Smart Meeting Systems: A Survey of State-of-the-Art and Open Issues

ZHIWEN YU

Northwestern Polytechnical University, China

and

YUICHI NAKAMURA

Kyoto University, Japan

8

Smart meeting systems, which record meetings and analyze the generated audio–visual content for future viewing, have been a topic of great interest in recent years. A successful smart meeting system relies on various technologies, ranging from devices and algorithms to architecture. This article presents a condensed survey of existing research and technologies, including smart meeting system architecture, meeting capture, meeting recognition, semantic processing, and evaluation methods. It aims at providing an overview of underlying technologies to help understand the key design issues of such systems. This article also describes various open issues as possible ways to extend the capabilities of current smart meeting systems.

Categories and Subject Descriptors: H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces (GUI)*

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: Smart meeting system, meeting capture, meeting recognition, semantic processing, evaluation

ACM Reference Format:

Yu, Z. and Nakamura, Y. 2010. Smart meeting systems: A survey of state-of-the-art and open issues. ACM Comput. Surv. 42, 2, Article 8 (February 2010), 20 pages.
DOI = 10.1145/1667062.1667065, http://doi.acm.org/10.1145/1667062.1667065

1. INTRODUCTION

Meetings are important events in our daily lives for purposes of information distribution, information exchange, knowledge sharing, and knowledge creation. People are

This research was supported by the National Natural Science of China, the Program for New Century Excellent Talents in University, and the Ministry of Education, Culture, Sports, Science and Technology, Japan, under "Cyber Infrastructure for the Information-Explosion Era".

Authors' addresses: Z. Yu, School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, China, 710072; email: zhiwenyu@nwpu.edu.cn; Y. Nakamura, Academic Center for Computing and Media Studies, Kyoto University, Japan, 606-8501; email: yuichi@media.kyoto-u.ac.jp.

DOI 10.1145/1667062.1667065 http://doi.acm.org/10.1145/1667062.1667065

usually unable to attend all the meetings they need to, and even when they attend, they often forget important information presented at the meeting. A typical solution is to take notes during the meeting for later dissemination and recall. However, traditional note-taking is insufficient to store all the relevant meeting events; it is subjective, often incomplete, and inaccurate [Jaimes and Miyazaki 2005]. This precipitates the need to automatically record meetings on digital media such as digital memories for future viewing [Bell and Gemmell 2007].

Smart meeting systems are designed for exactly this purpose. They aim at archiving, analyzing, and summarizing a meeting so as to make the meeting process more efficient in its organization and viewing. Smart meeting systems have attracted much attention from researchers in recent years. Many systems have been developed (e.g., Waibel et al. [1998]; Mikic et al. [2000]; Chiu et al. [2000]; Rui et al. [2001]; Lee et al. [2002]; Jain et al. [2003]; Stanford et al. [2003]; Wellner et al. [2004]; and Janin et al. [2004]. Building a smart meeting system relies on a variety of technologies, ranging from physical capture and structural analysis to semantic processing. Such technologies also involve many domains, including image/speech processing, computer vision, human–computer interaction, and ubiquitous computing. These research topics comprise various aspects of a smart meeting. They cannot fulfill a smart meeting system by themselves, but must be organically integrated into such a system.

Smart meeting systems are still not deployed very widely in our real-life settings. Several problems remain, such as a low recognition rate, lack of infrastructure, and limited features. Future trends should make them more robust, scalable, active, safe, and easily deployable. Hence, open issues such as improvement of the current technologies, infrastructure, active capturing, context awareness, real-time feedback, meeting activation, distributed meetings, and security and privacy, must be explored.

This article presents a survey of the state-of-the-art of smart meeting systems. The objective of this article is twofold. First is to provide an overview of underlying technologies so that researchers in the smart meeting domain can understand the key design issues of such a system; and second to present several open issues as possible ways to extend the capabilities of current systems.

This article is organized as follows. First, a survey of current research and technologies is presented. Sections 2, 3, 4, 5, and 6 describe smart meeting system architecture, meeting capture, meeting recognition, semantic processing, and evaluation methods, respectively. Open issues are then discussed in Section 7, and conclusions are provided in Section 8.

2. ARCHITECTURE OF A SMART MEETING SYSTEM

A generic architecture with basic modules of a smart meeting system is shown in Figure 1. In this design, Meeting Capture is the physical level that deals with capturing environment, devices, and methods. It records three major types of data: video, audio, and other context data. Meeting Recognition serves as the structural level, which analyzes the content of the generated audio-visual data. It mainly involves person identification, speech recognition, summarization, attention detection, hot spot recognition, and activity recognition. The recognition layer makes the meeting content meaningful and provides support for Semantic Processing, the semantic level. Semantic Processing handles high-level manipulations on the semantics such as meeting annotation, indexing, and browsing. Technologies for implementing these basic modules are described in detail in Sections 3, 4, and 5.

