National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*A Panoramic Video and Acoustic Beamforming Sensor for Videoconferencing **

Fiala, M., Green, D., and Roth, G.
October 2004

* published in the IEEE International Workshop on Haptic Audio Visual Environments and their Applications (HAVE'2004). Ottawa, Ontario, Canada. October 2-3, 2004. NRC 47364

Canada

A Panoramic Video and Acoustic Beamforming Sensor for Videoconferencing *

Mark Fiala, David Green, and Gerhard Roth
Computational Video Group, National Research Council, Ottawa, Canada K1A 0R6
mark.fiala@nrc-cnrc.gc.ca, dave.green@nrc-cnrc.gc.ca, gerhard.roth@nrc-cnrc.gc.ca.

Abstract

Videoconferencing systems in use today typically rely on either fixed or pan/tilt/zoom cameras for image acquisition, and close-talking microphones for good quality audio capture. These sensors are unsuitable for scenarios involving multiple users seated at a meeting table, or non-stationary users. In these situations, the focus of attention should change from one talker to the next, and if possible track moving users. This paper describes a multi-modal perception system using both video and audio signals for such a videoconferencing system. An omnidirectional video camera and an audio beamforming array are combined into a device placed in the center of a meeting table. The video and audio is processed to determine the direction of who is talking, a virtual perspective view and directional audio beam is then created. Computer vision algorithms are used to find people by motion and by face and marker detection. The audio beamformer merges the signals from a circular array of microphones to provide audio power measurements in different directions simultaneously. The video and audio cues are combined to make a decision as to the location of the talker. The system has been integrated with OpenH.323 and serves as a node using Microsoft NetMeeting.

1 Introduction

Videoconferencing is a growing modality for people to communicate, the addition of video imagery to the conventional telephone improves the quality of experience for users. To use today's videoconferencing system one must usually sit within the view of a narrow field of view video camera. Ideally a videoconferencing device would find who is talking and pan and zoom a camera around to the different people as they take turns talking. A "virtual cameraman" device should be able to automatically find people and track their positions as they move around. An automatic device, preferably with no moving parts, that can achieve this can contribute to videoconferencing.

Kapralos [4] uses a panoramic camera and a simple microphone array for videoconferencing. This work discusses pointing accuracy but does not address talker tracking. Cutler [1] describes a panoramic system composed of multiple cameras and a beamforming microphone array used to archive meetings. The system that we describe uses video to locate potential talkers, and a circular beamforming microphone array to continuously search the meeting space for sound cues. Only rudimentary calibration is required. In real-time, we can select the talker candidates in the video image using a combination of cues while directional audio is used to choose the active talker.

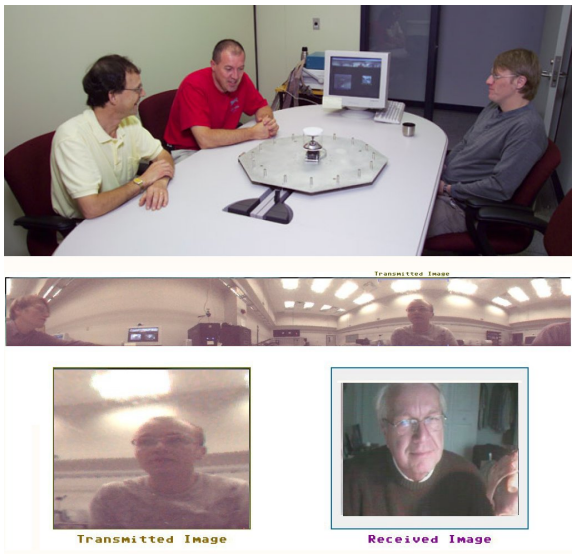

Figure 1: *Panocam* system used in a videoconference.

