A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose Tracking and Speaker Diarization

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

Kentaro Ishizuka
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho
Kyoto, 619-0237 Japan
ishizuka@cslab.kecl.ntt.co.jp

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

Martin Heinrich
NTT Communication Science Laboratories
3-1, Morinosato-Wakamiya
Atsugi, 247-0198 Japan

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

ABSTRACT

This paper presents a realtime system for analyzing group meetings that uses a novel omnidirectional camera-microphone system. The goal is to automatically discover the visual focus of attention (VFOA), i.e. "who is looking at whom", in addition to speaker diarization, i.e. "who is speaking and when". First, a novel tabletop sensing device for round-table meetings is presented; it consists of two cameras with two fisheye lenses and a triangular microphone array. Second, from high-resolution omnidirectional images captured with the cameras, the position and pose of people's faces are estimated by STCTracker (Sparse Template Condensation Tracker); it realizes realtime robust tracking of multiple faces by utilizing GPUs (Graphics Processing Units). The face position/pose data output by the face tracker is used to estimate the focus of attention in the group. Using the microphone array, robust speaker diarization is carried out by a VAD (Voice Activity Detection) and a DOA (Direction of Arrival) estimation followed by sound source clustering. This paper also presents new 3-D visualization schemes for meeting scenes and the results of an analysis. Using two PCs, one for vision and one for audio processing, the system runs at about 20 frames per second for 5-person meetings.

Categories and Subject Descriptors

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

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

General Terms

ALGORITHMS, HUMAN FACTORS

Keywords

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

1. INTRODUCTION

Face-to-face conversation is one of the most basic forms of communication in daily life and group meetings are used for conveying/sharing information, understanding others' intention/emotion, and making decisions. In the face-to-face setting, people exchange not only verbal messages but also nonverbal messages. The nonverbal messages are expressed by nonverbal behaviors in multimodal channels such as eye gaze, facial expressions, head motion, hand gesture, body posture and prosody; psychologists have elucidated its importance in human communications [2]. Therefore, it is expected that conversation scenes can be largely understood by observing people's nonverbal behaviors with sensing devices such as cameras and microphones.

In recent years, multimodal meeting analysis has been acknowledged as an emerging research area and intensive efforts have been made to capture meeting scenes, recognizing people's actions in meetings, and analyzing group interactions in meetings [7]. To date, virtually all research on meetings have focused on pre-recorded data and offline processing. However, realtime techniques for processing/analyzing meetings are necessary for realizing applications such as computer-mediated teleconferencing and interaction involving social robots/agents. Also, even for non-realtime applications such the archiving/browsing of multimodal meetings and psychological/social/clinical studies of human behaviors, realtime or near realtime processing would significantly enhance the effectiveness of the tasks including playback of meetings with reference to analyzed data created on the spot.

This paper presents a realtime multimodal system for analyzing face-to-face meeting scenes that uses a new omnidirectional camera-microphone system. The goal is to automatically estimate the visual focus of attention (VFOA), i.e. "who is looking at whom" and "who is attracting more gaze than others", in addition to speaker diarization, i.e. "who speaks and when", in realtime. To the best of our knowledge, this paper is the first to propose a realtime multimodal system to visually track not only face position, but also face pose in realtime for analyzing group meetings. A system that uses two PCs, one for vision and one for audio processing, runs at about 20 fps (frames per second) for 5-person meetings.

In this paper, we first describe our tabletop sensing device for round-table meetings; it consists of two cameras with two fisheye lenses and a triangular microphone array. From the high-resolution omnidirectional images that are captured, the position and pose of people's faces are estimated by STCTracker (Sparse Template Condensation Tracker), which realizes realtime robust tracking of multiple faces by utilizing GPUs (Graphics Processing Units). The face position/pose data output by the face tracker is used to estimate the focus of attention in the group. Using the microphone array, robust speaker diarization is carried out using VAD (Voice Activity Detection) and DOA (Direction of Arrival) estimation, followed by sound source clustering. This paper also presents new 3-D visualization schemes for meeting analysis and the results of some trials.

This paper is organized as follows. Section 2 overviews related works. Section 3 proposes our system, and Section 4 details the system configuration. Section 5 describes the experiments and visualization schemes. Finally, Section 6 presents our conclusion and some discussions.

2. RELATED WORKS

Omnidirectional vision system

To effectively capture round-table meeting scenes, omnidirectional cameras have been acknowledged as a reasonable solution, because they allow one image to cover the whole view [21, 24, 5]. However, omnidirectional vision has yet to supersede conventional camera systems, which capture meeting scenes using multiple cameras located at different viewpoints. One of the reasons is as follows. First, typical omnidirectional system employs catoptrics, i.e. equipping the camera with a mirror. Due to the low optical quality of the mirror and intrinsic complexity in the projection system, the effective resolution of peoples' faces and figures is relatively low. Second, the camera system is usually located at eye-level at the meeting; this obstructs natural gaze interaction. Placing the camera system on the table yields slanted views of people's faces. These problems limit the range of applicable computer vision techniques.

To avoid the problems triggered by mirror-based omnivision, another approach has been drawing attention; an omni-view image is created by fusing the images from the multiple cameras, which are embedded in a single body [14]. This type of system can provide near frontal images of people's faces with higher resolution than mirror-based systems. However, the main drawback of this system is the discontinuity created by image combination; the discontinuities hamper accurate face tracking.

