Eyes ‘n Ears: A System for Attentive Teleconferencing

B. Kapralos[1,3], M. Jenkin[1,3], E. Milios[2,3] and J. Tsotsos[1,3]

[1]Department of Computer Science, York University, North York, Canada M3J 1P3
[2]Department of Computer Science, Dalhousie University, Halifax Nova Scotia, B3H 1W5
[3]Centre for Vision Research, York University, North York, Canada M3J 1P3

{billk, jenkin, tsotsos}@cs.yorku.ca, eem@cs.dal.ca

Abstract

Various teleconferencing systems exist, including systems intended for multiple speakers in a conference setting. In such a multiple-speaker setting, a speaker must be localized and tracked in both the video and audio domains. Although many fast, reliable and economical video trackers capable of tracking humans exist, there are very few compact, portable and economical audio localization systems. On the contrary, most available audio localization systems are expensive, non-portable and require extensive audio arrays requiring substantial computational processing. Under the Eyes ’n Ears project, a simple, economical and compact method of sound localization for use in a teleconferencing system is being investigated. This paper describes the current status of the Eyes ‘n Ears project, and summarizes the hardware and software components that make up the system.

Introduction

Video teleconferencing has found a wide range of applications; from facilitating business meetings to aiding in remote medical diagnoses. Various commercial teleconferencing systems exist, including basic static systems for use by two participants (one at each end of the connection). There are also systems intended for multiple speakers (i.e. as in a conference setting) but these systems typically focus on a single user and provide limited, if any, automatic speaker tracking technologies.

Existing systems suffer from a number of limitations. Essentially, they provide a limited number of static or manually tracked views. As a consequence, in a multiple speaker setting, a speaker must either move into the camera's view or the camera must be manually commanded to track the speaker. Furthermore, in addition to video, teleconferencing systems must be able to capture and transfer audio (e.g. speaker’s voice). As a result, in a multiple speaker setting, the teleconferencing system must be able to localize a speaker. However, with the multiple speaker systems currently available, audio is not focused on the speaker. Although sound localization systems are available, many require extensive audio arrays [5]. Furthermore, integration with video is difficult especially in a multiple speaker setting.

Our research investigates the development of a teleconferencing system integrating both audio and visual cues. Our goal is to develop an affordable, limited maintenance and portable teleconferencing system capable of locating and tracking a speaker in a multiple speaker setting.

Description

Figure 1 below illustrates the Eyes ‘n Ears hardware set-up.

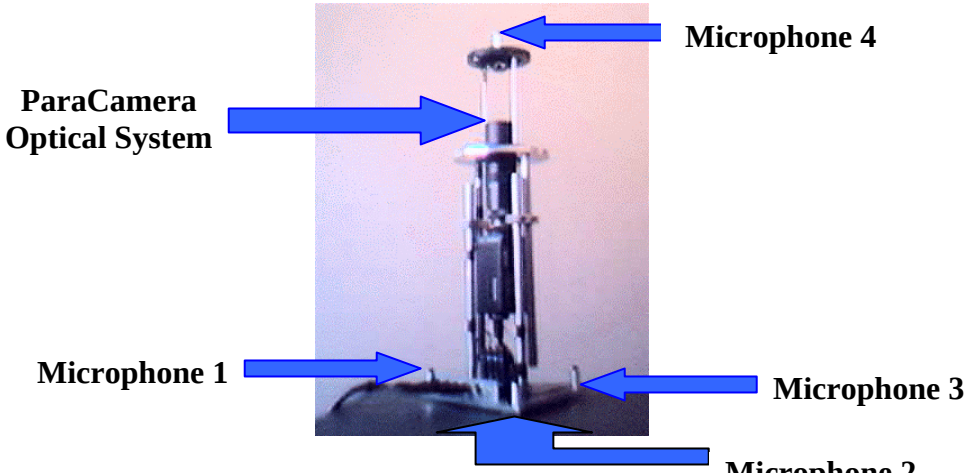

Figure 1. Eyes ‘n Ears Sensor

The following sections describe the hardware components in further detail.

Video System - ParaCamera

Typical camera lenses capture only a narrow field of view. To increase the visual field of the sensor, Eyes ‘n Ears utilizes Cyclovision's ParaCamera optical system. As shown in figure 2 below, the ParaCamera allows us to capture the entire hemisphere from a single viewpoint thereby providing multiple dynamic views. Once the hemispherical view has been obtained, it may be un-warped producing a panoramic view (figure 3). From this panoramic, perspective views of any size corresponding to different portions of the scene may be extracted easily (figures 4a, 4b).

