EC-VIP-MC 2003, 4th EURASIP Conference focused on
Video / Image Processing and Multimedia Communications, 2-5 july 2003 Zagreb, Croatia

IDENTIFICATION OF SOUND SOURCE IN MACHINE VISION SEQUENCES USING AUDIO INFORMATION

Abedin Vahedian
vahedian@irost.com
Iranian Research Org. for Science & Technology
Mashad Center,
Mashad, Iran

Abstract: *Identifying the sound source location in the acquired image for further processing is of concern in image/video processing especially when tracking objects of vibrating or sound making nature in the scene becomes necessary. On the other hand relying on the picture content to identify such objects is not an easy task especially when there are similar objects in the acquired image. In this paper, we present a technique for identifying the location of sound source using beam-forming techniques by means of an array of microphones based on the processing of audio from the source itself. Upon estimation of the location in 3-D space, a projection technique is used to accurately locate the position of sound source in 2-D image frame coordination.*

Keywords: *Source location, Beam-forming, Array processing, Microphone array, Audio processing*

1. INTRODUCTION

In processing of video images, identification of objects is as important as identification and extraction of patterns in the image, as much often it becomes necessary to identify and track a specific object in the picture in real-time fashion. Detection and/or estimation of the location of sound source or sound making objects in the acquired image are of great concerns as well. This often becomes a major requirement in applications such as machine vision, videophone, videoconference, security and monitoring.

Much research effort has been addressed at detection and position location problems. Most has attempted to use properties of the picture to identify the location of a specific object, which is then used to enable the intended system to track or apply further and complementary processing [1,2]. These techniques tend to be computationally very complex and make assumptions about the content of an image. It turns out that to accurately locate a point sound source, using the video information alone is a very difficult problem. Taking into account that in many cases the sound-making object may not have a characteristic in terms of image attributes or extractable pattern. In addition, in times the object could become hidden or covered in the captured scene or even appear within similar objects of its type and shape which make the identification of the sound making one from the others highly difficult. On the other hand, it is possible to use an e al source of information, namely the sound

information emanating from the source to identify its location in the acquired image by camera.

The sound-assisted technique presented in this paper is based on the idea of beam forming. Hence, by combining the sound and video information, it is possible to achieve a significant increase in identification of these types of objects without a considerable expense [9]. The approach implemented thus far, has been to use an array of microphones to locate the source of the sound using direction of arrival schemes. This new technique, which is in harmony with all existing video processing standards, has the potential to be easier to implement than previous techniques.

In the next section, the concept of using a microphone array in this particular application is explained. The implemented design to carry out the experiments is then presented. Performance evaluation in using the audio information is then discussed. Finally, the paper is concluded with the future work program toward the complete implementation of the proposed system.

2. IDENTIFYING OF SOUND SOURCE USING AN ARRAY OF MICROPHONES

In an audio DSP system, the directivity of a microphone array can be used to pick up sound emitted by a distant source while suppressing noise and reverberation arriving from other directions. There are many direction finding techniques and all these techniques exploit the fact that the time taken by a signal emitted by a source to reach different microphones is different due to the spatial spread of the microphones [3, 4, 5]. The position of a sound source may be located by considering different points on a spiral starting at the outer boundary of the area under consideration and measuring the delay in arriving from each point using the optimal delay estimation methods. Fig.1 shows a general set up.

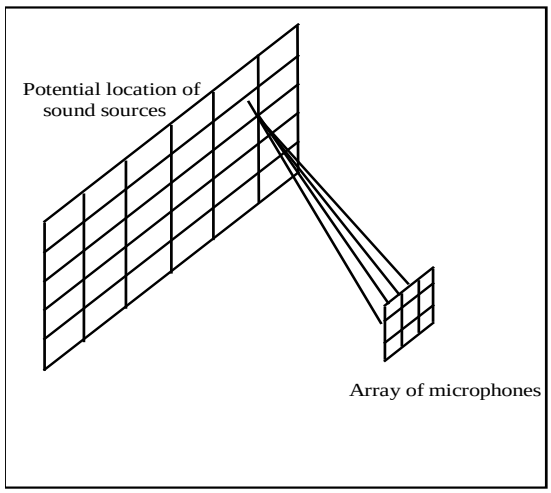

Fig.1 Array of microphones used to cover the area of interest.