Here, we introduce several representative architectures of existing smart meeting systems developed by Carnegie Mellon University [Waibel et al. 1998], Microsoft [Rui

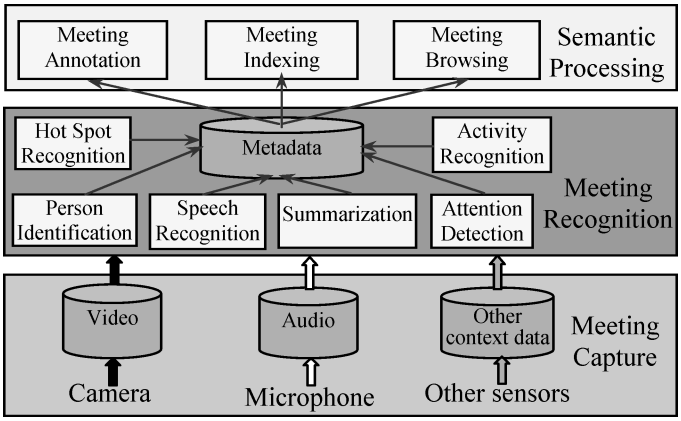

Fig. 1. Generic architecture of a smart meeting system.

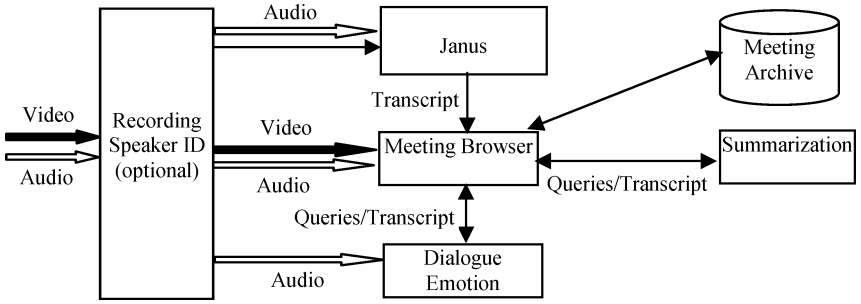

Fig. 2. Architecture of Carnegie Mellon University's smart meeting system [Waibel et al. 2001].

et al. 2001], University of California, San Diego [Mikic et al. 2000], and Ricoh Innovations Inc. [Lee et al. 2002]. Other important systems include the FXPAL's conference room [Chiu et al. 2000], Georgia Tech's TeamSpace [Richter et al. 2001] and Experiential Meeting System [Jain et al. 2003], the European Ferret system [Wellner et al. 2004], the ICSI's Meeting Corpus [Janin et al. 2004], and the NIST's Meeting Room [Stanford et al. 2003]. Particular components might be different, but the general concept of these systems' frameworks is similar to our generic architecture.

Carnegie Mellon University might be the first to propose the concept of meeting browser and design a smart meeting system. Figure 2 shows the components of its meeting room system [Waibel et al. 1998; 2001; Schultz et al. 2001]. It records a meeting with camera and microphone. Recognition functions include speaker identification, transcription, dialogue and emotion detection, and summarization. The meeting browser displays results from the speech recognizer (Janus), emotion and dialogue processor, summarization module; video, audio, and a meeting archive can also be navigated.

Microsoft developed a smart meeting system with a client–server architecture [Cutler et al. 2002; Rui et al. 2001], as shown in Figure 3. RingCam, an omnidirectional camera with an integrated microphone array, is designed for capturing meetings. The camera has a high resolution ($1300 \times 1030 = 1.3$ megapixels), and tracks meeting participants as well as captures video at 11 fps. The meeting server performs all

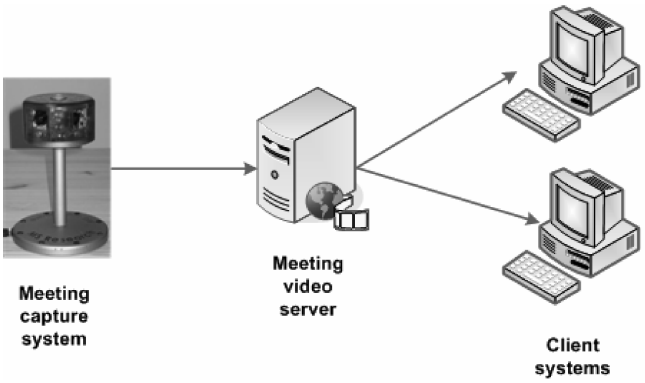

Fig. 3. Microsoft's smart meeting system architecture [Rui et al. 2001].

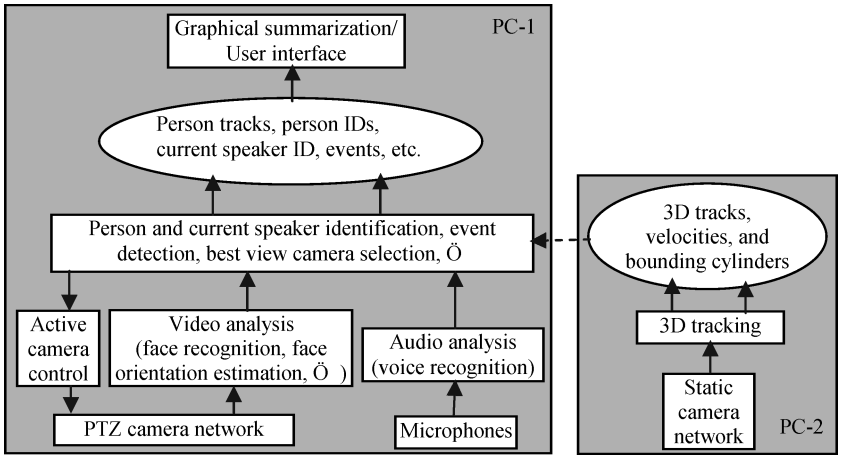

Fig. 4. Block diagram of AVIARY system [Mikic et al. 2000].