To maximize the image resolution and to minimize the image discontinuities, we developed an omnidirectional vision system composed of two cameras and two fisheye lenses. A fisheye camera can capture hemispherical view, and two fish cameras facing in opposite directions can provide approximately spherical coverage. This system can provide frontal face images with higher resolution than mirror-based systems (and even multi-camera-based system). More specifically, our system yields 4896 pixels for 360-degree coverage on the horizontal plane, and achieves the grabbing frame rate of 30.0fps. Its high resolution allows precise vision tracking and useful data for human observers.

Face tracking in meetings

Face tracking is a common task in visual meeting analysis, and tracking methods can be categorized in terms of the goal pursued; one estimates only face position (the localization problem), the other estimate face position and pose (also called head pose or face direction). Examples of the former include [20, 23] for normal cameras and [24, 5, 4, 14] for omni-cameras. Examples of the latter include [11, 8, 18] for normal cameras and [21] for omni-cameras. So far, almost all tracking methods for meeting analysis have targeted only offline processing of pre-recorded meeting videos; most papers did not even mention the speed of their methods. Pose tracking entails higher computational complexity than localization, especially for multiparty meetings, and it also requires higher image quality. So far, these reasons have prevented realtime face pose tracking from omni-images. Our system resolves these problems by using our new omni-vision system and a GPU-based face-pose tracking algorithm.

Visual focus of attention

The importance of measuring face pose arises from the fact that it is reasonable indicator of people's gaze and direction of visual attention. Among the nonverbal messages/behaviors possible, eye gaze is especially important because it has various roles such as monitoring others, expressing one's attitude/interest, and regulating conversation flow [9]. However, gaze direction during natural conversation is difficult to measure directly. Therefore, face direction is often used as a reasonable alternative. Moreover, face direction is more than just an alternative; by itself it is a useful indicator of people's attention to others during meetings. In addition, the temporal changes in face direction indicate head gestures such as nodding. Therefore, face direction is an important cue in analyzing meetings.

In recent years, authors have indicated that the conversation structure, i.e. "who is talking to whom, when", can be estimated by using the utterance pattern (the speech or silence of people) and the gaze pattern among meeting participants, i.e. visual focus of attention in a group [16, 18]. They use a dynamic Bayesian network to jointly estimate the gaze pattern and the conversation structures from observed utterances and face directions. Ba and Odobez have also focused on face direction as an important cue in estimating the focus of attention, and have built a gaze estimator [3].

Multimodal smart rooms

Recently, a number of multimodal systems for meeting applications have been developed by a number of research groups; they are often referred to as "smart rooms". AMI and AMIDA projects built a multimodal recording infrastructure for collecting meeting data; it uses microphone ar-

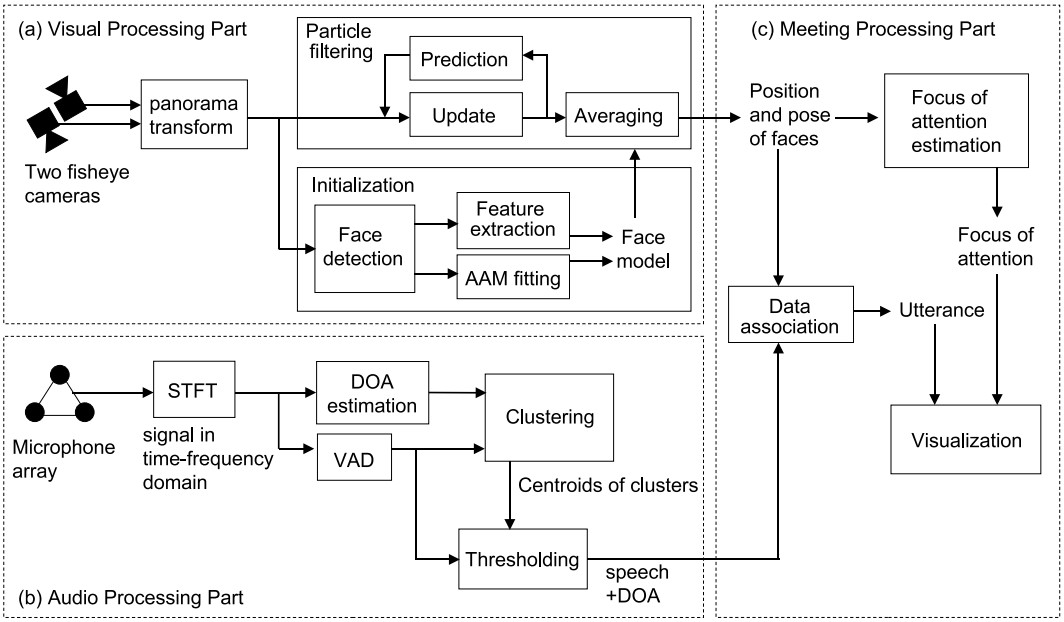

Figure 1: Diagram of system

rays, close-talk microphones, normal cameras, and omnidirectional cameras [19]. A team of USC has been developing a smart room equipped with microphone array, omnidirectional cameras, and far-field cameras [4]. They implemented a system that realizes a multiview localization and identification of people in meetings. The CHIL project is especially interested in using far-field active cameras and multiview techniques for tracking face pose to assess VFOA and realize personal identification [23]. In contrast to the "smart rooms" above mentioned, we aim to develop portable tabletop devices that do not require camera calibration or installation.

