A System for Synchronous Distance Learning

B. Kapralos[1,3], A. Barth[2], J. Ma[1] and M. Jenkin[1,3]

[1]Dept. of Computer Science, York University, Toronto, Ontario, Canada. M3J 1P3
[2]Dept. of Computer Science, University of Applied Science, Sankt Augustin, Germany
[3]Centre for Vision Research, York University, Toronto, Ontario, Canada. M3J 1P3

{billk, jenkin}@cs.yorku.ca

Abstract

Providing an effective mechanism for an instructor to interact with one or more remote classrooms is a complex problem in synchronous distance learning. Although a human operator at the remote classroom could be used to direct a video camera and hence provide feedback to the instructor as to the state of the class and to coordinate the instructor's attention to the class as questions are raised, more automated techniques are desirable. Here we describe a novel audio-visual sensor to support distance learning. This sensor provides the instructor with a panoramic view of the class and can automatically attend to students who signal their intent to interact with the instructor either by raising their hand or by asking a question vocally. This paper describes the sensor and the user interface that is provided to the instructor so (s)he can interact with the students in the remote class.

Keywords: distance learning, panoramic sensing, active vision.

1 Introduction

A central issue in developing synchronous distance learning technology, especially with students and instructors at a number of different locations, is enabling the remote classes and the instructor to interact with each other. There are really two parts to this problem. How to present the instructor to the remote classrooms, and how to present the remote classrooms to the instructor. The first part of this problem is perhaps the easiest to solve, as there is only one person (the instructor) who must be attended to. Attending to students in the remote classrooms is more difficult. Issues such as "How does a student 'raise his/her hand' for attention?" or "How can the instructor 'select' one student to converse with, and how to attend to the student once (s)he has been selected?" are complex problems which must be addressed if the classes and instructors are to interact in an effective manner. Providing human facilitators at each site is not cost effective and the option of physically wiring each seat with buttons for students to indicate that they have a question would require significant modifications to existing classroom spaces. An alternative would be to deploy a sensor system within the classroom that enables student interaction with the instructor and other classes. But how should the sensor attend to the person who wishes to ask a question? From a practical point of view, how should a sensor be constructed which has a wide enough field of view so that it can capture the entire class at once and still be able to attend to the person who wants to speak or ask a question? In addition, once a speaker has been selected, how should the sensor continue to track, localize and focus on the selected speaker?

In this paper we describe an ongoing research project that is investigating issues related to the development of a distance learning system that permits a remote classroom to interact with the instructor. This includes issues related to attending, in both the audio and visual domain, to individual students, finding students who wish to speak, permitting the instructor to view the entire remote class and to attend to audio and visual events within it. To this end, we have developed a novel sensor that combines directional audio and a panoramic sensor to locate students in the classroom who wish to interact with the instructor (e.g. ask a question), and signal their intent to do so using hand raising gestures or voice (e.g. speaking aloud). The sensor also utilizes a pan-focus-zoom camera in order to attend to individual students once a potential speaker has been identified.

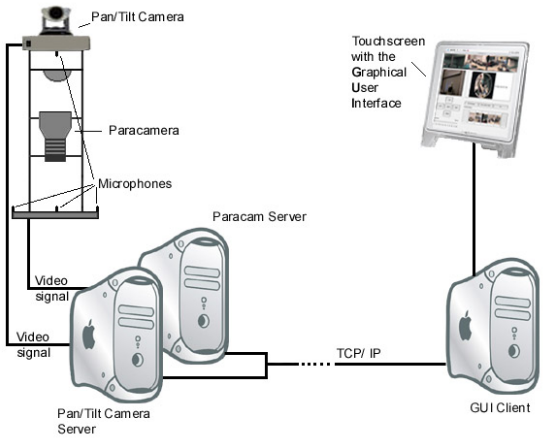


Figure 1: Architecture overview.

2 Overview

From a hardware point of view, the system is comprised of two main components (see Figure 1): (1) a sensor composed of an omni-directional audio sensor, an omni-directional video sensor and a camera with an electronically controllable zoom mounted on a pan-tilt-unit (hereinafter referred to as a pan-tilt-zoom camera), and (2) a touch-screen based graphical user interface (GUI) that permits the instructor to interact with the classroom using the sensor. The sensor is meant to be mounted on the ceiling of each remote classroom in order to permit capture of all potentially important audio and visual information but, this need not be the case. It can be placed anywhere in the room, provided it is placed at a height greater than any of the objects of interest in the scene (e.g. the students), to ensure that they are within its component's visual field of view. The touch-screen based GUI is used by the instructor and provides a simple interface to interact with the students of the remote classroom.