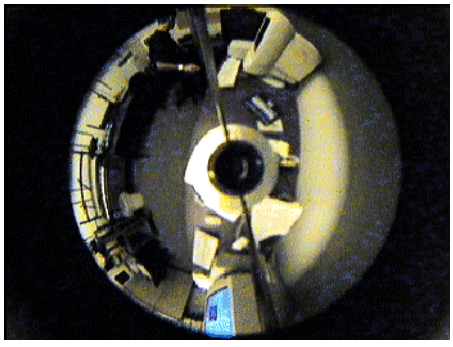

Figure 2. **Hemispherical View**

Figure 3. Panoramic View

Figure 4a
Perspective View

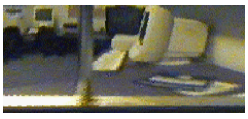

Figure 4b
Perspective View

Video Tracking

A good economical detection/tracking system must be able to locate the desired object quickly and reliably in the presence of noise and other objects in the environment. In

addition, it must run fast and efficiently, thereby tracking objects in real time, and run using inexpensive camera equipment [1].

The color of an object may be used as an identifying feature, which is local to the object and largely independent of the view and resolution. As a result, the use of color information may be used to detect objects from differing viewpoints. [8]. Furthermore, there are various fast and simple color based tracking systems available (see [4]).

Due to the considerations listed above, video tracking in our system is performed primarily using color information. Initially, a model is selected from an image obtained by the ParaCamera. The RGB intensity values of the model are converted to the Hue, Saturation and Value (HSV) values [2] thereby minimizing the negative effects introduced by changes in lighting conditions. A two dimensional histogram of the hue and saturation values is then computed (value is ignored as any changes in lighting will primarily correspond to changes in value). Once the model has been selected and its histogram computed, successive hemispherical images are obtained, image differencing is performed between their intensity differences to determine the regions of change due to the moving object(s). The sequence of images below, illustrates this process.

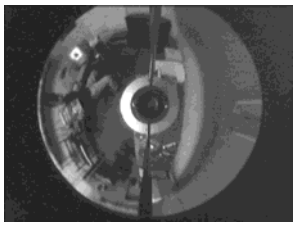

Figure 5. **View 1**

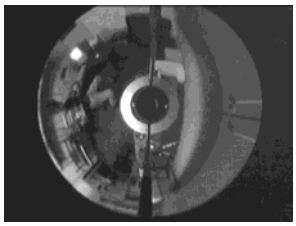

Figure 6. View 2

Region of Change

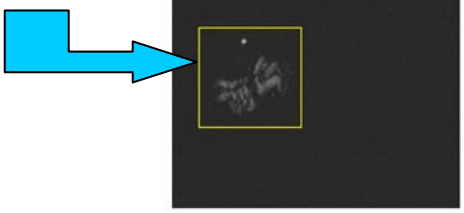

Figure 7. **Difference Image |View2 – View 1|**

Figure 8. **Threshold Applied to Difference Image**

Using a modified version of Histogram matching [7], a search for the model is performed within this bounded region of change. When the model is found, the region of the hemispherical image containing the model is un-warped thereby providing a perspective view of the model.

Audio System

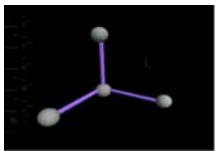

Figure 9
Microphone Set-up

Four omni-directional microphones mounted in a static pyramidal shape (see figure 9 to the left), about the base of the ParaCamera provide an economical and portable acoustic array capable of localizing speakers in 3-space [3]. Using beam-forming techniques, the audio system will be able to localize a speaker. Once the speaker's location has been determined, we may immediately obtain a

perspective view (from the hemispherical image) of the region including the speaker. Once the camera has focused on the speaker, they will be tracked in both the audio and video domain.

Sound Localization

Our sound localization system relies on beam forming techniques based on Interaural Time Difference (ITD) measurements between microphone pairs (baseline) to localize a sound source.

As shown in the figure 10, the ITD value of a single baseline, will place the location of the sound source to anywhere on the surface of a cone [6]. (*Cone of Confusion*). Each baseline will provide its own Cone of Confusion. By performing the intersection of three cones, Reid *et. al*, have determined the location of a sound source in 3-space fairly accurately [6].