processing required to broadcast, encode, and recognize meetings. Recognition modules include sound source localization, person detection and tracking, and speaker segmentation and clustering. The meeting browser is deployed at the client end. It provides users with customized viewing of meeting records: seeing all the meeting participants all the time, seeing the active participant only, controlling the camera themselves, or allowing the computer to take control. For the purpose of broadcast, the researchers in Microsoft also did some work to improve audio quality such as beamforming and noise reduction.

The AVIARY system [Mikic et al. 2000; Trivedi et al. 2000] developed by University of California, San Diego, uses two PCs, four static cameras, four active cameras, and two microphones for meeting capture. The block diagram is shown in Figure 4. One PC performs 3D tracking based on input from four static cameras. Results such as centroid, velocity, and bounding cylinder information are sent to the other PC, which is connected to a Pan Tilt Zoom (PTZ) camera network and microphones. It recognizes meeting data at two levels—the signal analysis and integrated analysis levels. The signal analysis level performs face recognition, face orientation estimation, and audio source localization. The integrated analysis level performs person identification, event detection, and camera selection in an integrated manner. Final results such as person

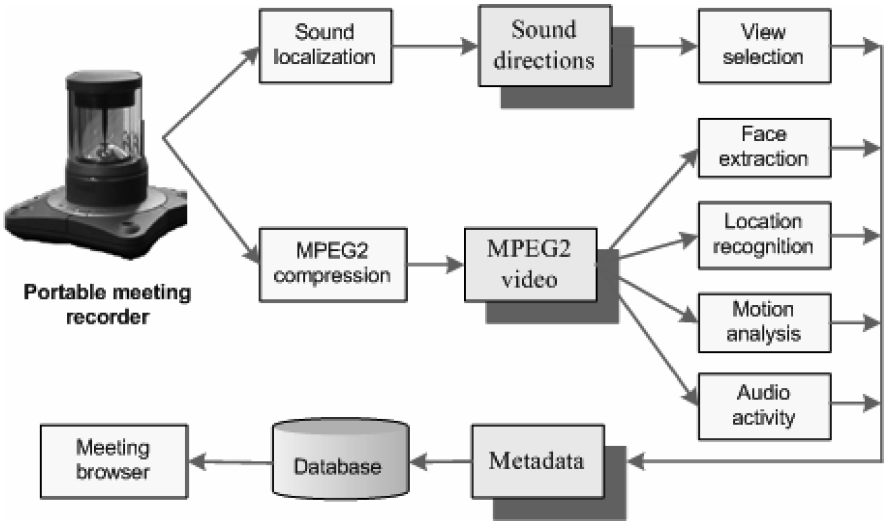

Fig. 5. System architecture of portable meeting recorder [Lee et al. 2002].

tracks, person IDs, and events are stored in a semantic database and retrieved later using a graphical user interface for browsing.

Ricoh Innovations, Inc. proposed a portable meeting recorder [Lee et al. 2002], as shown in Figure 5. A special device, composed of an omnidirectional camera in the center and four microphones positioned at the corners, is designed for capturing. Audio signals are processed in real time to determine the direction of speakers. Video data is used for face extraction, location recognition, motion analysis, and audio activity detection. Metadata is created from the recognition results. Meeting recordings and relevant metadata are presented to a meeting browser for viewing. The user can browse a meeting by reading the description page, listening only to the speakers that he or she is interested in, looking at the high-motion parts, searching for keywords in the transcription, looking at the presentation slides and whiteboard images, and so on.

3. MEETING CAPTURE

To capture a meeting, a smart environment must be equipped with a variety of sensors, such as cameras, microphones, motion sensors, and radio frequency identification (RFID). Sensor selection and set-up generally depends on the objective of the capture system and the way the contents will be used (e.g., speaker identification, activity detection, affective analysis, speech recognition) [Jaimes and Miyazaki 2005]. Meeting capture focuses on three major types of data that are valuable for meeting analysis and viewing: video data, audio data, and other context data derived from sensors (e.g., location, motion, and whiteboard notes).

3.1. Video Capture

Video plays a major role in capturing a meeting. It can record humans and any other objects including documents, whiteboards, and presentations. Four different types of cameras can be used to capture video data: static camera, moveable camera, camera array, and omnidirectional camera.