This paper discusses an experimental, combined panoramic video camera and microphone array which is placed at the centre of a meeting table which can dynamically detect and track the talkers seated around a table (Fig.1). This system we call *Panocam* sends audio and video of the current talker to a remote location. The video camera uses a commercial mirror/lens assembly to obtain a raw 360 degree image which is converted to cylindrical panoramic and virtual perspective views. Three com-

*NRC No. 47364

puter vision algorithms are used on the panoramic imagery; motion, face detection and marker detection. A circular array of 16 microphones is used for audio beamforming. The beamformer, which is implemented on a DSP, searches for sound sources in 15 directions simultaneously. Furthermore it is steered and acquires audio in any given direction under program control. By combining both video and audio derived cues, we have demonstrated a methodology to track in real-time, discussions involving several people. Finally, we have integrated the above functionality into an application that is compatible with Microsoft NetMeeting (http://www.microsoft.com/windows/netmeeting/) using OpenH.323 (www.openh323.org/). A screen capture of the computer at the panocam end of the link is shown in Fig.1(bottom).

2 Panocam

An experimental system combining an omni-directional video camera and a circular microphone array is placed in the centre of a meeting table, where it can detect and track in real-time people seated around a table. The system uses the panoramic video and beamforming microphone array to create a virtual perspective view and directional audio aimed at whoever is talking. This output directional audio and video is a virtual panning camera containing a perspective corrected section of the panoramic imagery for video and a directionally filtered audio output giving superior audio quality. In real-time, we can combine computer vision elements such as human face, motion and marker detection with a panoramic measure of sound power to decide in what direction to aim this virtual panning camera.

The video system is composed of a digital video colour camera fitted with a panoramic lens/mirror assembly. It captures a digital color video stream of which an annular region of 800 pixels diameter contains a 360 degree view of the meeting. This panoramic view, or cropped region thereof, is unnatural to present to a human viewer and thus has to be converted to the image that would have been seen by a traditional camera before being transmitted to the other end of the videoconference link. A first transformation warps the useful pixels in the raw image into a standardized panorama accounting for all device specific parameters such as focal length, radial profile, etc. A second transformation produces a final image with correct perspective. A number of cues were investigated to detect candidate talkers in the image. These included skin colour, motion, face based on OpenCV, and marker using ARToolkit. We have found motion detection to be the most robust. It is used to generate a set of azimuth angles that point to meeting members. These stages are shown in Fig.2.

Extensions to this basic videoconferencing prototype have been added which allow identification of who is talking, either by machine vision facial recognition, or by recognizing a unique marker pattern worn as a pin by the user. An experimental Face Server maintains a database of faces that have been previously captured. As a new face is detected, its position is noted and a matching algorithm selects the best candidate and labels the image. The optional marker module uses *ARToolKit* to analyze the image to detect unique markers carried by the users. This information can be used to locate the user and to annotate the display.

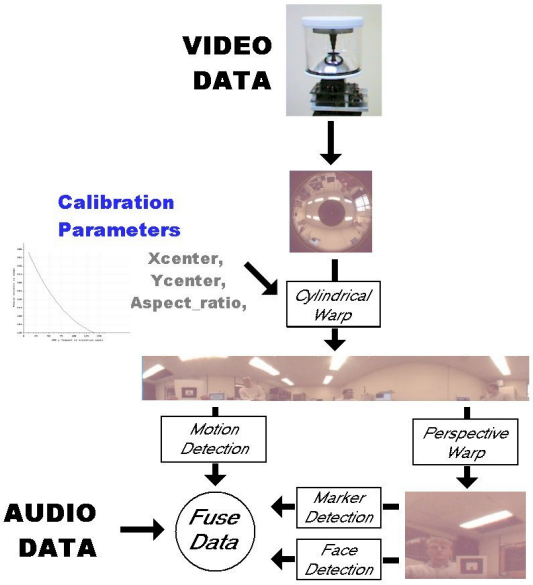

Figure 2: Processing of video data and integration with audio data.

The H.323 module manages communication over a local network or the Internet. This system has been demonstrated communicating remotely to other users using the NetMeeting program found on most Windows PCs. The remote user sees a rectangular image window aimed in the direction of who is talking, and hears an audio signal emphasizing sound from that direction.