3. PROPOSED SYSTEM

Our system consist of three parts as shown in Fig.1, (a)visual processing, (b)audio processing, and (c)meeting processing parts. This paper targets meeting scenes as shown in Fig.2. The visual processing part consists of our new omnidirectional camera system (Fig.3) and face tracking system using the image output by the camera system. For face tracking, we employ STCTracker [12, 13]; it includes an initialization part and particle filtering. The audio processing part employs a microphone array to capture the voices of the participants. Robust speaker diarization ("who speaks when" estimation) is carried out. Diarization is realized by a VAD (Voice Activity Detection) and DOA (Direction of Arrival) estimation followed by sound source clustering. Finally, the meeting processing part determines the utterance status (speaking or silent) of each meeting participant by cross-referencing the visual and audio information obtained in part (a) and (b). This data association is conducted by combining the face positions estimated by visual face tracking and the sound source locations as estimated by the audio processing part. Also, gaze direction (focus of attention) is estimated based on the position and direction of faces. This information is displayed on a monitor using our new visualization schemes.

3.1 Visual Processing

To estimate the face position and pose of each meeting participant in realtime, this paper employs STCTracker (Sparse Template Condensation Tracker) proposed in [12, 13]; its effectiveness for meeting analysis was verified in [17]. STCTracker consists of two main parts, initialization part and particle filtering part. According to [17], the advantages of STCTracker are its robustness against large head rotation, up to ±60 degrees in the horizontal direction, and its speed; it can track multiple faces simultaneously in real-time by utilizing a modern GPU (Graphics Processing Unit). Also, it can automatically build 3-D face templates upon initialization of the tracker.

In [17], STCTracker was applied to pre-recorded video sequences captured with normal cameras, which were carefully configured to catch each person's full face in each frame; large enough for precise tracking, but with enough room to permit reasonable head movement during conversations. In contrast, the system proposed in this paper newly incorporates a realtime STCTracker for simultaneously tracking people's faces on omnidirectional images captured in meetings.

3.1.1 Fisheye-based omnidirectional cameras

Fig. 3 shows our omnidirectional camera-microphone system. The camera part of the system 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. Our system captures only a horizontal strip as shown in Fig. 4(a), so that meeting participants are just covered by the image; this minimizes the transmission rate and affords high processing rates. Depending on lens type, image circle, and size of imaging device, two discontinuities and dead zones at right angles to the optical axes may occur. This is the main drawback of this camera system, and requires careful seating arrangements.

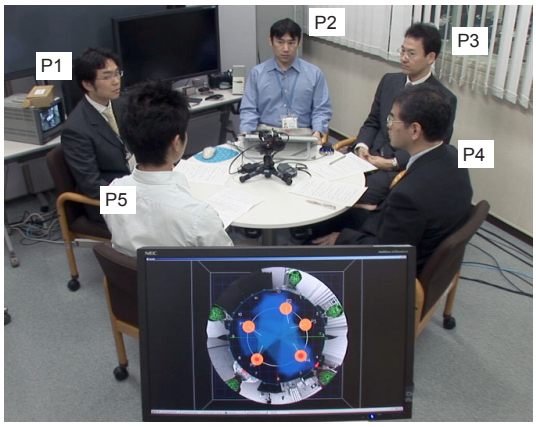

Figure 2: Meeting scene. LCD on near side shows the result of realtime processing.

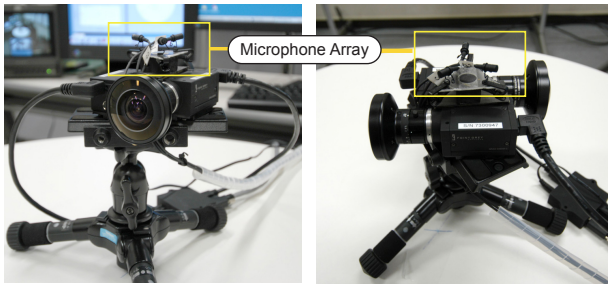

Figure 3: Omnidirectional camera-microphone system

Our system employs a fisheye lens with so called $f \cdot \theta$ projection, which is typical in fisheye lens. A point in the world is projected on to a point on the image plane; its distance from image center is proportional to the angle of incidence θ; f denotes the focal length. To resolve the distortion caused by this fisheye projection, and to obtain more suitable projection for face tracking, we perform a panoramic transformation. Fig. 4(b) shows an example of panorama images converted from fisheye images (Fig. 4(a)).

3.1.2 STCTracker

The basic idea of STCTracker is combining template matching with particle filtering. In contrast to traditional template matching, which assesses all pixels in a rectangular region, sparse template matching focuses on a sparse set of feature points within a template region. The state of a template, which represents the position and pose of the face, is defined as a 7 dimensional vector consisting of 2-DOF(Degree of Freedom) translation on the image plane, 3-DOF rotation, a scale (we assume weak-perspective projection), and an illumination coefficient. The particle filter is used to sequentially estimate the posterior density of the template state, which is represented as a particle set. The weight of each particle is calculated based on matching error between input images and the template whose state is assigned by each particle; higher weight is given to particles with smaller matching error. STCTracker has significant speed owing to the sparseness of the feature points and robustness owing to robust template matching combined with multiple-hypothesis generation/testing by the particle filter framework. Although the face model (template) is rigid, it can accept a certain amount of facial deformation caused events such by utterances and expression changes.