A static camera can capture only a fixed view with a constant angle and distance. A moveable camera, also called a PTZ camera, can pan, tilt, and zoom. Since either of these types can cover only a limited area, recently most researchers have tended

to use a camera array or an omnidirectional camera. A camera array is composed of multiple cameras, static or moveable, and can capture multiple views. FlyCam [Foote and Kimber 2000] uses four cameras to construct a panoramic view of the meeting room. Other systems that utilize a camera array include those of Yang et al. [1999]; Trivedi et al. [2000]; Jaimes et al. [2004]; Bounif et al. [2004]; and McCowan et al. [2005]. Recent advances in omnidirectional cameras have inspired many researchers to use them for meeting capture Lee et al. [2002]; Chen and Yang [2002]; Busso et al. [2005]. Omnidirectional cameras can capture a 360 degree view of a meeting, and they have been widely used in the computer vision research community for 3D reconstruction, visualization, surveillance, and navigation [Rui et al. 2001].

For efficient video capture, Rui et al. [2004] propose several video production rules to guide camera placement, including rules for camera positioning, lecturer tracking and framing, audience tracking and framing, and shot transition. Other studies that deal with optimal camera placement and control include Horster and Lienhart [2006]; Ram et al. [2006]; Zhao and Cheung [2007]; and Singh et al. [2007].

3.2. Audio Capture

Microphones are used to capture meeting audio data. Capturing high-quality audio in a meeting room is challenging, since it requires that various noises and reverberation be removed and that the gain be adjusted for different levels of input signal [Cutler et al. 2002]. In general, these issues can be addressed by choosing the types, locations, and number of microphones.

There are six basic types of microphones: cardioid, omnidirectional, unidirectional, hypercardioid, supercardioid, and bidirectional [Mboss 2007]. System requirements determine which type of microphone should be used. For instance, if the application needs all-around pick-up, pick-up of room reverberation, low sensitivity to pop (explosive breath sounds), low handling noise, and extended low frequency, it is appropriate to use omnidirectional microphones.

In general, microphones can either be attached to the ceiling [Chiu et al. 2000], placed on the table [AMI 2007], or worn by attendees Kern et al. [2003]; Janin et al. [2004]. Using close-up or wearable microphones reproduces sound accurately and is to use simple but is intrusive. Several systems integrate cameras and microphones within a single device placed in the center of the table to record video and audio simultaneously. RingCam Cutler et al. [2002] consists of a 360 degree camera and an 8-element microphone array at its base, which is used for beamforming and sound source localization. The portable meeting recorder [Lee et al. 2002] is composed of an omnidirectional camera in the center and four microphones positioned at the corners. Budnikov et al. [2004] and Raykar et al. [2005], provide solutions for common problems in microphone array systems, such as determining locations and synchronizing audio sampling across different devices.

3.3. Other Context Capture

Besides video and audio data, context also provides important data for smart meetings, and is useful in generating annotations for audio-visual meeting records. Furthermore, human memory of a meeting could be enhanced by contextual information such as the weather and light Kern et al. [2003]; Jaimes et al. [2004]. Typical context in a meeting room includes the locations, motions, seat positions, emotions, and interests of participants as well as the room layout, meeting agenda, and whiteboard notes. This data can be captured through other types of sensors (e.g., RFID, pen strokes, head-tracking sensors, motion sensors, physiological sensors, lighting sensors, pressure

sensors, etc.). The advantages of using such sensors lie in their small size, low cost, and easy processing.

Kaplan [2005] employs RFID to detect who is present at the meeting and who is making a presentation. The Conference Assistant [Dey et al. 1999] uses radio-frequency-based tags with unique identities and multiple antennas to determine when users enter the meeting room. Liwicki et al. [2006] utilize a normal pen in a special casing to write on an electronic whiteboard and then acquire the written text. Mimio [2007] tracks the location of a pen at high frequency and infers the content of the whiteboard from the history of the pen coordinates. Equipped with a magnetic pose and position tracking mechanism, a head-mounted ISCAN system [Stiefelhagen and Zhu 2002] detects the subject's head position and orientation, eye orientation, eye blinks, and overall gaze (line of sight). To detect postures and movement in parts of the body, Kern et al. [2003] deploy a network of 3-axis accelerometers distributed over the user's body. Each accelerometer provides information about the orientation and movement of the corresponding body part. Aizawa et al. [2001] adopt a wearable sensor to capture physiological signals, that is, brain waves, that are used to measure the wearer's level of interest, and to evaluate video shots for summarization.

4. MEETING RECOGNITION

Meeting recognition is responsible for low-level analysis of video data, audio data, and context captured in a meeting environment. It detects structural features from the meeting content and provides support for high-level semantic processing such as indexing, retrieval, and browsing.

4.1. Person Identification

Person identification addresses the problem of who is doing what at a meeting, for example, speaking or writing. Face, speaker, and writer identification is often used to achieve this goal.

Numerous face recognition algorithms have been introduced in recent years. The Eigenface approach [Turk and Pentland 1991] is one of the most widely used. The major challenges in identifying human faces in a meeting room include low-quality input images, poor illumination, unrestricted head poses, and continuously changing facial expressions and occlusion [Gross et al. 2000]. Gross et al. [2000] propose the dynamic space warping (DSW) algorithm, which combines local features under certain spatial constraints and specifically addresses the problem of occlusion.