The sensor is used to automatically detect and attend to students wishing to interact with the instructor. Students who wish to interact with the remote instructor may signal their intent via voice (e.g. by speaking aloud) or using hand raising gestures. Hand raising gestures are detected using a combination of color and motion cues over a sequence of omni-directional images while sound localization techniques with a steerable microphone array allows for detection of audio signals. The system identifies "attention seeking" actions in the audio and video domain, and then presents potential speakers to the remote instructor. The instructor can then select which one of the potential speakers (including speakers who have not sought attention overtly) and the sensor will then attend to that speaker in both the audio and visual domains. A high resolution view of the speaker and a beamformed audio signal can then be presented to the remote instructor.

2.1 Vision Subsystem

The sensor has two competing requirements, to be able to view the entire classroom in order to not miss any salient visual cues, and to attend to specific students. Note that multiple speakers may request attention simultaneously, and thus it is essential that the sensor be capable of dealing with multiple events simultaneously. We have chosen to address this issue using two separate sensors, an omni-directional camera to view the entire classroom, coupled with a pan-tilt-zoom camera to attend to individual speakers.

Cyclovision's Paracamera omni-directional camera consists of a high precision paraboloidal mirror and a combination of special purpose lenses. By aiming a suitably equipped camera at the face of the paraboloidal mirror, the optics assembly permits the Paracamera to capture a 360^o *hemispherical* view of all students at the remote site, from a single viewpoint. Once the hemispherical view has been obtained, it may be easily un-warped [9] producing a *panoramic* view. From this panoramic view, a *perspective* view of any size corresponding to portions of the scene can be easily extracted.

The ability of the Paracamera to capture an image of the entire visual hemisphere makes it very attractive for a variety of applications. For example, Stiefelhagen et. al. [11] and Yong et. al. [13], use an omni-directional camera to capture the simultaneous video of each participant in a small group meeting. Omni-directional devices have also been used in many other vision based applications, including surveillance [3, 5], autonomous robot navigation [14], virtual reality [12], telepresence [12], remote view from a Dolphin [2], and pipe inspection [1].

A Directed Perception pan-tilt unit is mounted directly above and in line with the Paracamera. A computer-controllable zoom camera is mounted on this pan-tilt stage. The pan-tilt unit is mounted as close as possible to the hemispherical mirror of the Paracamera, and the pan axis of the pan-tilt unit is aligned with the horizontal axis of the Paracamera. Given this alignment, and assuming that the two sensors will be directed at targets at some distance from the two visual sensors, it is straightforward to calibrate the Paracamera and the pan-tilt-zoom cameras to a common directional coordinate system.

2.2 Audio Subsystem

The audio system consists of four omni-directional microphones (m_1, m_2, m_3 and m_4), mounted in a static, pyramidal shape about the Paracamera and provides an acoustic array, capable of localizing sound sources in three-dimensional space [4]. Through beamforming [6], it is possible to steer the audio array in different directions in order to attend to specific audio events (see [7] for details). In addition to providing steered audio from the remote class to the instructor, it is also possible to attend to novel audio events using sound localization techniques [10, 4] with the four microphones. For example, a student saying "excuse me" at a volume significantly above the background noise in the environment, could be attended to by the audio system without a cueing visual event. Once again however, when a speaking student has been localized, beamforming can be used to focus on the student.

The audio coordinate system must be calibrated with the visual coordinate system. The location of the four microphones m_i in the Paracamera coordinate system is obtained through physical measurement and by observing the three lower microphones in the Paracamera view. During system calibration, an image obtained by the Paracamera (see the left image in Figure 2) is presented and the coordinates of the microphones are determined manually by "pointing and clicking" (using the mouse). Once the image coordinates have been specified, the method described in [5] is used to determine the position (x, y and z coordinates) of the three microphones in the real world relative to the video system coordinate system.

Since the fourth microphone is not visible in the Paracamera image, this procedure cannot be used to determine its position. However, as shown in Figure 1, microphone four is mounted directly above the vertex of the paraboloidal mirror and therefore corresponds to a simple translation (which is easily measured) on the z-axis. Therefore, its real world position can also be determined.

