A Multi Agent System for Audio-Video Tracking of a Walking Person in a Structured Environment.

Enzo Mumolo [1], *Massimiliano Nolich* [1], *Emanuele Menegatti* [2], *Enrico Pagello* [2]
[1] DEEI, University of Trieste, Italy
[2] DEI, University of Padova, Italy

Sommario

In questo articolo viene descritto un sistema intelligente per realizzare un inseguimento audio-video di una persona che cammina in un ambiente chiuso, come ad esempio una esposizione, una mostra o un museo. Il sistema è formato da alcuni dispositivi robotici ciascuno costituito da una schiera di microfoni e da una telecamera. L'algoritmo di inseguimento è in parte pilotato dalla percezione acustica. Le informazioni di inseguimento possono essere utilizzate per meglio organizzare le esposizioni stesse.

Abstract

In this paper an intelligent audio-video tracking system, composed by several robotic devices provided with microphone arrays and video-cameras, suitable for monitoring the presence of a walking person in a structured environment, is presented. The tracking is mainly driven by acoustic perception. The main contribution of this paper is a definition of an architecture for managing the data provided by each robotic device driven by acoustic perceptions.

Key words: *multi agent system, audio-video tracking, acoustic source localization, beamforming.*

1 Introduction

In this paper we deal with the following problem: is it possible to monitor the path followed by people visiting a showroom, a museum or an exposition? An answer to this question could be important to know what are the most visited products in an exposition, or how long a picture is viewed in a museum. Of course, this knowledge could be important to better organize the exposition of some sort of products or the placement of artistic objects in a museum. Also, a similar system can be very useful for example in surveillance applications. In this paper, we describe a system which is able to track the position of a person on the basis of audio-video measurements. Since the audio-video devices cover overlapping areas, a data fusion mechanism should be used to obtain the results. From this point of view, the system described in this paper is preliminary, in the sense that the devices are used in sequence, one at a time, as no data fusion is performed, leaving it to future development. When a person exits from the device's area of coverage, it is tracked by another device, namely the one whose coverage better detects the person. In the following, audio-video devices are called robots, and the showroom, museum or exposition are called generically "environments". The system presented in this paper allows to track the path of walking persons without requiring to transport intrusive devices.

Many researchers have attempted to integrate multiple senses. Most of their implementations process each sense separately and integrate the overall results in the final step. The system described in [1] 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 partecipate in the localization of objects. It is used simply to take images of the sound source after it has been localized. The system can be used for videoconferences. In [2] a multi-modal sound localization system is illustrated. The system utilizes two cameras and a 3-element microphone array. Test results show a significant improvement in the integrated vision and sound localization over that of the stand alone microphone array based sound localization system to accurately localize sound sources in low signal to noise situations.

The paper is organized as follows: in Section 2 the

scenario is described, while in Section 3 the tracking algorithm is summarized; Section 4 deals with some preliminary experimental results and in Section 5 some conclusions are given.

2 Scenario

In an indoor exposition, the system described in this paper automatically tracks the sounds produced by the movements of a walking person inside an environment. In this work we assume that:

- the environment is completely covered by a set of robots, each of them provided with microphone array and video-camera;
- each robot knows the absolute map of the environment and where it is placed;
- each robot communicates with the others through TCP/IP channel;
- a single visitor enters in the environment at a time;
- the floor is made by material which makes noise when a person walks;
- low background noise is present and is furthermore reduced by beamforming.

In fig. 1, a generic environment with three robots is shown. Roughly speaking, the areas of coverage of each robot are represented as gray tones. Clearly, the areas overlap to some extent; for example a section were the first two robots overlap is marked with 2 and a section covered by all the three robots is marked with 3. Hence, the robots are placed in the environment providing a coverage of the regions of interest. Initially the Robot A camera points to the entrance and waits for a movement which it is assumed to come from a person entering the environment. An acoustic localization algorithm is then started to localize the noise produced by the steps. After the localization, a beamforming algorithm is used to direct the microphone array to the acoustic source, i.e. to the person. Also the camera is moved towards the localized position. 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. When the person moves its position is tracked by acoustic localization. The microphone array and the camera are both steered towards the position and the acoustic signal is used to train an HMM.

The same procedure is performed when the person moves after a stop and when another person enters the

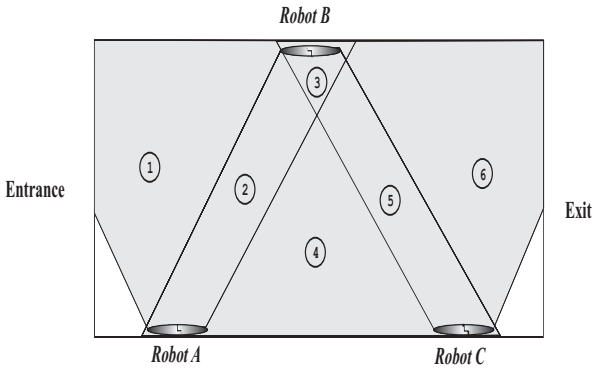

Figure 1: Three robot placed inside an environment and the areas covered.

environment. The HMMs trained so far are used to reveal with a given probability if the person is a different one. In this way the system can provide also the number of different persons visiting the environments. When a person moves from a stop – at the last localized position – its movements are detected by the videocamera, which, again, starts acoustic localization. The sequence of localized positions, with the related time stamps, are then stored for tracking purposes. The acquired audio-video can be transmitted in streaming for recording. When another robot detects the visitor, it asks the neighboors the parameters of the HMM for acoustic classification and the last known position in order to update its information.