Speaker identification aims at knowing which participant is talking and where that person is located. Many systems use audio-based sound source localization (SSL) to locate the speaker [Rui 2002; Valin et al. 2003; Liu et al. 2007].

Writer identification is useful to determine who is writing. Liwicki et al. [2006] identify the writer based on handwriting data acquired through an electronic whiteboard.

Since the accuracy of any single identification method is usually not high, many researchers combine two or more approaches for the purpose of identification. Mikic et al. [2000]; Cutler et al. [2002]; and Busso et al. [2005] integrate face and speaker recognition for robust person identification. Yang et al. [1999] identify participants in a meeting by fusing multimodal inputs, including face ID, speaker ID, color ID, and sound source direction ID.

4.2. Speech Recognition

Speech recognition, also called transcription, determines the content of a speaker's speech. Using a single microphone per person for transcription is problematic due to

significant overlapping speech from other participants. Thus, microphone arrays are widely used in meeting applications for their ability to discriminate among multiple competing speakers based on their location [Moore and McCowan 2003; Busso et al. 2005]. Other challenges in transcribing meetings lie in their highly conversational and noisy nature and the lack of domain-specific training data [Yu et al. 1999].

Waibel et al. [1998] first developed a speech transcription engine based on the JANUS recognition toolkit. They then used vocal tract-length normalization and cluster-based cepstral mean normalization to compensate for speaker and channel variations [Waibel et al. 2001]. Baron et al. [2002] adopt lexical and prosodic features to detect sentence boundaries and disfluency interruption points in meetings. To access information from the audio data streams that result from multichannel recordings of meetings, Renals and Ellis [2003] utilize word-level transcriptions and information derived from models of speaker activity and speaker turn patterns. Akita et al. [2006] present an automatic transcription system for meetings of the National Congress of Japan. As topics change frequently at such meetings, they use two approaches for topic adaptation: a PLSA-based approach and a trigger-based one. The former is based on probabilistic latent semantic analysis (PLSA) that performs adaptations turn-by-turn-to, emphasize topics in individual pairs consisting of a question and an answer. The latter adaptation stresses words relevant to a history; thus, a long-distance context can be reflected in a language model.

4.3. Summarization

There are two types of meeting summaries. One covers the entire meeting, that is, it provides a meeting summary in a user interface for quick review [Mikic et al. 2000]. This mainly involves visualization or browsing of a meeting (which is discussed separately in Section 5.3). The other type is speech summarization followed by speech transcription. Speech summarization is used to generate condensed informative summaries of a meeting based on transcripts generated manually or automatically by recognition mechanisms. Waibel et al. [2001] propose a summarization system for audio data access. It consists of five major components, including disfluency detection and removal, sentence boundary detection, detection of question–answer pairs, relevance ranking with word error rate minimization, and topic segmentation. Several approaches for automatic speech summarization are discussed in Murray et al. [2005], such as maximal marginal relevance (MMR), latent semantic analysis (LSA), and feature-based approaches. MMR and LSA are borrowed from the field of text summarization, while feature-based methods using prosodic features are able to utilize characteristics unique to speech data. They perform well in different meeting summarization situations. For example, MMR that is good at query-based summarization and multidocument summarization is suitable for users who wish to query-based summaries of meetings.

4.4. Attention Detection

Attention detection addresses the problem of tracking the focus of attention of participants in a meeting, that is, detecting who is looking at what or whom during a meeting [Stiefelhagen 2002]. Knowing the focus of attention is useful for understanding interactions among objects within a meeting and to index multimedia meeting recordings. Attention detection involves recognizing head and eye orientation. For example, eye-tracking needs to be combined with tracking head-position to determine if one person is looking at another, and especially when the other person is talking. Stiefelhagen et al. [1999] first employ hidden Markov models (HMMs) to characterize the participants' focus of attention, using gaze information as well as knowledge about the number and

positions of people in a meeting. They then use microphones to detect who is speaking currently, and combine the acoustic cues with visual information to track the focus of attention in meeting situations [Stiefelhagen et al. 2001; Stiefelhagen 2002; Stiefelhagen et al. 2002]. This approach has been verified as a more accurate and robust estimation of the participants' focus of attention.

4.5. Hot Spot Recognition

It is desirable to provide the most informative and important parts of a meeting to users who are browsing the meeting. To do so we must be able to recognize hot spots. Hot spots refer to parts of a discussion in which participants are highly involved (e.g., heated arguments, points of excitement, etc.) [Wrede and Shriberg 2003a; 2003b]. Gatica-Perez et al. [2005], use the concept of the level of group interest to define relevance or the degree of engagement that meeting participants display as a group during their interaction. Wrede and Shriberg [2003b] find that there are relationships between dialogue acts (DAs) and hot spots; involvement is also associated with contextual features such as the speaker or the type of meeting. Acoustic cues such as heightened pitch have been employed successfully for automatic detection of hot spots in meetings [Wrede and Shriberg 2003a; Kennedy and Ellis 2003]. Gatica-Perez et al. [2005] propose a methodology based on HMMs to automatically detect high-interest segments from a number of audio and visual features. Yu et al. [2007] use the degree of crossover of utterances and gaze changes, as well as the number of nods and responses, as indicators for recognizing hot spots at a meeting.