Despite the potential for errors, the method described above seems sufficient for the task at hand. Each microphone itself occupies a very small portion of the image and as long as the center of the microphone is chosen to be somewhere in the image region corresponding to it, the error will be rather small. Furthermore, informal lab surveys indicate that after repeating the "point and click" operation several times, the difference in the image coordinates of the microphone center is actually small, usually between one or two pixels in each of the i, j image coordinates. Finally, the exact center of the microphone in the image may not actually be required. The microphone itself

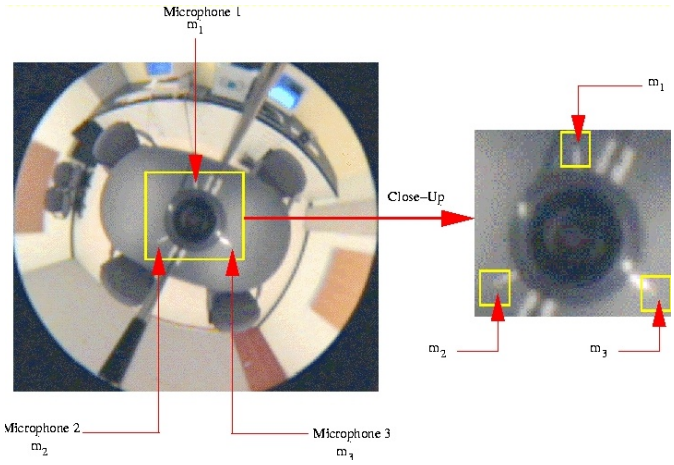


Figure 2: Image coordinates of the three microphones are obtained by clicking to the center of each individual microphone in the image. The position of the microphones in the real world can then be determined using the procedure described in [5].

is certainly not a "point receiver" and as a result, the center of the microphone may not necessarily define its position in the real world correctly. Therefore, it is sufficient to chose the image coordinates anywhere in the region occupied by the microphone in the Paracamera image.

3 Attending to Students

In a traditional classroom setting, students wishing to interact with the instructor signal their intent by raising their hand in some manner (e.g. hand raising, hand waving etc.) or, although less common, simply speak aloud. Students use these methods of interaction (especially hand raising gestures), from a very young age and have therefore become accustomed to them. In order to maintain, as much as possible, a natural classroom setting, rather than requiring students to learn a new method, in this system, students wishing to interact with the instructor can do so using these two "traditional" methods. Hand raising gestures are detected by the video system while students seeking the instructor's attention by speaking aloud are detected by the audio system. Once a student wishing to interact has been detected, the direction to this potential speaker is transmitted to the remote instructor and (s)he can choose to direct the pan-tilt-zoom and audio systems to attend to this speaker. The following sections provide greater details regarding the operation of the audio and video student detection systems.

3.1 Visual Event Detection

Hand raising gestures are detected in the video domain over a sequence of Paracamera (hemispherical view) images using a combination of color and motion cues. Color models for both skin and non-skin color classes have been previously constructed by the authors for the automatic detection of faces of the participants of a video teleconferencing session [8]. The models were constructed by manually classifying regions in Paracamera images as either skin or non-skin. A simple Bayesian model is used to detect regions of skin present in the Paracamera (hemispherical view) images. The detected skin regions correspond to hands, arms, head and other exposed regions of skin of people in the visual field of view of the sensor and any objects in the environment which may have similar color, including corkboards (as shown in Figure 3), cardboard boxes etc. However, it is the raising arm which is of interest and not the detection of the head/face or any other skin colored objects. Given that the majority of the detected skin regions, whether they correspond to actual exposed regions of skin or to other objects having similar color, will exhibit little or no movement when examined over several image frames, motion cues are used to disambiguate the raising hand from other detected skin regions.

Examination of hand raising gestures over a sequence of Paracamera images reveals that regardless where in the image the participant may be situated or whether the hand is raised straight upwards or diagonally to the left or right, the skin regions corresponding to the raising hand follow a trajectory moving outwards from the center of the Paracamera image. As a result, only detected skin regions exhibiting such motion patterns are considered. Motion is detected using image differencing over two or more image frames. Once all such regions have been determined, spatially close regions are grouped together to form a single cluster. A convex hull is then formed around each cluster allowing each cluster to be treated as a single entity.