3 Video Processing

3.1 Motion Detection

We have found motion detection to be the most robust of the three computer vision algorithms employed to find potential talkers, indeed motion detection alone was found to be sufficient to locate talkers and was the only algorithm activated for most of our tests and videoconferences performed with the system.

Motion detection is performed on the cylindrical image and uses image subtraction of successive images, thresholding of image differences, and aggregation to produce single estimates for the center of areas of large motion. Fig.3 below shows a screen capture when four people are in the

scene. The thresholded temporal difference image is shown below with black showing regions of greater motion. The instantaneous motion trigger a *motion event* which remembers that motion was present at that azimuth angle. When new motion is observed within a set number of degrees, this motion event is modified. If the motion is observed outside the "same talker" parameter, a new motion event is created. The motion events time out after a set time interval if no new motion is observed.

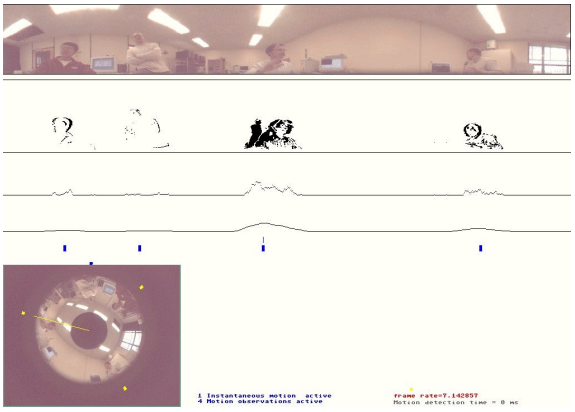

Figure 3: Motion detection used to locate potential talkers.

3.2 Marker and Face Detection

Motion provides only the indication that an unknown talker is present at an azimuth angle, with marker detection an identity can be ascribed to this talker allowing a name to appear along with the image similar to interviews on television. 2-D *self identifying markers*, equivalent to 2D barcodes, can be worn or set in front of meeting partipants. When these markers are identified in the image a corresponding ASCII string is ascribed to this potential talker. The marker system used is the popular ARToolkit [6, 5] (http://mtd.fh-hagenberg.at/depot/graphics/artoolkit/).

Face detection is the third computer vision algorithm employed to detect the presence of potential talkers. If facial recognition is available (face server) then an identity can be ascribed to the talker as with marker detection. The face detection algorithm is an OpenCV function (www.intel.com/research/mrl/research/opencv/overview.htm), which implements the work of Viola and Jones [8]. The face detection is applied to a perspective warp image. Depending on processing power available, a system can be looking simultaneously in all directions for faces, or take turns each frame to look in only some or one direction as does our implementation. Our system has a single rotating perspective view window in which face and marker detection is done once per frame.

Fig.4 shows an example screen shot from our system. A box is shown around the face in the perspective view as detected by the OpenCV function, and a thumbnail of the face appears below the panoramic strip. If the face identification server is connected, the face is resampled to a standard size and sent to the face server for identification.

The performance of the marker and face detection suffers from the low image resolution and contrast respectively. Markers have to be quite large to be detected. Fig.4 shows the typical size of the required marker. We plan to improve the image resolution so that small patterns that clip onto a talker's shirt can be used.

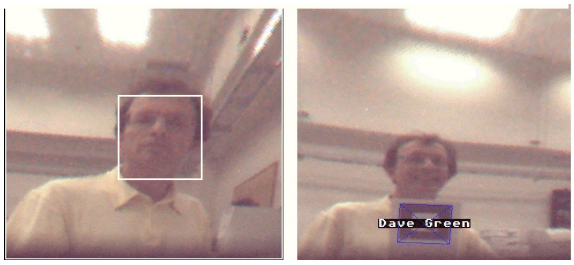

Figure 4: Face (left) and marker (right) detection used to locate potential talkers. The marker allows for talker identification, as does the face detection if the facial recognition system is employed.

4 Audio Processing

The responsibility of the audio processing subsystem is to report the speech power as a function of azimuth angle around the camera. An example of the data given by audio processing is shown in Fig.5.

