Realtime Meeting Analysis and 3D Meeting Viewer Based on Omnidirectional Multimodal Sensors

Kazuhiro Otsuka
NTT Communication Science Laboratories
3-1, Morinosato-Wakamiya
Atsugi, 247-0198 Japan
otsuka@eye.brl.ntt.co.jp

Shoko Araki
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho
Kyoto, 619-0237 Japan
shoko@cslab.kecl.ntt.co.jp

Dan Mikami
NTT Communication Science Laboratories
3-1, Morinosato-Wakamiya
Atsugi, 247-0198 Japan
mikami.dan@lab.ntt.co.jp

Kentaro Ishizuka
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho
Kyoto, 619-0237 Japan

Masakiyo Fujimoto
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho
Kyoto, 619-0237 Japan

Junji Yamato
NTT Communication Science Laboratories
3-1, Morinosato-Wakamiya
Atsugi, 247-0198 Japan

ABSTRACT

This demo presents a realtime system for analyzing group meetings. Targeting round-table meetings, this system employs an omnidirectional camera-microphone system. The goal of this system is to automatically discover "who is talking to whom and when". To that purpose, the face pose/position of meeting participants are tracked on panorama images acquired from fisheye-based omnidirectional cameras. From audio signals obtained with microphone array, speaker diarization, i.e. the estimation of "who is speaking and when", is carried out. The visual focus of attention, i.e. "who is looking at whom", is esimated from the result of face tracking. The results are displayed based on a 3D visualization scheme. The advantage of our system is its realtimeness. We will demonstrate the portable version of the system consisting of two laptop PCs. In addition, we will showcase our meeting playback viewer with man-machine interfaces that allow users to freely control space and time of meeting scenes. With this viewer, users can also experince 3D positional sound effect linked with 3D viewpoint, using enhanced audio tracks for each participant.

Categories and Subject Descriptors

H1.2 [Models and Principles]: User/Machine System — Human Information Processing

General Terms

ALGORITHMS, HUMAN FACTORS

Keywords

realtime system, meeting analysis, omnidirectional cameras, fisheye lens, face tracking, speaker diarization, microphone array, focus of attention

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.
ACM 978-1-60558-772-1/09/11.

1. REALTIME MEETING ANALYSIS

In recent years, multimodal meeting analysis has been acknowledged as an emerging research area and intensive efforts have been made to analyze meetings. So far, almost all research on meeting analysis has focused only on pre-recorded data and offline processing. However, realtime techniques for processing/analyzing meetings are becoming important because they are essential to realizing applications such as computer-mediated teleconferencing systems and social robots/agents. In last year's ICMI, the authors proposed the first realtime system for analyzing group meetings that uses a omnidirectional camera-microphone and integrates speaker diarization and visual face pose tracking[3].

In this demo session, we will show our realtime meeting analysis system as a preliminary research step toward future communications. The main differences of the demo system from our system in [3] are as follows. First, we have implemented an improved face pose tracker using Memory-Based Particle Filter (M-PF) [2], which increase robustness against abrupt head motion and the recoverbility of lost track caused such by occlusions. Second, we implemented semi-realtime speech enhancement method for enhancing each person's voice during meetings [1]. Third, we have developed a meeting playback viewer with man-machine interfaces that allow users to intuitively control space and time, of meeting scenes.

This demo targets round-table meetings, as shown in Fig. 1(a), and displays realtime output on a PC display. At the center of the table in Fig. 1(a), our omnidirectional camera-microphone system is placed (Fig. 1(c)). The camera part consists of two cameras with fisheye lenses, which are facing in (180 degree) opposite directions. Since each fisheye lens covers a hemispherical region, the camera system can capture a near spherical region. The microphone array consists of three tiny microphones placed at the vertices of a triangle, and is located atop the camera unit. This demo runs with two laptop PCs; one is for vision and the other is for audio processing.