Some preliminary results of the hand raising gesture detection operation are provided in Figure 3, where three sequential frames of a student raising their hand are shown. The region corresponding to the raising hand, as detected by the system, is outlined in yellow. Despite the presence of other skin colored regions (potential distracters) such as the corkboard and the subject's face, the system is capable of correctly detecting the raising hand. The corkboard is actually incorrectly classified as skin however, since its position remains static over time and therefore exhibits no motion patterns, it is not classified as a hand raising gesture. Similarly, the subject's head is also detected as skin but, the head does not necessarily remain stationary over time and will move slightly (or even substantially) over a short period of time. However, preliminary results suggest that head movements are typically small, do not persists for a prolonged period of time and do not necessarily follow the same trajectory as a raising hand. As a result, head movements have not posed a problem to date.

3.2 Audio Event Detection

Novel audio events that are distinct from the background can be used to identify the direction from which the sound originated. Assuming that the background noise is low relative to attention-seeking sounds, it is straightforward to identify corresponding sounds recovered by each of the four microphones (see [10]). The delay between the arrival times at the four microphones identifies the direction from which the audio event originated. The deployment of the audio event detection system is currently underway.

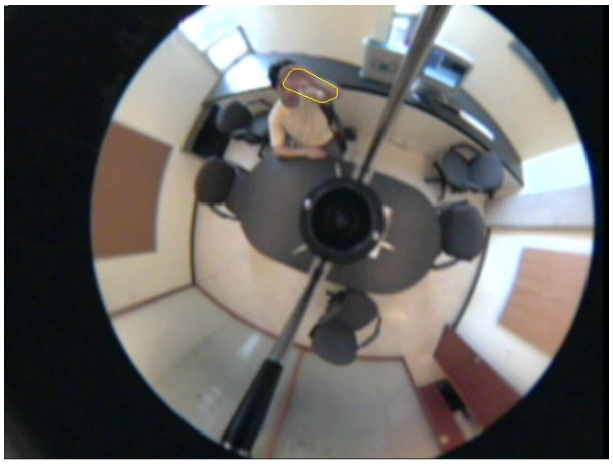
Note that it is not trivial to identify the direction to sounds that are not 'novel' due to the fact that determining the correspondence between the audio events at each of the microphones can be very difficult.

3.3 Summary

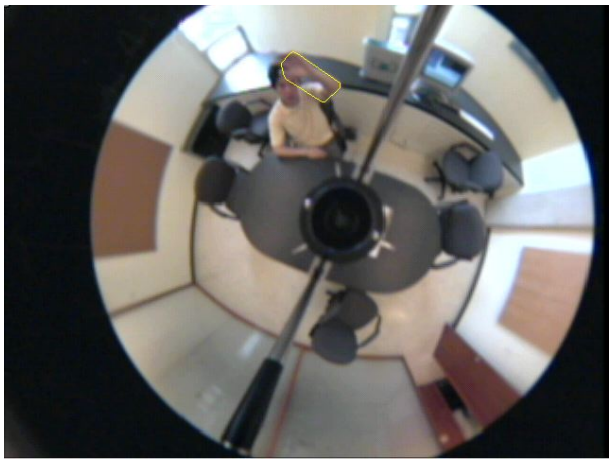
The audio and video event detection systems run in parallel and can identify multiple targets in parallel. For each audio or visual target, the direction to that target is obtained and this information is forwarded to the remote instructor. The omni-directional visual sensor and the static beamforming audio sensor are capable of looking in all directions simultaneously, thus permitting multiple potential speakers to be identified. The instructor's user interface (see below) is then used to attend to specific speakers, either potential speakers identified by the audio or visual event detection systems, or speakers identified by the instructor him/herself.

4 Instructor User Interface

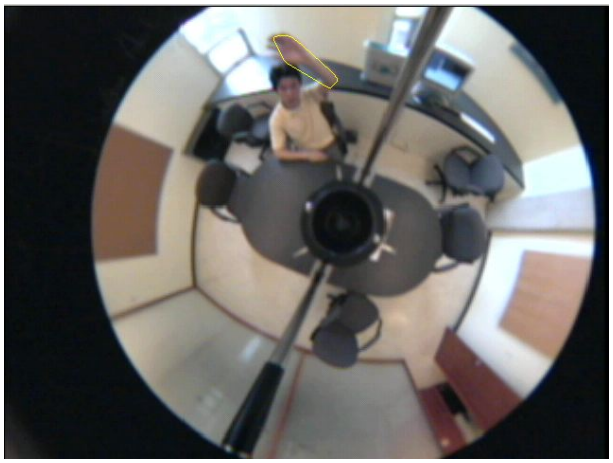
The instructor interacts with the system through a touch-screen based graphical user interface (GUI). The interface provides live video obtained from the omni-directional sensor and pan-tilt-zoom sensors, as well as directing the audio signal that the instructor will hear. As shown in Figure 4, the interface consists of the following five components.



