PANOCAM: COMBINING PANORAMIC VIDEO WITH ACOUSTIC BEAMFORMING FOR VIDEOCONFERENCING[1]

David Green and Mark Fiala
Institute for Information Technology, National Research Council of Canada, M-50, 1200 Montreal Rd, Ottawa ON Canada K1A 0R6, dave.green@nrc-cnrc.gc.ca

1. INTRODUCTION

Videoconferencing systems in use today, typically rely either on 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 or should track a moving talker. This paper describes an experimental, combined panoramic video camera and microphone array which is placed at the centre of a meeting table and which can detect and track in real-time the talkers seated around a table.

Kapralos [1] uses a panoramic camera and a simple microphone array for videoconferencing. This work discusses pointing accuracy but does not address talker tracking. Cutler [2] 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. 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/).

2. PANORAMIC VIDEO SYSTEM

The video system is composed of 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.

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" [3] based on OpenCV (www.intel.com/research/mrl/research/opencv/overview.htm), and "marker" using ARToolkit (http://mtd.fh-hagenberg.at/depot/graphics/artoolkit/). We have found motion detection to be the most robust. It is used to generate a set of azimuths that point to candidate talkers. As each candidate is detected, a video sub-image with correct perspective is produced, Figure 2. Audio information is then used to select one of these sub-images for output.

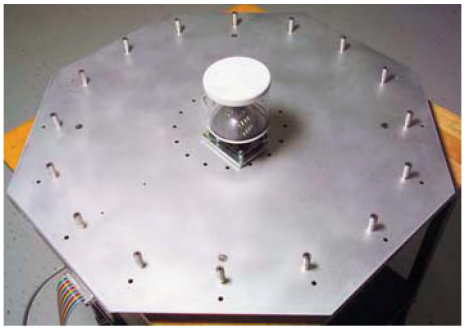

Fig. 1. Panoramic Audio/Video Sensor. A video camera with a panoramic lens is surrounded by a 16-element microphone array.

3. CIRCULAR BEAMFORMING MICROPHONE ARRAY

The circular microphone array is composed of 16 omnidirectional sensors uniformly distributed on a diameter of 57 cm, Figure 1.

The raw 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 delay-and-sum computations.

The DSP computes 16 "look directions" in each sample interval, Figure 2. One direction is specified for audio capture and is used for analog output via a DAC (steered audio). The remaining 15 look directions which are equally

[1] NRC No. 47171

distributed, are used to estimate speech power in each direction [4].

4. VIDEOCONFERENCING OVERVIEW

While the system software runs on a desktop PC, the microphone array controller runs on the DSP. The DSP receives steering commands from the main application, and returns the directional power estimates. Steered audio is available as an analog output signal. 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 [3]. 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. The H.323 module manages communication over the Internet. The system has been demonstrated communicating remotely with other users using NetMeeting. The remote user sees a rectangular image window aimed at the talker, and hears the steered audio.

5. TALKER TRACKING

A key requirement is to be able to automatically select the current talker in real-time. We address this by combining information from candidate sub-images with the audio search results.

When motion is detected in the panoramic image, a "potential talker" is created, and will persist according to some simple rules. A candidate persists for 120 seconds without motion. A non-moving candidate persists for 30 seconds after being the loudest audio direction. The system will track several potential talkers, Figure 2. The vision processing maintains a list of potential talkers, of which one is selected based on audio information. Every 11.6 ms, the DSP provides a frame of audio intensity as a function of heading angle, Figure 2. The final output viewing direction is chosen by selecting the potential talker with the highest audio response. A perspective view is warped in the direction of the selected talker. The selected view and the steered audio are transferred to the H.323 module. In this way sensory fusion is achieved; audio information chooses the general direction and video processing fine-tunes the steering direction.

Figure 2 shows an example with 3 candidate talkers participating in a discussion. The figure illustrates how video sub-images are selected using motion cues, and how the microphone beamformer is used to determine the instantaneous direction to the current talker.

6. DISCUSSION

The image quality of the prototype is limited by two factors. First, with the current optics, the raw image only occupies about 36% of the camera pixels. Sub-images therefore are of low pixel resolution. Second, the average light level determines the camera exposure. Therefore a very bright region on the raw image will produce poor contrast in other regions. We are exploring solutions to these problems.

7. CONCLUSIONS

We have described a panoramic audio/video sensor for videoconferencing applications. Problems related to low audio beam resolution and to reverberation are mitigated by the use of video-based face detection. The system has been demonstrated to dynamically detect and track the active talker amongst a group participating in a round-table discussion. We have experimented with face detection and identification, and marker detection for image annotation. The system includes H.323 functionality and so is compatible with Microsoft NetMeeting.

Fig. 2. Panocam. The raw video image appears in the lower left. The initial corrected panoramic image appears at the top. The real-time audio search response shows an active talker near 12:00 in the left centre frame and three virtual images are shown in the centre frame. The audio and corresponding virtual video (lower image, "6-10") is selected for output.

REFERENCES

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

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

[3] Gorodnichy, D. (2003) Facial Recognition in Video. Proc. of IAPR Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA'03), LNCS 2688, 505-514.

[4] Flanagan, J.L., Johnston, J.D. Zahn, R. and Elko, G.W. (1985) Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am. 78(5), 1508-1518