4.6. Activity Recognition

Activity recognition is of at most importance for understanding a meeting, but it is also the most complicated, as it involves multiple modalities and a variety of the technologies described above. Meeting activity can mainly be divided into two classes, individual and group activities.

Individual activity consists of the actions or behavior of a single person. Zobl et al. [2003] detect and recognize single-person actions at a meeting, including sitting down, getting up, nodding, shaking the head, and raising the hand, via image processing. The system consists of three processing steps: feature extraction, stream segmentation, and statistical classification. Mikic et al. [2000] identify three types of activities: a person standing in front of the whiteboard, a lead presenter speaking, and other participants speaking, on the basis of voice recognition, person identification, and localization. Individual activity can also be inferred through postures and body-part motions [Kern et al. 2003]. For example, a person presenting a talk is likely to be standing up, possibly slowly walking back and forth, moving his arms, and gesticulating. Actions such as entering, exiting, going to the whiteboard, getting up, and sitting down can be recognized via head tracking [Nait-Charif and McKenna 2003].

Due to the group nature of meetings, it is important to recognize the actions of the group as a whole, rather than simply those of individual participants [McCowan et al. 2005]. Group activities are performed by most of the participants in a meeting. They can be recognized by directly segmenting the meeting content or indirectly deducing from individual actions. For example, Dielmann and Renals [2004] and Reiter and Rigoll [2005] segment meetings into a sequence of events: monologue, discussion, group note-taking, presentation, and presentation at the whiteboard, among which discussions and group note-taking are group activities. Kennedy and Ellis [2004] detect laughter events where a number of participants laugh simultaneously by using a support vector machine (SVM) classifier trained on mel-frequency cepstral coefficients (MFCCs), delta

MFCCs, modulation spectra, and spatial cues from the time delay between two desktop microphones. McCowan et al. [2005] recognize group actions in meetings by modeling the joint behavior of participants based on a two-layer HMM framework.

Human social activity outside of the physical behaviors and actions discussed above has received increasing attention recently. Meetings encapsulate a large amount of social and communication information. Hillard et al. [2003] propose a classifier for the recognition of an individual's agreement or disagreement utterances using lexical and prosodic cues. The Discussion Ontology [Tomobe and Nagao 2006] was proposed for obtaining knowledge such as a statement's intention and the discussion flow at meetings. Both systems are based on speech transcription that extracts features such as heuristic word types and counts. Yu et al. [2008a] propose a multimodal approach for recognizing human social behaviors such as proposing an idea, commenting, expressing a positive opinion, and requesting information. A variety of features, including head gestures, attention, speech tone, speaking time, interaction occasion (spontaneous or reactive), and information about the previous interaction, are used. An SVM classifier is adopted to classify human social behaviors based on these features.

5. SEMANTIC PROCESSING

Given the recognized features, high-level semantic manipulations are required to quickly access information of interest from the meeting record. Such semantic processing technologies include meeting annotation, indexing, and browsing.

5.1. Meeting Annotation

Annotations play an important role in describing raw data from various viewpoints and enhancing the querying and browsing process [Bounif et al. 2004]. Meeting annotation is the process of creating labels for the meeting shots. It consists of acquiring descriptors and then labeling data with them. To acquire annotations, explicit and implicit approaches can be used [Geyer et al. 2003]. Explicit annotation capture refers to users manually detecting people, places, and events in the shot. On the other hand, implicit capture extracts annotations automatically by using sensors [Kern et al. 2003] or media recognition [Gatica-Perez et al. 2003]. To assign labels in accordance with the meeting data, Reidsma et al. [2004] present several annotation schemas, for example, manual annotation, efficient interface for manual annotation (if manual annotation is unavoidable, the efficiency of creating this annotation depends heavily on the user interface of the annotation tool), semiautomatic annotation (e.g., hand-gesture labeling), and integrating annotations from a third party. Techniques for automatic image annotation [Datta et al. 2008] would help to relieve the human burden in the process of meeting annotation.

5.2. Meeting Indexing

There are different levels of indexing meeting records. At the raw data level, a temporal search can be resolved on a sensor stream along the time axis. However, finding detailed information in a meeting record is difficult when using only time to aid navigation [Geyer et al. 2005]. With recognized data and annotations, a semantic database is required to store them and retain links with raw data (audio and video). Higher-level indexing is required for efficient organization, management, and storage of the meeting semantics. Geyer et al. [2005] discuss various ways for indexing meeting records (e.g., online and offline, explicit versus derived) and then propose creating indices based upon user interactions with domain-specific artifacts. Bounif et al. [2004] adopt a metadictionary structure to manage annotations and create various indexes to semantics

(e.g., meeting overview, participant, document, and utterance). Due to its importance and popularity, the event is widely used as an index to access meeting information. For example, Jain et al. [2003] propose event-based indexing for a experiential meeting system, in which events are organized at three levels: domain, elemental, and data. The AVIARY database [Trivedi et al. 2000] stores semantic events of associated environmental entities and uses them to retrieve semantic activities and video instances.