The upper-left part (a) of Fig. 1 shows the framework of STCTracker; it consists of initialization and particle filtering parts. The initialization stage detects faces in images, generates templates, and initializes the particles. The initialization stage first detects newly appeared frontal faces by using the Viola & Jones face detector [22]. Next, the Active Appearance Model (AAM) is used to locate facial parts and contours from the facial subimages detected by the face detector. AAM represents the combination of eigen face textures and eigen face shapes. Next, a set of feature points is extracted from each facial region. Finally, from the personalized 2-D AAM model and an average 3-D face model, the depth value of each feature point is calculated to form a face template (face model) for tracking. Note that before AAM and feature extraction, the detected face region is converted to its perspective projection equivalent.

Particle filtering consists of update stage, prediction stage, and averaging stage. The update stage calculates particle weight based on matching error of each template assigned by a particle state. The resulting particle distribution represents the posterior distribution of template states. The prediction stage resamples particles and predicts the particle distribution at the next time step. The update stage and prediction stage are alternately repeated for each image frame. The averaging stage calculates point statistics from the posterior particle distribution output by the update stage. Note that, in the update stage, when matching the face template against input images, the face template is projected as a panoramic image to compensate the difference in projection system between template and input image.

3.2 Audio Processing

Speaker diarization is done based on the audio signals from a microphone array as shown in Fig. 3. The array consists of three tiny microphones placed at the vertices of a triangle with 4cm sides, and is located atop the camera unit. To realize diarization, this paper employs methods proposed in [1]; the key parts are a noise-robust voice activity detector (VAD), a direction of arrival (DOA) estimator, and a DOA classifier. Traditionally, the GCC-PHAT technique [10] has been employed for estimating DOA in meeting situations as in ICSI and CHIL. However, the GCC-PHAT technique sets the constraint of just one DOA per frame and it often fails to detect speakers correctly in the case of overlapped speech. To avoid this problem, this paper employs the time-frequency domain DOA called TFDOA, which was recently proposed in [1].

The goal of speaker diarization is to determine "who is speaking" at each time step, from temporal signals captured by the microphone array. The observation process assumes that the signals observed by the microphones can be formed by convoluting the source signals and the impulse response of mixing process, and adding stationary background noise. That is, speaker diarization is the inverse of the observation process, i.e. find the source signal from the observed signal. Here, we assume that no prior knowledge is available as regards the number of people, speech sources, and the mixing process.

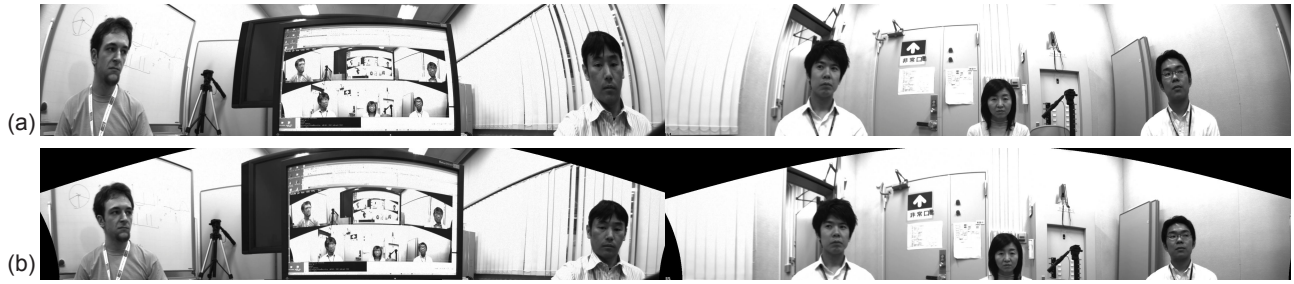

Figure 4: Camera images, (a)Fisheye images, (b)Panorama images. Two images from each camera are aligned side by side to form 360-degree view (image size = 4896×512 pixels).

The process flow of the diarization is shown in the lower-left part (b) of Fig. 1. First, short-term Fourier transformation (STFT) yields the time-frequency representation of the observation. Here f and τ denote frequency and time-frame index, respectively. Second, speech activity (human speech or noise) is estimated from a continuously observed signal by using a VAD. The speech activity of each sound source (potential speaker) is then determined by clustering/classifying the direction of arrival (DOA). The DOA is estimated by the TFDOA method [1].

3.2.1 Voice Activity Detection (VAD)

This paper uses a VAD method called "Multi Stream Combination of Likelihood Evolution of VAD" (MUSCLE-VAD)[6]. It employs two speech/non-speech discriminators; one is aperiodic component ratio-based detection (PARADE), the other is the SKF(switching Kalman filter)-based method. PARADE is robust against burst noise and SKF is robust against stationary and non-stationary noise. Therefore, this VAD method is robust against a wide variety of noises.