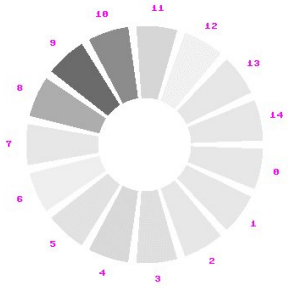

Figure 5: Speech power measurements as a function of azimuth angle. Each sector represents the audio power in one of the beamformer directions.

The circular microphone array is composed of 16 omnidirectional sensors uniformly distributed on a diameter of 57 cm. Delay-and-sum beamforming is used for directional audio capture. A digital signal processor is used to implement the beamformer. This system computes 16 listen directions in each sample interval. One direction is specified for the

final audio output and is variable. The remaining 15 beamformer directions are fixed in equally distributed angles, and are used to analyze speech power in these directions. The main system and video software runs on a desktop PC, and the microphone array controller runs on a high performance embedded DSP which makes possible the real-time, simultaneous search and output beamforming functionality. The DSP sends a measure of the speech power for all listen directions to the main application, and receives steering commands from the main application to aim the variable audio beam.

The circular microphone array is shown in Fig.6. The raw electret microphone signals are pre-amplified and filtered at 4.8 kHz before being digitized by a 16-channel ADC sampling at 11.025 kHz per channel. A 200 MHz TMS320C6201 DSP PCI card (http://www.innovative-dsp.com/) performs the beamforming calculations.

An array of microphones together can be used for *audio beamforming*, where a virtual steerable sensor can be created from an array of stationary sensors. A delay-and-sum beamformer [7] is used for Panocam. In general, a circular beamformer can generate a single main lobe which can be dynamically steered in both azimuth and elevation. For this application, we fix the elevation angle to zero degrees so that the main lobe is limited to the horizontal plane. We can then steer the beam to any direction on the horizontal.

Figure 6: The Panocam Sensor. A video camera and panoramic lens is situated at the centre of a 16-element microphone array. The array is 57 cm in diameter and generates a main lobe which can be dynamically steered along the horizontal plane.

Simulations of both nearfield and farfield responses of the beamformer (Fig.7) indicate that it has a 3 dB beamwidth of about 25 degs. at 1000 Hz and with side lobes about -6 dB. The simulation shows that because of the aperture size of the array (57 cm), nearfield effects are minimal at distances greater than 1m from the centre of the array. Fig.8 shows the polar response of the array at 1000 Hz.

The DSP can compute 16 look directions in each sample interval. One direction (steered audio) is specified for audio capture and is used for analog output via a DAC. This audio signal is sent to the input of the soundcard of the PC. The remaining 15 look directions which are equally distributed, are used to search for the current talker. Estimates of speech power are computed for each direction.

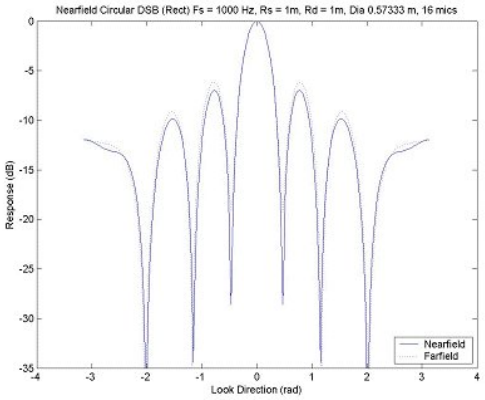

Figure 7: Simulation of the Circular Microphone Array. The beamwidth of the main lobe at 1000 Hz is about 25 degs. The nearfield and farfield responses are very similar with sound sources at 1 m or greater from the centre of the array. At this distance, the side lobes of the farfield response are only about 1 dB higher the nearfield response. Because this difference is not significant for this application, we use farfield beamforming because it is simpler to implement.

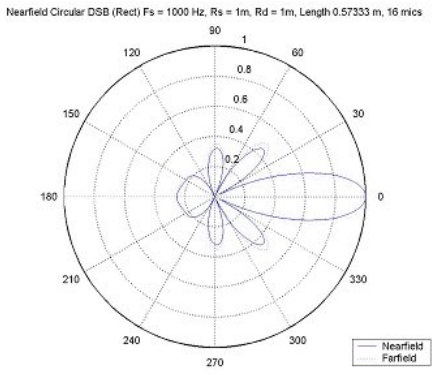

Figure 8: Simulation of Array Polar Response at 1000 Hz.