5.3. Meeting Browsing

Meeting browsing acts as the interface between the smart meeting system and end users. It consists of visualizing, navigating, and querying meeting semantics as well as media content. Waibel et al. [1998] first proposed the concept of a meeting browser for summarizing and navigating meeting content. Tucher and Whittaker [2004] present a survey of browsing tools and classify them into four groups: audio, video, artifact, and discourse browsers, according to their focus on navigation or attention. Here, we categorize meeting browsers into two classes, depending on whether they are designed with basic browsing functions (visualization and navigation) or enhanced with more functionality (e.g., querying and retrieval).

From the perspective of visualization, graphical user interfaces are designed to present various kinds of information with various representations (colors and shapes) for easy review and browsing. Mikic et al. [2000] use 3D graphic representations for events and participants in an intelligent meeting room. Ferret [Wellner et al. 2004] provides interactive browsing and playback of many types of meeting data, including media, transcripts, and processing results, such as speaker segmentation. DiMicco et al. [2006] present visualization systems for reviewing a group's interaction dynamics (e.g., speaking time, gaze behavior, turn-taking patterns, and overlapping speech at meetings). To review and understand human interactions efficiently, Yu et al. [2008b] present a multimodal meeting browser for interaction visualization. It mostly offers visualization of group social dynamics, such as human interaction (activity level), speaking time (participation level), and attention from others (importance level). The media content can also be navigated. The browser helps people to quickly understand meeting content (or topics) and human interactions in a multimodal and interactive manner via graphics, video images, speech transcription, and video playback. Other similar systems include Bett et al. [2000]; Lee et al. [2002]; and Geyer et al. [2003].

To enhance meeting navigation, query and retrieval functions are generally added. Rough'n'Ready [Colbath and Kubala 1998] provides audio data browsing and multivalued queries (e.g., specifying keywords as topics, names, or speakers). Jaimes et al. [2004] propose a meeting browser enhanced by human memory-based video retrieval. It graphically represents important memory retrieval cues, such as room layout, the participants' faces and seating positions. Queries are performed dynamically, that is, as the user manipulates the cues graphically, the query is executed, and the results are shown. For instance, moving the face icon to a location in the room layout immediately shows all keyframes for videos in which the person was sitting at the specified location. Thus, the system helps users to easily retrieve video segments by using memory cues.

6. EVALUATION METHODS

The criteria used to evaluate a smart meeting system include user acceptance, accuracy, and efficiency. User acceptance is measured by conducting user studies and analyzing user feedback. Accuracy tests the recognition mechanisms by means of different objective methods. Efficiency is usually employed to evaluate a meeting browser, that is, whether it is useful for understanding the meeting content quickly and correctly.

Rui et al. [2001] conducted user studies to test their meeting interface. The subjects were asked to rate a variety of statements (e.g., "I like this user controlled interface," the computer did a good job of controlling the camera, the overview window was helpful," etc). The disadvantage of user studies lies in their subjective nature; user tasks and questions are often loosely defined, and hence final scores are open to considerable interpretation [Wellner et al. 2005].

Other than user studies, most current smart meeting systems use objective methods to evaluate the accuracy of their recognition mechanisms. Evaluation metrics can be categorized as the recognition rate, precision and recall, and other measures. The recognition rate is the ratio between the number of correctly recognized objects and the total number of objects. For example, Liwicki et al. [2006] use the recognition rate to evaluate writer identification based on different features. Gross et al. [2000] compare the recognition rates of the classical Eigenface approach and DSW algorithm in face recognition. Zobl et al. [2003] calculate an average recognition rate to evaluate the action recognition system. Some systems, for example, Kennedy and Ellis [2004] and Gatica-Perez et al. [2005], borrow precision and recall [Rijsbergen 1979] from the field of information retrieval to evaluate recognition performance. In general, precision can be used as a measure of the ability of the system to identify correct objects only. Recall is used to test the ability of the system to recognize all correct objects. Besides the above methods, some systems define special measures of their own. For example, Zhang et al. [2004] use the action error rate (AER) and the frame error rate (FER) as criteria to evaluate the results of group action recognition and individual action recognition, respectively. AER is defined as the sum of the insertion (Ins), deletion (Del), and substitution (Subs) errors divided by the total number of actions in the ground truth. On the other hand, FER is defined as one minus the ratio between the number of correctly recognized frames and the number of total frames, which is similar to the recognition rate in nature.

To assess the efficiency of meeting browsers, Wellner et al. [2005] propose a browser evaluation test (BET). It uses the number of observations of interest found in the minimum amount of time as the evaluation metric. The method first instructs observers to view a meeting and produce a number of observations. Second, both true and false statements of the observations are presented as test questions. Third, subjects are invited to use the browser to review the meeting and answer as many test questions as possible in a short time. Finally, the method compares the subjects' test answers to the original observations, and computes a score for the browser. This method is objective, independent, and repeatable. It may help the researchers evaluate smart meeting systems from a different perspective.