The scheme uses the time difference of arrival (TDOA) estimates between the source and a spaced array of microphones forming a] ray. This difference is easily converted to

the range difference since sound waves propagate at a relatively constant speed, at least inside a larger room. Therefore, each TDOA measurement between two microphones determines that the position of sound source must lie on a hyperboloid with a constant range difference between the two microphones. Fig.2 depicts the solution where two hyperboloids are formed from TDOA measurements at three fixed microphones to provide an intersection point that locates the position of source. The equation of this hyperboloid is given by:

$$R_{i,j} = \sqrt{(X_i - x)^2 + (Y_i - y)^2 + (Z_i - z)^2} - \sqrt{(X_j - x)^2 + (Y_j - y)^2 + (Z_j - z)^2} \quad (1)$$

with $R_{i,j}$ being the range difference between the microphones *i* and *j*. The coordinates (X_i, Y_i, Z_i) and (X_j, Y_j, Z_j) represent the two fixed microphones and make up the unknown coordinate of the sound source position.

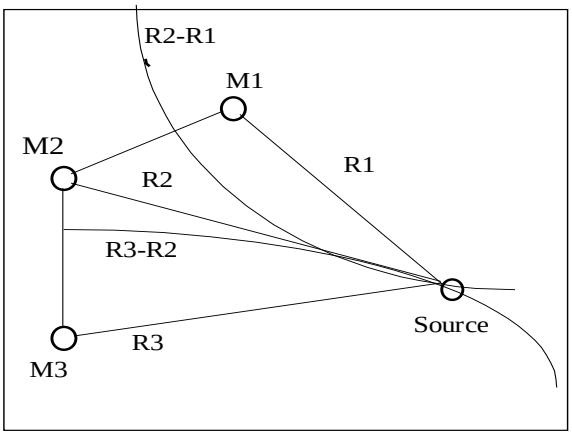

Fig.2 Hyperbolic position location with 3 microphones for a 2-D solution.

In the general case, in which the 3-D position of the source is required, at least four independent measurements need to be made.

The technique is implemented in two stages: First, estimates for TDOA values are computed from sound signals. Then estimated TDOA values are processed to determine a location estimate.

These steps are now explained:

1-*Computing TDOA Estimates:* To evaluate the hyperbolic range equations, we need to obtain the estimate of the range differences, or equivalently the TDOA.

A signal emanating from a distance source and monitored in the presence of noise at two spatially separated microphones can be mathematically modelled as:

$$\begin{aligned} x_1(t) &= s_1(t) + n_1(t) \\ x_2(t) &= \alpha s_1(t + D) + n_2(t) \end{aligned} \quad (2)$$

where the signal and noise are assumed to be real, jointly stationary random processes. Signal *s1*(t) is also assumed to be uncorrelated wit ɔise signals.

One common method of determining the time delay (D) and hence, the arrival angle relative to the microphone axis is to compute the cross-correlation function [6]

$$Rx_1x_2(\tau) = E[x_1(t)x_2(t-\tau)], \tag{3}$$

where E denotes expectation. The argument that maximises (3) provides an estimate of delay. This implies that in implementation, the A/D and the data logger must share a precise time-base with the reference signal but does not impose any requirement on the signal transmitted by the source. Because of the finite length of recorded samples, however, $Rx_1x_2(\tau)$ can only be estimated. In order to improve the accuracy of the delay estimate, it is necessary to prefilter x_1 and x_2 prior to the cross-correlation process. An estimated cross-spectral density function can also be computed in the frequency domain, and then the estimated cross-correlation function is obtained via an inverse Fourier transform. The later approach has been adopted in this study as the frequency domain processing lends itself well to filtering of the signal prior to computation of the cross-correlation function.

The Phase Transform (PHAT) processor has been used for this purpose:

$$\hat{R}x_1x_2(\tau) = \int_{-\infty}^{\infty} \frac{\hat{G}x_1x_2(f)}{|Gx_1x_2(f)|} e^{j2\pi f\tau} df. \tag{4}$$

where the expression $Gx_1x_2(f)$ denotes the cross power spectrum of each pair of signals. For the model in (2) with uncorrelated noise (i.e. $Gn_1n_2(f) = 0$),

$$|Gx_1x_2(f)| = \alpha Gs_1s_2(f). \tag{5}$$

and for a relatively long time monitoring, when
$\hat{G}x_1x_2(f) = Gx_1x_2(f)$, then

$$\frac{\hat{G}x_1x_2(f)}{|Gx_1x_2(f)|} = e^{j\theta(f)} = e^{j2\pi fD} \tag{6}$$

and the cross-correlation estimation approaches a delta function. In practice, performing the above delay estimator on a pair of relatively long audio samples (compared to the sample delay expected) leads to a peak in the cross-correlation function.

2-Position Estimation by solving the Hyperbolic Equations: Once a reliable estimate for TDOA values has been computed, it is substituted into the hyperbolic equations, which are then solved for the Cartesian coordinate of the sound source, with respect to a set of objects (i.e. the coordination of microphones, or one reference microphone assumed on the (0, 0, 0) location).