3.2.2 Direction of Arrival (DOA) estimation

The TFDOA(Time-frequency domain DOA) conducts DOA estimation for each time-frequency slot (f, τ), instead of the time domain DOA offered by GCC-PHAT [10]. TFDOA outputs DOA vectors $\boldsymbol{q}(f, \tau)$, whose component includes azimuth angle and elevation angle of arrival for each slot of frequency f and time frame τ. Next, an online clustering algorithm, known as leader-follower clustering, is applied to DOA vectors so that one cluster corresponds to one sound source, which should be a potential speaker. Finally, the speech activity of individual speakers is determined by thresholding distance between the newly detected sound source and the cluster's centroid.

3.3 Meeting Processing

The meeting processing part uses the outputs of the visual processing and audio processing parts. Currently, our system implements utterance detection and estimation of focus of attention. The presence/absence of utterances of each person is determined by combining the DOAs of speech from the audio processing part and the face positions from the vision processing part. This process is a data association problem; it aims to find the visual source responsible for utterances or noise. Here, we simply tackle this problem by the nearest neighbor rule with thresholding.

The face position/pose from the face tracker is used to estimate the visual focus of attention in the group. More specifically, this paper focuses on the discretized gaze direction of each person, i.e. looking at one person among all, or looking at no one. To estimate the gaze direction, we introduce a likelihood function that represents the distribution in his/her head direction when a person looks at a target. Here this paper employ a Gaussian function, which was also used in [16]. The method used in [16, 18] tries to estimate the parameters of the Gaussian likelihood function because the relative position of people is not available due to the lack of camera calibration. Although our system can not obtain the exact 3-D position of each person, it does provide angles relative to the camera system. In a round-table setting, we assume that the distance to each person from the camera is approximately the same.

Fig. 5 illustrates the relative position among people; the camera is located at the origin of this coordinate system. In Fig. 5, the position of each person is indicated by angles $h_i, (i = 1, \cdots, N)$, which can be obtained from horizontal position in the panorama image; N denotes the number of meeting participants. The horizontal head direction of each person can be represented as an angle; when a person points his/her face directly towards the camera, the angle of head direction is zero. Let $\varphi_{i,j}$ be the head direction of person P_i, when he/she looks straight at person P_j. This angle $\varphi_{i,j}$ can be calculated as

$$\varphi_{i,j} = -\tan^{-1}\left[1/\tan\left((h_i + h_j)/2\right)\right] \quad (1)$$

Using the face angle from one person to another, $\varphi_{i,j}$, the likelihood function of head direction $h_{i,t}$ of person P_i when person P_i is looking at person P_j can be written as,

$$L(h_{i,t}|X_{i,t} = j) := N(h_{i,t}|\kappa \cdot \varphi_{i,j}, \sigma^2), \quad (2)$$

where $X_{i,t}$ denotes the gaze direction of person P_i at time t, and $N(\cdot|\mu, \sigma^2)$ represent a Gaussian distribution with mean $\mu = \kappa \cdot \varphi_{i,j}$ and variance σ^2. In Eq.(2), κ denotes a constant (set here to 1). In addition, the likelihood function representing the person averting his/her gaze from everyone is defined as a uniform distribution. Using the likelihood functions, gaze direction is determined by the maximum likelihood scheme. Next, the focus of attention in the group at each time step can be determined by counting the number of gazes that each person receives from the others.

4. SYSTEM CONFIGURATION

Fig. 6 shows the hardware configuration of our system. The specifications of the vision-processing PC are as follows. CPU is Intel Core 2 Extreme QX9650 3.0GHz, and GPU is NVIDIA GeForce9800GX2 (two GPU cores are set in one package). OS is Windows XP SP2. The cameras are Point Grey Research's Grasshopper (B/W 5.0 Mega pixel model,

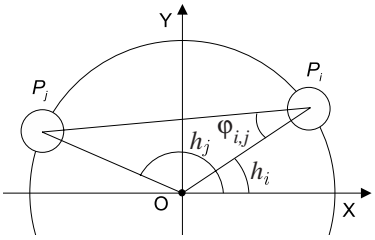

Figure 5: Spatial configuration of participants and their relative angles

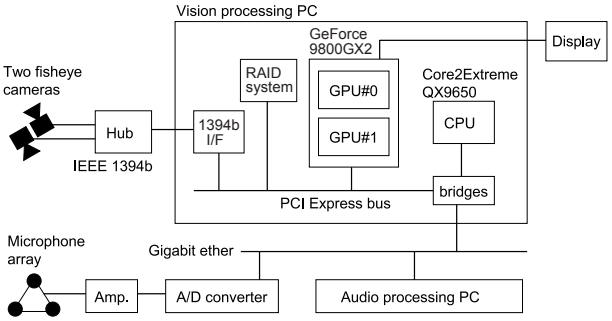

Figure 6: Hardware configuration

2/3" CCD). The fisheye lens is Fujinon's FE185C086HA-1(f=2.7mm). The camera and PC are connected by IEEE1394b links. The audio processing PC uses an AMD Athlon 64, 2.4 GHz, and its OS is Linux. The two PC's are connected via a Gigabit ether network. The basic program of the vision system was written in Microsoft Visual C++ 8. The GPU program was written in NVIDIA CUDA 1.1. The VAD component was written in C language and the speaker diarization part uses MATLAB6.5.

