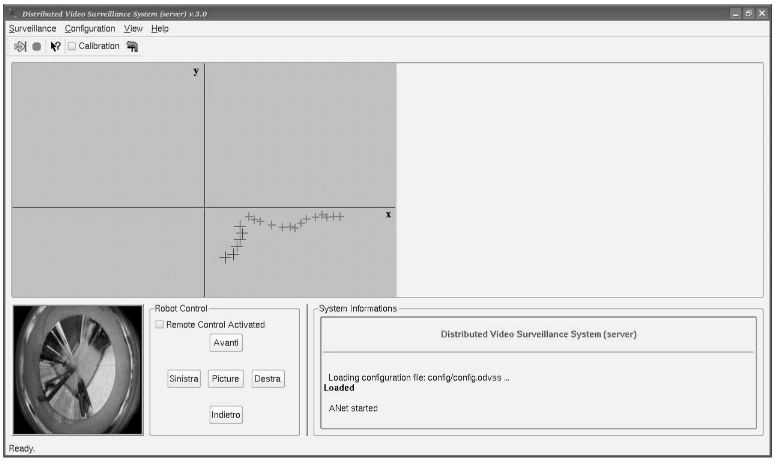


Fig. 16. Three pictures taken during the preliminary experiments: (Top) the robot is patrolling; (Middle) an intruder enters in the surveilled room; (Bottom) the robot approaches the intruder directed by the Static VA on the right of the picture and recognize it in its omnidirectional image.

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 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 has reached the position communicated by the SVA, it analyses its current images to identify the intruder. 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 displays it to the operator, as shown in Fig. 17.

7. Experimental results

For testing the data fusion and tracking system some preliminary experiments were performed. In the first one, an intruder enters the surveilled room from the left in Fig. 16. Once the position of the intruder is acquired, the mobile robot moves toward the intruder, and a close-up image of the intruder is grabbed and sent back to the monitoring station that can display it to the remote operator with the graphical interface depicted in Fig. 17. The graphical interface displays also the paths followed by the intruder and by the robot, obtained fusing the measurements of the different sensors.


Distributed Video Surveillance System (server) v.3.0
Surveillance
Configuration
View
Help
□ Calibration
y
x
Robot Control
□ Remote Control Activated
Avanti
Sinistra
Picture
Destra
Indietro
System Informations
Distributed Video Surveillance System (server)

```
Loading configuration file: config/config.odvss ...
Loaded

ANet started
```

```
Ready.
```

Fig. 17. A screenshot of the graphical user interface at the server. (Top left) The tracks of the intruder (light gray) and of the robot (dark gray) obtained fusing the measurements of the static vision agent and of the static acoustic agent. (Bottom left) The image grabbed by the mobile vision agent. (Bottom centre) Five buttons to remotely control the robot. (Bottom right) Status bar and system information display.

To evaluate qualitatively the performances of the different sensors and of the sensor fusion module a second experiment is presented in Fig. 18. In this experiment, the person is walking in a room making a loop of $4 \times 2m$. The position of the person along time is measured by two sensors: the Static Vision Agent and the Acoustic Vision Agent. The position of the person is calculated by the Static Vision Agent locating the position of the feet on the floor (triangles in Fig. 18) and by the Acoustic Vision Agent by locating the noise of the step of the person (circles in Fig. 18). The plot in Fig. 18 shows how both sensors are noisy and how several measurements underestimate or overestimate the distance of the person from the sensors. However, the fusion of the two kind of measurements and the integration in time performed by a Kalman filter produce a reliable tracking of the walker. In this experiment the walker moves at normal walking speed (about 1.3 m/s).