3 Audio-video tracking algorithm

Each robot is an embedded device (fig. 2) composed by the following parts:

- a microphone array;
- a video camera that can be steered;
- a DSP based board for acoustic acquisition;
- a frame grabber;
- a PC based board for implementing the agent software.

The signal processing technologies adopted in each device are: neural algorithms for acoustic localization;

Figure 2: Robot setup.

beamforming in frequency domain for reducing directional noises; frame differences for detecting walker movements; Hidden Markov Model (HMM) for detecting the step sounds of the walker.

Each device communicates each other in order to cooperate in determining the whole path of the walker and saving the audio-video of his/her path.

The tracking system is provided with a linear array (composed of 8 microphones) and a video camera that can be steered. Each robot performs the tracking algorithm described in the following.

The tracking algorithm is mainly driven by acoustic perceptions as depicted in fig. 3. The main part of the

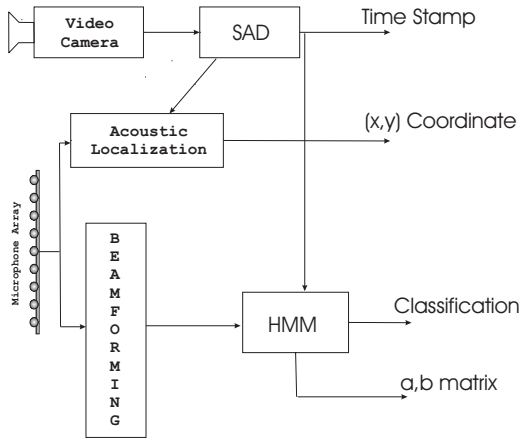

Figure 3: Audio-video tracking algorithm.

algorithm are:

- Step Activity Detection (SAD): the movement of the walking person are perceived using the video input;
- acoustic localization: the noise produced by the footsteps of the walker are localized using a microphone array;
- beamforming: the inputs of the microphone array are combined in order to obtain a directional microphone;
- HMM classification: the steps of a person are described with an HMM so that it is possible to infer the number of different persons visiting the environment.

Step Activity Detection. The Step Activity Detection is implemented revealing the differences between different acquisitions of the video camera. The difference between two acquisitions is computed, and if it is greather than a threshold, a movement is detected and the other computations of the tracking algorithm are performed.

Acoustic Localization. The localization algorithm is performed using a neural network based algorithm as described in [3, 4], that has as input the generalized cross correlation of signals acquired and as output the (x, y) coordinates of the acoustic source detected.

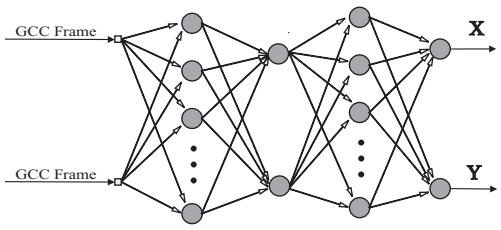

Figure 4: Block diagram of acoustic localization.

Beamforming. For removing directional noise, beamforming in frequency domain is performed using a 8 microphones linear array, obtaining a directional main lobe in the reception diagram as presented in [5]. In fig. 5 a reception diagram is reported; in this case the array is steered towards a -30 degree direction and the interferring noise coming from the broadside direction (0 degree) is de-emphasized. The beamforming algorithm is schematically depicted in fig. 6.

The adaptive algorithms for beamforming apply a vector of weights W to the vector of observations, that is the

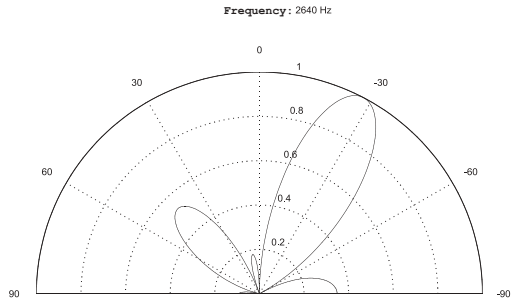

Figure 5: Example of the reception diagram of the microphone array.