4.1 Data and Processing

Although the original image size of each camera is 2448 × 2048 pixels, we captured only a horizontal strip (2448 × 512 pixels) that covers the upper-body of meeting participants; this contributes to the achieved frame rate of 30.0 fps. Pixel depth was 8 bits (256-step grayscale). The two cameras were synchronized. The panorama projection from fisheye images was executed on a GPU core and face tracking was performed on another GPU core. The panorama images were written on a hard disk at 30.0 fps without any image compression. For audio processing, the sampling rate was 16 kHz and the frame size for STFT was 64ms; frame shift was 32 ms. For each frame, detected speech activity data and sound data were transmitted to the vision processing PC by TCP/IP over the Gigabit ether network.

5. EXPERIMENTS

Experiments were conducted to verify the performance of the proposed system. We targeted a round-table meeting with 5 participants. Fig. 2 shows the meeting environment and the participants. To observe and verify the realtime speed of the system, another camera system was set up to jointly capture both the participants' behaviors and a PC monitor that displayed the result of the realtime analysis, as shown in Fig. 2 [1]. Fig. 7 is taken from an actual screenshot of the PC display during a meeting. In Fig. 7, green meshes illustrate the result of face tracking, and the red dots along the axes indicate the DOA of voice. Also, speaking person is indicated by a red frame around his/her face. The number of particles for face tracking was 1500 per face. It was confirmed that the system ran at around 20 fps for 5 people; no system known to the authors has achieved comparable speed. The latency of visual processing (including image grabbing, panorama transform, tracking, and display) was about 170ms. The latency of audio processing (including A/D convert, STFT, DOA estimation, VAD, and transmission over ethernet) was about 80ms on average.

[1]Demonstration movies are available from our website http://www.brl.ntt.co.jp/people/otsuka/ICMI2008.html

5.1 Visualization

To visualize the conversation scenes for meeting observers, who could be remote meeting participants in the teleconferencing situation or users of meeting archive systems, two visualization schemes with a manually configurable interface were implemented, as shown in Fig. 8. Fig. 8(a) shows the cylindrical visualization of panoramic images and the relative position of each meeting participant (indicated as a circle). Also, Fig. 8(a) shows the approximate field of view as (blue) translucent triangles; overlapped fields of views indicate where people pay attention to each other. Moreover, the voice activity of each participant is displayed by the red dot in each person's circle in Fig. 8(a).

Fig. 8(b) is an example of the output of the second visualization scheme, called piecewise planar representation; the face image of each person is mapped to a planer surface, which is arranged to indicate the relative position of the participants. This visualization provides the viewers with larger face images which enables better understanding of the individual's expressions, while still clearly indicating their interpersonal positioning and interactive behaviors. In addition to field of view and voice activity included in Fig. 8(a), Fig. 8(b) shows discretized gaze directions of each person by arrows and the focus of attention, people who are attracting the gaze of more than a person, is indicated by a circle(s).

For both visualization schemes, our system offers a maneuverable interface by using a 3-D mouse, 3Dconnexion's SpaceNavigator. With this device, the users can freely and intuitively manipulate their viewpoints, as shown in Fig. 8(c) and 8(d). The rotation operation can choose the person (by literally rotating the knob of SpaceNavigator) and zooming operation can control the focus (by pushing/pulling the knob); from one-person (Fig. 8(d)) to all people, as in Fig. 8(b). Due to the high-resolution imaging provided by our new system, zoom-up face images retain sufficient details.

5.2 Quantitative Evaluation

To quantitatively evaluate system performance in a preliminary test, this paper targeted a 5-person conversation of 3 minute duration.

Table 1 shows the evaluation results of the speaker diarization. Table 1 includes diarization error rate (DER),

$$\text{DER} = \frac{\text{Wrongly estimated speaker time length}}{\text{Entire speaker time length}} \times 100[\%],$$

which has been established by NIST [15]. Table 1 also presents the diarization errors including the missed speaker