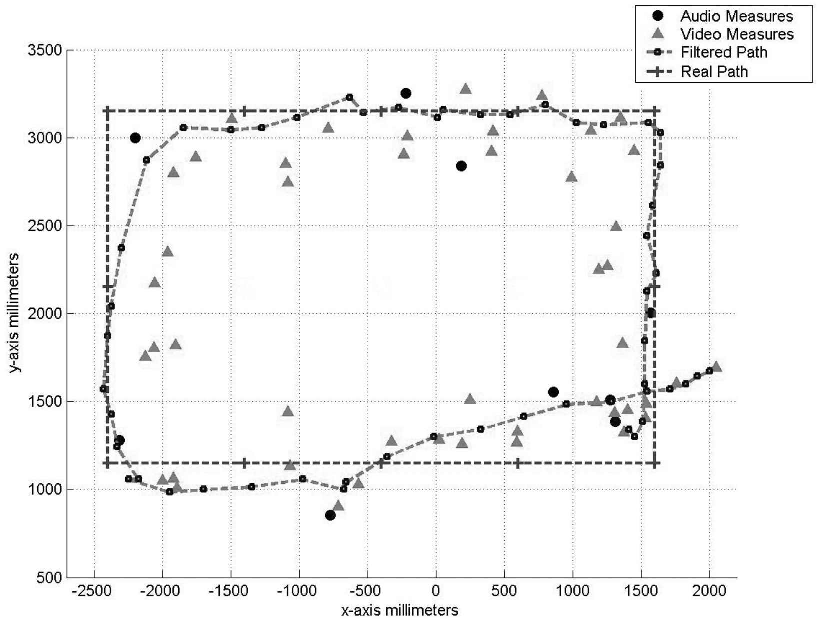


Fig. 18. The actual and the estimated path followed by a person walking in the environment. The static acoustic agent and the static vision agent are placed in the origin of the coordinate system.

8. Conclusion and future work

In this work, we presented an intelligent surveillance system able to autonomously monitor a room and to locate and 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 a static acoustic agent, but it has been designed in order to connect any number of sensory agents. The experiments reported in this paper were limited to qualitative tests of the system. More detailed experiments will produce a quantitative evaluation of the system. In addition, even if the experiments produced in this paper are limited to the tracking of one intruder, the system is designed in order to track several intruders at the same time.

In the current implementation the mobile sensorial agents can only transmit their data to the monitor, so a human operator can have a closer view of a particular location. Future developments will be devoted to the fusion of the sensorial data. One of the next steps will be to mount the omnidirectional acoustic sensor on the robot to have a robot fitted with an omnidirectional camera and an omnidirectional microphone. Moreover, the system is designed to integrate several mobile robots in order to have a team of several surveillance robots that can "go and seek" for different intruders.

9. Acknowledgements

The authors wish to thank: the students of the IAS-Lab, especially Nicola Milani, Nicola Brisotto, 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.

References

1. P. Aarabi and S. Zaky. Robust Sound Localization using Multi-Source Audio-Visual Information Fusion. *Information Fusion*, 2:209–223, 2001.
2. M. S. Brandstein and H. F. Silverman. A Practical Methodology for Speech Source Localization with Microphone Arrays. *Computer Speech and Language*, April 1997.
3. 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.
4. B. Chen, M. Meguro, and M. Kaneko. Probabilistic integration of audiovisual information to localize sound source in human-robot interaction. *Proceedings of the 2003 International Workshop on Robot and Human Interactive Communication*, 2003.
5. R. Collins, A. Lipton, and T. Kanade. A system for video surveillance and monitoring. Technical report, Robotics Institute at Carnagie Mellon University, 2000.
6. 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.
7. S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE IEEE Transactions on Multimedia*, 2(3):14–151, 2000.
8. E. M. E.Pagello, A. DAngelo. Cooperation issues and distributed sensing for multi-robot systems. *IEEE Proceedings of IEEE*, (in press).
9. 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.
10. D. Gutchess, A. K. Jain, and Sei-Wang. Automatic surveillance using omnidirectional and active cameras. In *Asian Conference on Computer Vision (ACCV)*, January 2000.

11. S. Haykin. *Neural Network A Comprehensive Foundation.* Macmillan College Publishing Company, New York, second edition, 1998.
12. 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.
13. D. H. Johnson and D. e. Dudgeon. *Array Signal Processing - Concepts and Techniques.* Prentice Hall, 1993.
14. C. H. Knapp and G. C. Carter. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Trans. ASSP*, ASSP-24(4):320–327, August 1976.
15. 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.
16. 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, March 2004.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. D. Rabinkin, R. Renomeron, J. French, and J. Flanagan. Estimation of Wavefront Arrival Delay Using the Cross-Power Spectrum Phase Technique. *J. Acous. Soc. Am.*, Vol. 100(N. 4 Pt. 2):2697, October 1996.
23. M. Riedmiller and H. Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In *ICNN*, 1993. San Francisco.
24. D. C. Schmidt. ACE: an object-oriented framework for developing distributed applications. In *Proceedings of the th USENIX C++ Technical Conference*, (Cambridge, Massachusetts), April 1994. USENIX Association.
25. D. C. Schmidt, D. L. Levine, and S. Mungee. The design of the TAO real-time object request broker. *Computer Communications*, 21(4), 1998.