5 Sensor Fusion: Integration to Determine Active Talker

The different modalities of video and audio are integrated to both find the talker, and produce this virtual pan-tilt output. Each candidate is detected by the three methods of video information listed above and audio information, together they are used to select one of these candidate talker directions for output.

Fig.9 is a screen shot showing the integration of data to make the final talker decision. Motion, face and marker events are shown overlaid over the panoramic image (bottom left) and the cylindrical image (top). The audio information is shown in middle left. This shows the audio history as a function of angle, the most recent speech power readings are shown in the center.

The potential talkers are shown as rectangular views in the lower right part of Fig.9. The system then decides between these potential talkers as a function of the speech power in the direction of the talker.

Figure 9: Simulation of Array Polar Response at 1000 Hz.

6 Video Hardware and Image Quality

Three different omnidirectional cameras were assembled for our system. They are all *catadioptric* sensors (contain both lenses and mirrors in the optical path) created by replacing the lens in a digital video camera with a commercial mirror/lens. The components and their useful image parameters are shown in Figure 10. The best image of the three systems was created using a Pixelink digital video colour camera (http://www.pixelink.com/)) fitted with a Remote Reality NetVision Assembly B panoramic lens/mirror assembly (http://www.remotereality.com/) (Figure 1.). It captures a color image of 1280x1024 pixels of which an annular region of 800 pixels diameter contains the panoramic image. The unused space is due to this model of lens/mirror assembly being designed for both $\frac{1}{2}$ and $\frac{1}{3}$ inch CCD's.

A system of similar pixel resolution was created using a colour 1024x768 pixel dragonfly IEEE 1394 camera, however the image noise was higher giving a poorer subjective perception. Finally an NTSC camera was used which gives the least high quality image, but is still sufficient for talker tracking.

The resolution of the virtual perspective views is the limiting factor on resolution, the effect of a lower resolution perspective view is poorer subjective performance for the users and diminished performance for the face and marker detection vision algorithms. However, motion works sufficiently to detect the talking direction using even the NTSC image.

With the Pixelink and Dragonfly IEEE cameras, providing a useful annular image of diameter 800 pixels, the perspective warp has an equivalent pixel density to a 320x240 image. This is sufficient for videoconferencing systems like NetMeeting but for a higher resolution system a higher resolution is necessary. We are currently integrating a "Ladybug" multi-CCD IEEE 1394 camera from Point Grey Research which provides a nearly seamless hemispherical view with 6 separate cameras enclosed in a small package.

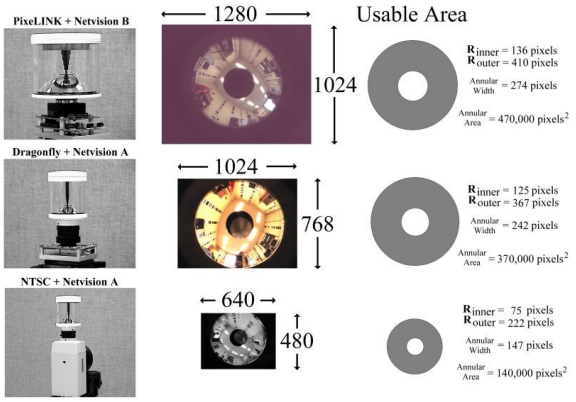

Figure 10: Resolution with three panoramic camera configurations.

7 Other Applications of Panocam

The face detection can be combined with a face recognition system to automatically recognize and label talkers. When the panocam system detects a human face, an image of the face can be cropped out and sent over the network to a server containing a database of faces and their names. The server would return an ASCII string identifying the person if the face matched one in its records. A diagram of the complete system developed at our facility is shown in Fig.11. Face recognition systems compatible with panocam are being developed in parallel in our institute [3, 2].