(a)



(b)



(c)

Figure 3: Sample hand raising detection sequence over three consecutive frames. The raising arm, as detected by the system, is indicated by the yellow outline.

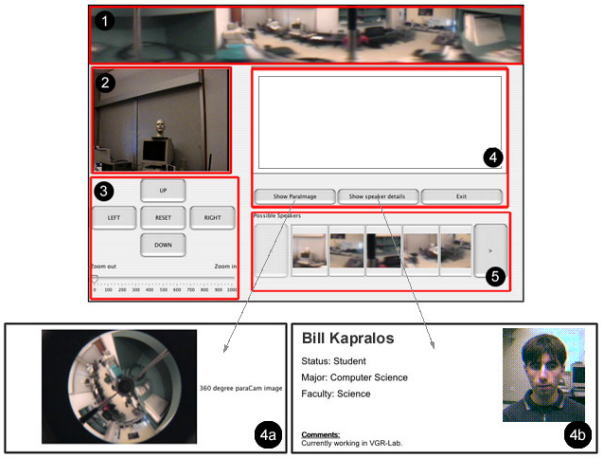


Figure 4: The five component's comprising the instructor's GUI.

4.1 Panoramic View Panel

The panoramic image is obtained by "un-warping" the original hemispherical view obtained with the Para-camera. This view provides the instructor with an overview of the entire classroom. Given its poor resolution, it may be difficult for the instructor to identify any particular students or capture any facial expressions. However, by "touching" the portion of the panoramic view corresponding to the region of interest in the "real world", the instructor can have the sensor (pan-tilt-zoom camera and microphone array), focus on this direction and thereby obtain a higher resolution view of this particular region.

4.2 High Resolution Image Panel

A live video image from the pan-tilt-zoom camera is displayed in this panel. The image resolution is greater than the resolution of the panoramic (or hemispherical) views however, this view does have a narrower field of view.

4.3 Pan-Tilt-Zoom (PTZ) Camera Control Panel

This panel consists of five buttons and one scroll pane to allow for manual control of the pan-tilt-zoom camera. **Up** and **down** will move the camera in an upwards and downwards direction by increasing and decreasing the "tilt" of the PTU respectively. **Right** and **left** will move the camera to the right and left by "panning" the PTU to the right and left respectively. **Reset** will position the pan-tilt-zoom camera to its initial position (e.g. its position during start-up) and set the camera zoom value to one.

Camera zoom value can be manually adjusted by using the scroll pane located at the bottom of this panel. The amount the camera actually moves in either direction (up, down, right or left) using the four buttons, is dependent on the current zoom value.

4.4 Multi-functional Information Panel

This panel is used to provide additional information to the instructor and to exit the program. Currently, two pieces of information are provided. The instructor can choose between viewing the hemispherical (un-warped) view obtained with the Paracamera (*View Paracamera Image* button) or, information related to a particular student currently viewed in the high resolution panel (*View Speaker Details* button). As shown in Figure 4, student information can include for example, name, status, department, course major or any other information required by the instructor.

4.5 Potential Speaker Panel

At any one time, there may be multiple students competing for the instructor's attention. The button scroll pane provides the instructor a "list" of all students seeking his/her attention and provides a simple method of choosing which student to attend to. Each button in this scroll pane corresponds to a student wishing to interact with the instructor as detected by the sensor using either hand raising gestures or sound localization techniques. For each detected student, a region in the panoramic image (e.g. un-warped hemispherical view) surrounding the student is extracted, placed on a button and added to this scroll pane. The instructor can then have the system focus on one particular student by simply "touching" the button corresponding to this student. After choosing a student, the pan-tilt-zoom camera and microphone array are steered accordingly and the interaction can begin.