Our system consist of visual processing, audio processing, and meeting processing parts. The visual processing part conducts face pose tracking on the panorama images

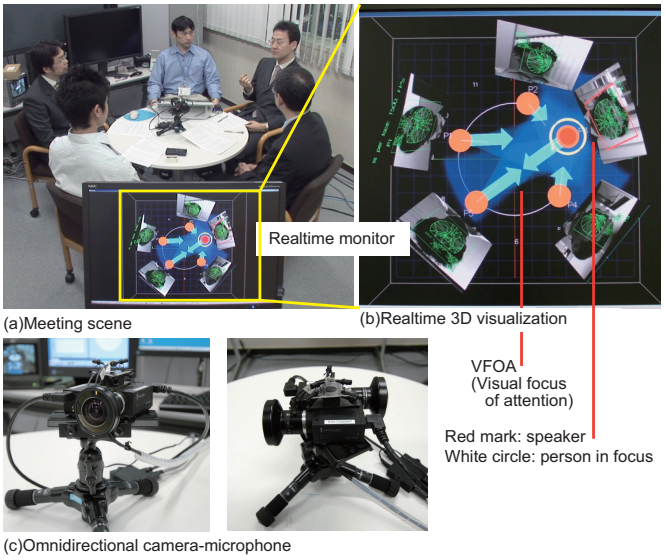

Figure 1: Overview of demo system

converted from fisheye images from the cameras. For face tracking, we employ a particle filter-based template matching, called STCTracker [2]. The audio processing part conducts a robust speaker diarization, which consisitf of a VAD (Voice Activity Detection) and DOA (Direction of Arrival) estimation followed by sound source clustering. The meeting processing part determines the utterance status (speaking or silent) of each participant by data association of audio and visual information. Moreover, the visual focus of attention is estimated from the positions and directions of the faces.

Finally, the result of analysis is displayed using our 3-dimensional visualization schemes, as shown in Fig. 1(b)[1]. The face images of each person is arranged in accordance with the actual position of participants around the meeting table. The diagram on the center of meeting space indicates the state of meeting including utterance status, head directions, gaze directions, and person in focus.

2. MEETING PLAYBACK VIEWER

A meeting viewer is developed for offline playback of meeting scenes recorded/analyzed with our realtime system, assuming possible users of meeting archive system. The meeting viewer provides several views including 2D panorama, tiled faces, 3D views, as shown in Fig. 2(a)∼(d). In addition, a timeline window is created to display transcript, detected voice intervals, gaze directions, facial expression, and so on, as shown in Fig. 2(e). Fig. 3 shows a laptop PC and user-interface devices including 3D mouse, Jog/Shuttle dial, and MIDI controller. These interfaces allows users to freely and intuitively manipulate their viewpoints in 3D meeting space and time, for better understanding of meeting scenes.

Using MIDI controller, users can control the volume and balance of enhanced audio tracks for each person. In addition, the volume and balances of the enhanced audio tracks can be linked with 3D viewpoint manipulated by the users; users can hear voice of person on right(left) in 3D view from right(left) loudspeaker. This 3D sound reproduction enables us an immersive meeting viewing. Furthermore,

[1]Demonstration movies are available from http://www.brl.ntt.co.jp/people/otsuka/ICMI+MLMI09.html

Figure 2: Screenshot of meeting viewer

Figure 3: Meeting viewer and interface devices

this software allow users to annotate behaviors of meeting participants such as gaze directions, facial expressions, and voice activity. The 2D+3D visualizaions with flexible viewpoint/time control interfaces contribute to decrease annotator's cognitive load and to increase the work efficiency.

3. REFERENCES

[1] S. Araki, H. Sawada, and S. Makino. Blind speech separation in a meeting situation with maximum SNR beamformers. systems. In Proc. ICASSP, pages 41–44, 2007.

[2] D. Mikami, K. Otsuka, and J. Yamato. Memory-based particle filter for face pose tracking robust under complex dynamics. In Proc. IEEE CVPR'09, pages 999–1006, 2009.

[3] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In Proc. 10th ICMI, pages 257–264, 2008.