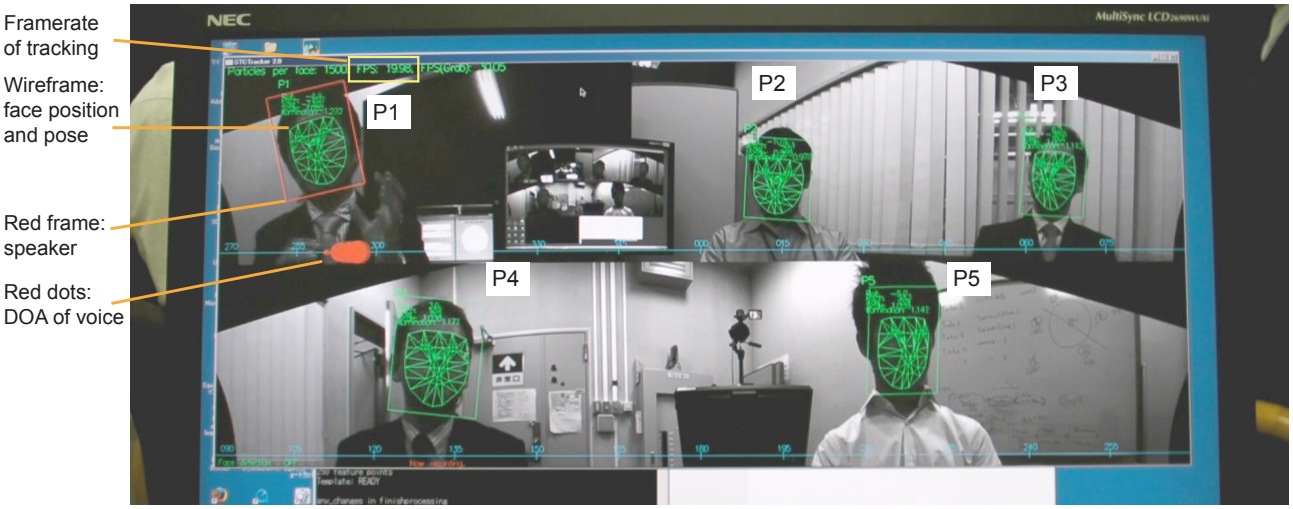

Figure 7: Screenshot of system monitor displaying face tracking and VAD results

Table 1: Evaluation results of speaker diarization[%]

DER	MST	FAT	SET
4.0	0.9	3.0	0.1

Table 2: Average accuracy of gaze directions[%]

All	P1	P2	P3	P4	P5
55.9	69.5	55.9	20.6	57.6	74.6

time (MST), the false alarm speaker time (FAT), and the speaker error time (SET). Table 1 confirms that our system can realize accurate diarization. The scores are at least comparable and can even outperform our previous work [1]. This is partially due to the fact that we targeted a formal-style meeting with less overlapped speech.

To evaluate the accuracy of gaze direction (VFOA) estimates, we first developed an annotation tool based on the visualization scheme proposed in this paper. The annotation was created by an annotator. Table 2 shows the frame-based accuracy of gaze direction, which is the ratio of frames in which gaze estimates coincided with the human annotation. Except for P3, the gaze estimates are reasonably successful given past studies [16, 18] that also estimated gaze direction from head direction. The main reason for the error is that humans can move his/her gaze without moving his/her head, e.g. looking sideways and casting down one's eyes. Also, when turning gaze from one to another, the eye moves first and the head then follows. The low accuracy of P3 results from the fact that he restlessly and meaninglessly let his eyes rove throughout the meetings. For all participants P1∼P5, about the half of the inconsistencies between the estimates and the human label were related to gaze aversion, i.e. the labels indicate looking at no one. About 45% of the inconsistencies were cases in which the person looked at was sitting next to the estimated target.

6. DISCUSSION AND CONCLUSION

This paper proposed a realtime system for multimodal meeting analysis by combining face pose tracking and speaker diarization. A novel data capturing device for group meetings was proposed based on two fisheye cameras that can provide omnidirectional views and a triangular microphone array. A realtime face tracker based on particle filtering and its GPU implementation provides face position and direction of each meeting participant; the data is used to estimate the focus of attention in meetings. From the acoustic signals captured by the microphone array, robust speaker diarization realized by VAD (Voice Activity Detection) and DOA (Direction of Arrival) estimation followed by sound source clustering. The system runs at about 20 frames per second for 5-person meetings, given two PCs, one for vision and one for audio processing.

Future works include the following. First, we need to increase the range of head rotation accepted and to better follow rapid head motions. Second, more robust and accurate diarization is required, because current VAD+DOA can be degraded by the reverberations created by various objects such as computer displays and white boards. In addition, it is necessary to conduct more comprehensive evaluations in various meeting scenes with different numbers of participants and different seating arrangements. Furthermore, it is necessary to verify the effectiveness of the visualization schemes, e.g. how clearly meeting content can be delivered to viewers.

Although the meeting analysis implemented in our current system is currently at a primitive level, the main contribution of this paper is to foster a new field, realtime multimodal meeting analysis. Authors believe that VFOA and utterance information are essential information for advanced meeting scene analysis such as the role of participants (speaker, addressees, and side-participants), conversation structures, dialogue act, floor control, and the detection of dominant speaker. Furthermore, it is important to discover and build real applications based on the output of meeting analysis, such as multimodal meeting archival systems, computer-mediated teleconferencing systems, and human-robot interaction.

7. REFERENCES

[1] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino. A DOA based speaker diarization system for real meetings. In *Proc. HSCMA2008*, pages 29–32, 2008.