5 Discussion

This paper has presented a snapshot of an ongoing project, aimed at developing a system for synchronous distance learning to enhance interaction between an instructor and students at (potentially) many remote locations. A new attentive visual sensor with fused high-resolution and low-resolution components, tailored to remote learning applications has been designed. The sensor consists of an omni-directional video sensor (Paracamera), a microphone array and a computer-controlled zoom camera mounted on a pan and tilt unit. Using a combination of color and motion trajectory cues, students at each remote location attempting to capture the instructor's attention via hand raising gestures can be detected. The system can then focus on these students with the pan-tilt mounted camera and microphone array to provide the instructor with a higher resolution image and to capture their speech respectively. Alternatively, sound localization techniques can be used to detect students who signal their intent to interact with the instructor vocally.

5.1 Future Plans

As mentioned, this project is currently ongoing and despite the promising results so far, considerable work remains. Most importantly, in order to test the sensor, field trials will be scheduled using a sequence of graduate level distance learning courses. These field trials will be used to refine the sensor and its software and obtain feedback with respect to the instructor's touch-screen interface (e.g. what can be added, what should be removed etc.).

A more accurate method of calibration between the Paracamera and the pan-tilt-zoom camera will also be implemented. A method permitting for accurate calibration between the Paracamera and a pan-tilt mounted camera has been described by Boult et. al. [3] and we are currently working on incorporating this technique into our work. We are also exploring the development of a faster/smaller/lighter sensor that has the same capabilities as the "laboratory bench sensor" that we have been using to date. With respect to the instructor's interface, future versions can involve keeping an image database of each student in a remote classroom along with a database of relevant information regarding each student. During normal system operation, after detecting a particular person, face recognition software may be used to reference them in the stored database. The instructor can have all the relevant information regarding a particular student available to him or her. This information can then be displayed in this information panel.

We are also in the process of collecting sample sequences of hand raising gestures by subjects. These sequences will be analyzed to construct potential statistical models which capture any information which is common to such hand raising gestures in sequences of Paracamera images (e.g. such as the trajectory followed by the raising hand, size of the convex hull surrounding the skin regions corresponding to the raising hand etc.).

Acknowledgments: The financial support of NSERC, NCE IRIS, IBM and CRESTech is gratefully acknowledged.

References

[1] A. Basu and D. Southwell. Omni-directional sensors for pipe inspection. In *IEEE Trans. Syst. Man Cybern*, volume 25, pages 3107–3112, 1995.

[2] T. E. Boult. Dove: Dolphin omni-directional video equipment'. In *Proc. of IASTED Conf on Robotics and Automation*, August 2000.

[3] T. E. Boult, R. J. Micheals, X. Gao, P. Lewis, W. Yin C. Power, and A. Erkan. Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets. In *Proc. IEEE Workshop on Visual Surveillance*, pages 48–55, Fort Collins, Colorado, USA, June 1999.

[4] K. Y. Guentchev and J. J Weng. Learning based three dimensional sound localization using a compact non-coplanar array of microphones. In *AAAI Spring Symp. on Int. Env.*, Stanford CA, March 1998.

[5] D. Gutchess, A. Jain, and S. Cheng. Automatic surveillance using omni-directional and active cameras. In *Proc. Asian Conf. Comput. Vis.*, 2000.

[6] D. Johnson and D. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice Hall, USA, 1993.

[7] B. Kapralos. Eyes 'n ears: A system for attentive teleconferencing. Master's thesis, Department of Computer Science, York University, Toronto, Ontario, Canada, April 2001.

[8] B. Kapralos, M. Jenkin, and E. Milios. Eyes 'n ears face detection. In IEEE, editor, *Proc. 2001 IEEE ICIP*, pages 66–69, Thessaloniki, Greece, October 2001.

[9] V. Peri and S. Nayar. Generation of perspective and panoramic video from omnidirectional video. In *Proc. DARPA Image Understanding Workshop*, pages 243–245, New Orleans, LA USA, 1997.

[10] G. Reid and E. Milios. Active binaural sound localization. In *EUSIPCO*, volume IV, pages 2353–2356, Rhodes, Greece, 1998.

[11] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing. In *Proc. ACM Mult. '99*, pages 3–10, Orlando, FL USA, October 1999.

[12] Y. Yasushi. Omni-directional sensing and its applications. *IEEE Trans. Inf. & Syst.*, E82-3, March 1999.

[13] R. Yong, A. Gupta, and J. Cadiz. Viewing meetings captured by an omni-directional camera. In *ACM Trans. Comput.-Hum. Interact.*, March 2001.

[14] J. Zheng and S. Tsuji. Representation for route recognition by a mobile robot. *Int. Conf. Comput. Vis.*, 9(1):55–76, 1992.