Solutions for hyperbolic and related position fixes have attracted attention for applications in mobile stations and navigation systems [7,8]. In machine vision application, however, there is

not only the challenge for a resolution of, for example ±3cm within a whole of scene range of ±1*m*, but also the processing time of the algorithm is of importance. However, there exist assumptions on the dimensions of the area of interest for inspection/monitoring which allow us to reduce the complexity of the computations to a linear level, as the hyperbolic equations are non linear and their solution is nontrivial, particularly when range estimates may be inconsistent due to the noise or multipath arrival or echo.

The position estimate from the above two step procedure, is now converted to the image properties in terms of pixels in vertical and horizontal directions where one single point within the picture is determined as being the location of the centre of the sound source. Subsequent segmentation is carried out to implement any required processing once a picture cell is assumed to be the sound making object's center.

3. IMPLEMENTATION OF THE EXPERIMENT

A rectangular array of four microphones spaced *20* cm from each other has been used in this work. The camera is located in the center of the microphone plane such that its focal plane and the microphone plane are the same as shown on Fig.3. A digital audio/video recorder facilitated the recording of video sequences synchronised with the audio. Four audio channels have then been transferred to the computer through a *4*-channel simultaneous A/D data acquisition card. A *4* kilo sample section of each audio channel was used in the TDOA estimation procedure. At a sampling rate of *32ks/s*, a time delay equivalent to one sample then represents a range difference of approximately 1*cm* between two microphones, which in turn represents a range difference of about *5cm* along the scene where a 5-cm speaker playing a 2-khz tone was placed in the scene as the sound source at a distance of *1m* from microphone array.

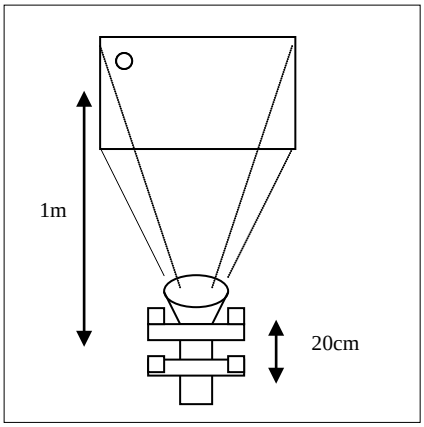

Fig.3 Experiment set-up

4. RESULTS AND DISCUSSIONS

In the audio processing section, using PHAT provides reliable TDOA estimation results. When the sound source was displaced in the scene or the camera made panning over the scene, a displacement of ±5cm was detected. The projection task performed in software according to the known geometry and the coordination reference also made a fairly reliable identification. A video sequence recorded on tape will accompany this paper that clearly demonstrates the impact of the position estimation algorithm on the video sequences.

Further work to be carried out as part of this project includes:

* techniques to analyse the audio data so as to achieve the best minimum resolution in terms of detectable range in a scene of interest.
* cancelling echo and reverberation and resolving multipath arrivals of signal using beam-forming techniques.
* increasing the sensible range in both the vertical and horizontal directions.

5. CONCLUSION

A new idea has been demonstrated to exploit the relationship between sound and video in machine vision applications. It has been shown that the performance of source identification task can be improved with little or no increase in the overall complexity. The continuation of this work can open up a new dimension in using every possible correlation between sound and video information in video based applications.

REFERENCES

[1] R. Mech and P. Gerken, “Automatic Segmentation of Moving Objects”, Doc. ISO/IEC JTC1/SC29/WG11 MPEG97/1949, Bristol, UK, April 1997.
[2] B. Falcidieno and T. L. Kunii (Eds), “Modeling in Computer Graphics”, Springer, Berlin, 1993.
[3] B. Ottersten and T. Kailath, "Direction of Arrival Estimation for Wide band signals ...", IEEE Trans, ASSAP, Vol. 38, No. 2, Feb. 1990, pp. 317-327
[4] S. R. de Graaf and D. H. Johnson, "Capability of Array Processing Algorithms to Estiamte Source Bearing", IEEE Trans. ASSAP, Vol. 33, No.6, Dec. 1985, pp. 1368-1379.
[5] Barry D. Van Veen and Kevin M. Buckley , "Beamforming: A versatile Approach to Spatial Filtering", IEEE ASSP Magazine April 1988.
[6] C. H. Knapp "The Generalized Correlation Method for Estimation of Time Delay", IEEE Trans. ASSP, Vol. ASSP-24, No.4, August 1976.
[7] B. T. Fang "Simple Solutions for Hyperbolic and Related Position Fixes", IEEE Trans. Aero. and Elec. Sys., Vol. 26, No. 5, Sep. 1990.
[8] T. S. Rappaport et al, "Position Location Using Wireless Communications on Highways of the Future", IEEE communications Magazine. October 1996.
[9] A. Vahedian, et al, “Improving Videophone Subjective Quality Using Audio Information”, International Journal of Imaging Systems and Technology, Vol 10, 86-95 (1999).