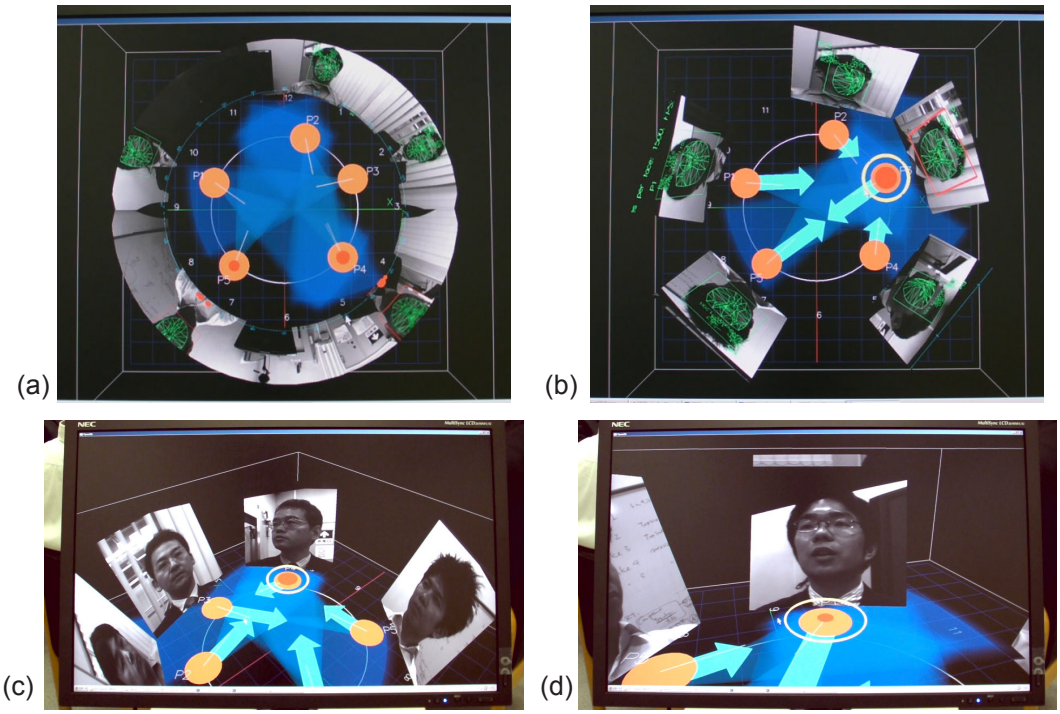

Figure 8: Visualization schemes, (a)Cylindrical visualization, (b)Piecewise planar visualization, (c)Viewpoint maneuver to middle range, (d)Viewpoint maneuver to close-up range

[2] M. Argyle. *Bodily Communication – 2nd ed.* Routledge, London and New York, 1988.

[3] S. O. Ba and J.-M. Odobez. A study on visual focus of attention recognition from head pose in a meeting room. In *Proc. MLMI2006*, pages 75–87, 2006.

[4] C. Busso, P. G. Georgiou, and S. S. Narayanan. Real-time monitoring of participants' interaction in a meeting using audio-visual sensors. In *Proc. ICASSP2007*, pages 685–688, 2007.

[5] D. Douxchamps and N. Campbell. Robust real time face tracking for the analysis of human behaviour. In *Proc. MLMI2007*, pages 1–10, 2007.

[6] M. Fujimoto, K. Ishizuka, and T. Nakatani. A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme. In *Proc. ICASSP2008*, pages 4441–4444, 2008.

[7] D. Gatica-Perez. Analyzing group interactions in conversations: a review. In *Proc. IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent Systems '06*, pages 41–46, 2006.

[8] D. Gatica-Perez, J.-M. Odobez, S. Ba, K. Smith, and G. Lathoud. Tracking people in meetings with particles. Technical Report IDIAP-RR 04-71, IDIAP, 2004.

[9] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[10] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. ASSP*, 24(4):320–327, 1976.

[11] L. Chen, et al. Vace multimodal meeting corpus. In *Proc. MLMI2006*, pages 40–51, 2006.

[12] O. Mateo Lozano and K. Otsuka. Real-time visual tracker by stream processing. *Journal of Signal Processing Systems*, DOI 10.1007/s11265-008-0250-2, 2008.

[13] O. Mateo Lozano and K. Otsuka. Simultaneous and fast 3D tracking of multiple faces in video by GPU-based stream processing. In *Proc. ICASSP2008*, pages 713–716, 2008.

[14] Y. Matsusaka, H. Asoh, and F. Asano. Multi human trajectory estimation using stochastic sampling and its application to meeting recognition. In *Proc. MVA2007*, pages 16–18, 2007.

[15] NIST Speech Group. Spring 2007 (RT-07) rich transcription meeting recognition evaluation plan. Technical Report rt07-meeting-eval-plan-v2, NIST, 2007.

[16] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *Proc. ICMI'05*, pages 191–198, 2005.

[17] K. Otsuka and J. Yamato. Fast and robust face tracking for analyzing multiparty face-to-face meetings. In *Proc. MLMI2008*, 2008.

[18] K. Otsuka, J. Yamato, and H. Murase. Conversation scene analysis with dynamic Bayesian network based on visual head tracking. In *Proc. ICME'06*, pages 949–952, 2006.

[19] S. Renals, T. Hain, and H. Bourlard. Interpretation of multiparty meetings the AMI and AMIDA projects. In *Proc. HSCMA2008*, pages 115–118, 2008.

[20] K. Smith, S. Schreiber, I. Potúcek, V. Beran, G. Rigoll, and D. Gatica-Perez. Real-time monitoring of participants' interaction in a meeting using audio-visual sensors. In *Proc. MLMI2006*, pages 88–101, 2006.

[21] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting index based on multiple cues. *IEEE Trans. Neural Networks*, 13(4), 2002.

[22] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[23] M. Voit and R. Stiefelhagen. Tracking head pose and focus of attention with multiple far-field cameras. In *Proc. ICMI2006*, pages 281–286, 2006.

[24] F. Wallhoff, M. Zobl, G. Rigoll, and I. Potucek. Face tracking in meeting room scenarios using omnidirectional views. In *Proc. ICPR2004*, 2004.