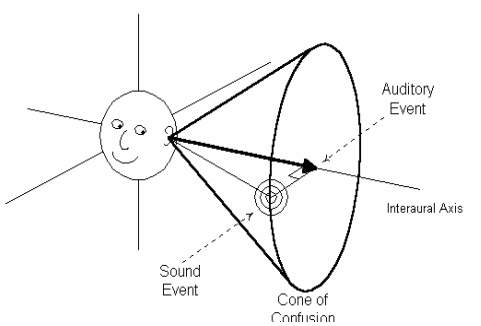

Figure 10.Cone of Confusion [8]

Sound Localization – Current Status

All hardware and software issues regarding the simultaneous input of sound from the four microphones have been resolved. We are currently capable of detecting a sound source on all four microphones, filtering the sound to remove noise and calculating the ITD value associated with each baseline using cross correlation.

As an example, a sound source ("Dropping Middle C" generated with *Sound Effects* on an Apple Power PC) was placed at an equal distance from each of the two microphones of a single baseline. As figures 11 and 12 below illustrate, both microphones detected the same sound. Furthermore, as expected, the plot of the correlation values (figure 13 – Time Shift in number of samples vs. Correlation value), indicate there is no time shift between the signals received by each microphone as the maximum value returned by the correlation function occurs at a time shift of zero.

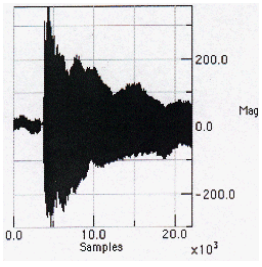

Figure 11.
Signal at Microphone 1

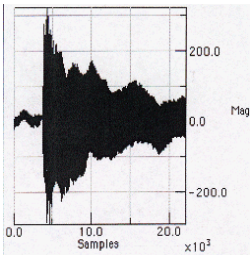

Figure 12
Signal at Microphone 2

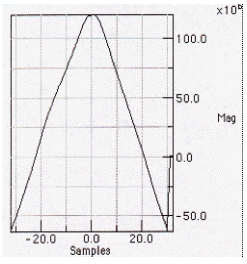

Figure 13
Time Shift vs. Correlation Value

Current Status

This paper has described a multi-speaker teleconferencing system, which will be able to

localize a speaker in a multiple speaker setting using both video and audio cues. Although the system has not yet been completed, progress has been made. A color-based tracker capable of tracking objects as well as humans in the video domain has been developed. Furthermore, progress improvement is evident with regards to an audio localization system capable of locating a speaker in 3-space using ITD values. We are currently capable of detecting a sound with all four microphones, filtering the sound to eliminate unwanted noise and performing correlation between the sounds received by the two microphones of each baseline.

Future research will focus on automating the video human detector. Rather than manually selecting a model, the system will be able to automatically detect humans using color information. In addition, the audio localization system will be completed, thereby allowing the location of a speaker to be determined in 3-space. In order to accomplish this, beam-forming techniques will be used. We are also experimenting with a similar approach to Reid *et. al*, whereby the location of a sound source is determined by geometrically taking the intersection of each baseline's cone of confusion.

Finally, once the sound localization system has been completed, we will integrate both audio and visual cues to allow tracking of a speaker in both the audio and video domain.

References

[1] Bradski, R. Gary. (1998). *"Computer Vision Face Tracking for Use in a Perceptual User Interface"*. http://developer.intel.com.

[2] Foley, James D., Andries Van Dam, Steven K. Feiner and John, F. Hughes. (1996). "Computer Graphics Principles and Practice". Addison-Wesley Publishing Company. USA.

[3] Guentchev, K. Y. and John J, Wong. (1998). *Learning-Based Three Dimensional Sound Localization Using a Compact Non-Coplanar Array of microphones.* American Association for Artificial Intelligence.

[4] Herpers, R. G. Verghese, K. Derpanis, D. Topalovic, J.K. Tsostos. (1999). "*Detection and Tracking of Faces in Real Environments*". Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems. Korfu Greece. September 26-29 1999.

[5] Rabinkin, D. (1996). *A DSP Implementation of Source Location Using Microphone Arrays.* In 131st meeting of the Acoustical Society of America. Indiana USA. May 15 1996.

[6] Reid L. Greg. *Active Binaural Sound Localization: Techniques, Experiments and Comparisons*. Master of Science Thesis. York University, Department of Computer Science. April 28, 1999.

[7] Swain, J. Michael and Dana H. Ballard. (1991). "Color Indexing". International Journal of Computer Vision. Volume 7, pp. 11-32.

[8] West, James R. (1998). *Five Channel Panning Laws: An Analytical and Experimental Comparison.* Master of Science in Music Engineering Technology Thesis. Faculty of Music. Coral Gables, Florida.