The system has uses beyond just videoconferencing, with the face recognition system and speech recognition in-

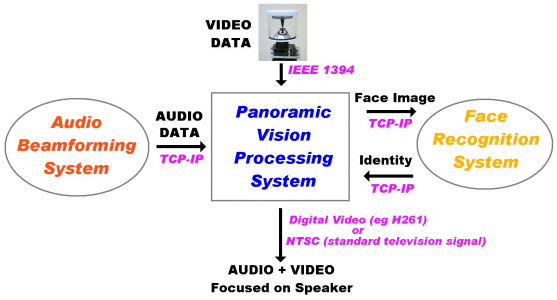

Figure 11: System diagram.

tegrated, it can be used to annotate minutes of a meeting. Text can be generated from the audio and correlated with the talker's name, and possibly a snapshot of their face as that speech was uttered. An example meeting record is shown in Fig.12 (this has not been implemented, this is just an example of what such a system could generate).

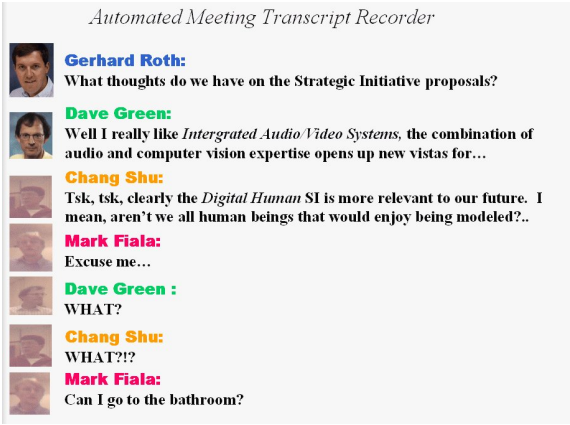

Figure 12: *Panocam* used as a meeting transcript recorder.

8 Conclusions

We have described a panoramic audio/video sensor for videoconferencing applications. This working prototype system demonstrates the successful fusion of information from two different modalities; video and audio, to construct a robust videoconferencing system whose automatic camera-person functionality would not be possible with only one modality alone. Problems related to low audio beam resolution and to reverberation are mitigated by the high resolution accuracy of video. Likewise, the many possibilities for who may be talking merely by video input is complemented by the audio power measurements which rule out the detected people who are not producing noise. We have experimented with face detection and identification, and marker detection for image annotation. The system includes H.323 functionality to transport the virtual view and audio to a remote videoconferencing node such as Microsoft Netmeeting.

References

[1] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed meetings: a meeting capture and broadcasting system. In *Proc. 10th ACM Intl. Conf. On Multmedia*, pages 503–512, 2002.

[2] D. Gorodnichy. Face recognition in video. In *Proceedings of International Association for Pattern Recognition (IAPR) International Conference on Audio and Video-Based Biometric Person Authentication (AVBPA'03), LNCS 2688*, pages 505–514, Guildford, United Kingdom, June 9-11 2003.

[3] D. Gorodnichy and O. Gorodnichy. Using associative memory principles to enhance perceptual ability of vision systems.

[4] B. Kapralos, M. Jenkin, and E. Milios. Audiovisual localization of multiple speakers in a video teleconferencing setting. In *Intl. Jour. Imaging Systems and Technology*, volume 13(1), pages 95–105, 2003.

[5] H. Kato and M. Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Proc. the 2nd IEEE and ACM International Workshop on Augmented Reality*, pages 85–94, San Francisco, CA, USA, Oct 1999.

[6] H. Kato, M. Billinghurst, and I. Poupyrev. *ARToolkit User Manual, Version 2.33*. Human Interface Technology Lab, University of Washington, 2000.

[7] B. van Veen and K. Buckley. Beamforming: A versatile approach to spatial filtering. In *IEEE ASSP Magazine*, volume 5, pages 4–24, 1988.

[8] Paul Viola and Michael Jones. Robust real-time object detection. *Proceedings of the Second International Workshop on Statistical Learning and Computational Theories of Vision Modeling, Learning, Computing and Sampling*, July 2002.