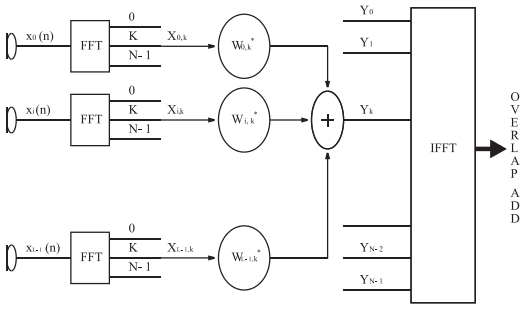

Figure 6: Beamforming algorithm.

signals coming from the microphones in the frequency domain, in order to minimize the mean square value of the weighted observations,

$$E[(w'y)^2], \tag{1}$$

subjected to some given constraint, described by

$$Cw = c, \tag{2}$$

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

$$R = E[yy']$$

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

$$w_{opt} = R^{-1}C' \left(CR^{-1}C'\right)^{-1} c. \tag{3}$$

Our frequency domain constraint, represented by equation (2) consists in a signal emphasys 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. Besides emphasizing the signal, the beamforming aims at reducing noise and reverberation.

HMM Classification. The signal acquired by the microphone array and enhanced using beamforming is feeded into a HMM classification algorithm, described in [5], that detects if the steps of the walker are unknown, using a threshold on the probability output of the HMM. If the walker is a new one, the HMM parameter are trained for the following 5 acquisition of the robot. If the walker is known or already trained, the classification result is provided by this block.

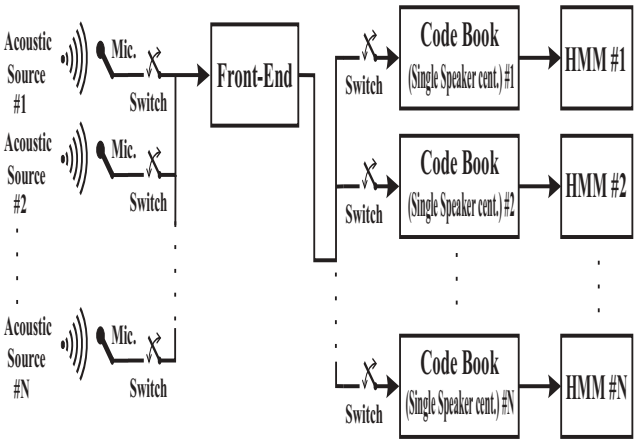

Figure 7: HMM training and execution.

Multi-agents communication The system is composed by agents that communicate the following data:

- a time stamp;
- (x, y) coordinates of the detected person;
- the class of the detected person;
- the parameters (two real matrices) of the model of the new detected person.

The agents should perform a data fusion among the stored data ([6, 7]).

4 Experimental results

For testing the tracking system some sample signals were acquired.

An example of a visitor path revealed by the tracking system is depicted in fig. 8. The time stamps acquired during the tracking are printed over the path; in the case of fig. 8, the visitor moves slowly (mean velocity of about 0.5 km/h) but continously from the entrance to the exit of the environment.

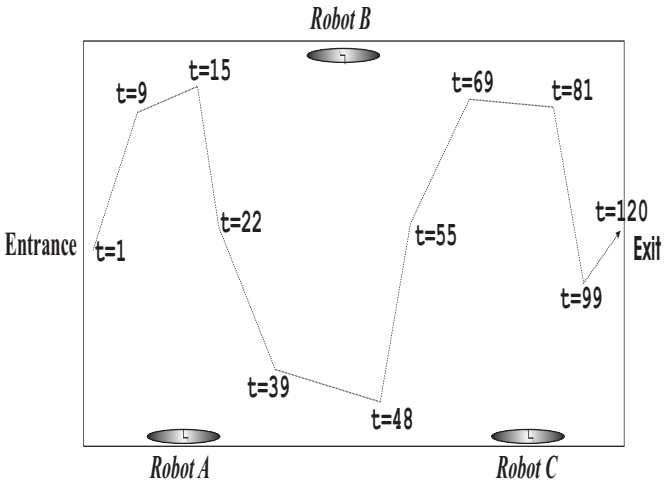

Figure 8: Example of trajectory computed by the multi-agent tracking system.

5 Conclusion and future works

The tracking system presented in this paper is based on a pipeline computation. Future developments concern the fusion of the sensorial data provided by several robot in order to build a more accurate trajectory of the walker, using deeper interaction between the software agents.

References

[1] D.V. Rabinkin, R.J. 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.

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

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

[4] Enzo Mumolo and Massimiliano 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.

[5] Enzo Mumolo and Massimiliano 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.

[6] Emanuele Menegatti, Alberto Scarpa, Dario Massarin, Enrico Ros, and Enrico Pagello. Omnidirectional Distributed Vision System for a Team of Heterogeneous Robots. In *Omnivis 2003: Workshop on Omnidirectional Vision and Camera Networks*, June 2003.

[7] Enzo Mumolo, Massimiliano Nolich, Gianfranco Fenu, and Ervin Ceperic. Kalman Data Fusion in Robot Simulation using a Neural Network Model of System Dynamics. In *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, June 8-11 2003.