7. OPEN ISSUES

Although many smart meeting systems have been developed in the past decade, they still require further improvements to extend the effectiveness of current capabilities and to be more applicable in real-life settings. Open issues include improving current technologies, offering infrastructure support, and providing more features such as active capturing, context-awareness, real-time browsing, meeting activation, distributed meetings, and security and privacy.

7.1. Improving Current Technologies

Due to the dynamics and diversity of meetings, processing meeting data automatically is challenging [Yu et al. 1999; Cutler et al. 2002; Geyer et al. 2005]. Most of the current techniques described in Sections 3, 4, and 5 are lacking in many respects. Recognition

rates for speech, events, interaction, and person identification are sometimes low in real-world settings. This is one reason that smart meeting rooms are still not very widely used. Hence, one important open issue is to improve current technologies to a level that is robust and usable in everyday settings. Multimodal approaches for sensing and recognition would be promising [Yang et al. 1999; Bett et al. 2000; Bounif et al. 2004; McCowan et al. 2005]. The purpose of multimodal sensing is to collect rich and complete information about a meeting. Multimodal recognition achieves robust and reliable results, as it recognizes meeting objects and events from multiple aspects.

7.2. Infrastructure Support

Smart meeting applications usually involve a broad spectrum of sensors and software for capture and recognition, which makes the development of a smart meeting system difficult [Jaimes and Miyazaki 2005]. While most current studies on smart meetings have been focused on developing particular techniques to process the generated audio–visual content automatically, infrastructure is required to integrate such devices and software. This will provide middleware support for rapid application development and deployment. Scalability is critical in building smart meeting infrastructures. Since a variety of sensors (e.g., cameras, microphones, motion sensors, RFID, etc.), are used to capture a meeting, the system must scale up to a large number of devices. Furthermore, the amount of data to be stored and accessed may be very large. It is essential to provide efficient processing mechanisms that can handle many devices and large amounts of data as the size of the meeting space grows. High-level abstraction of sensors and programs (e.g., capturers, accessors, storers, or transducers) [Truong and Abowd 2004; Pimentel et al. 2007] will be useful in the design of a smart meeting infrastructure.

7.3. Active Capturing

Current meeting recordings are passive, that is, sensors are fixed and preset to capture given objects and scenes throughout the meeting. Although some kinds of structural features, such as who is where at what time, could be recognized, more semantic information is often necessary. Such semantic information (e.g., what a person is doing and why, what is important, or what should be paid attention to) is particularly important, and is often difficult to obtain if a meeting is captured passively. Furthermore, people do not know what is captured and how the recording is going on unless they watch the video through a monitor. Systems should enable the devices to actively record the objects that the user is currently interested in and wants recorded. Such active capturing can help to record the right content at the right time according to the user's requirements. The Virtual Assistant [Ozeki et al. 2008] enhances content acquisition by allowing users to interact with the capture system through an artificial agent that prompts users to provide feedback by asking questions or back-channeling. It would be useful to explore this direction.

7.4. Context-Awareness

User context plays an important role in determining what information which using media types and in which sequence should be presented to which user [Jain 2003]. Existing smart meeting systems provide the same browsing interface and content to the users. However, what a user is interested in and wants to watch largely depends on the user's current contexts (e.g., user purpose, role, location, and seat position). For example, user purpose plays an important factor in viewing meetings. If the user just wants to know what the conclusion was, he or she would merely focus on the result; but if the user wants to know how the conclusion was reached, he or she may be interested

in the procedure, event, and human communication (e.g., did all members agree on the outcome, how long did it take before a decision was made, did everyone have the chance to give an opinion etc.). Context-aware browsing of meetings will exploit various contexts to present different meeting information to different users. It essentially represents a query to the meeting record that answers by presenting the results in a form users find desirable and immediately usable. Personalized event experience [Jain 2003] might be helpful to achieve context awareness in meetings. It allows users to explore and experience events from multiple perspectives and revisit these events from their own perspective. Context-aware recommendation [Yu et al. 2006] that performs multidimensional information recommendation according to different types of context could also be useful in addressing this open issue.

7.5. Real-Time Browsing

Most existing systems focus on analyzing and viewing meetings after they are finished. A real-time display that allows the participants to take a bird's-eye view of the current meeting is sometimes needed. First, it helps in organizing the meeting: for instance, by knowing the current status of a meeting (e.g., did all members agree on the outcome, who was quiet, who was extroverted etc.), the organizer can perform adjustments to make the meeting more efficient [Yu et al. 2007]. Second, it helps a person who misses a meeting's beginning and joins it in progress to know what has already happened and what topics were discussed. It further improves people's skill in participating in a meeting in a balanced manner. Balanced participation is essential in properly solving problems. Through real-time browsing, the members become aware of their own and others' behavior in a discussion (e.g., one person speaks for a long time; two people always hold their discussions in a subgroup), and then make adjustments to increase the satisfaction of the group with the discussion process [Kulyk et al. 2005]. The individual, the group, and the whole organization could benefit from participants' awareness of their own behavior during meetings [Pianesi et al. 2008]. For real-time browsing, automatic and rapid recognition, annotation, and summarizing are required.

7.6. From Understanding to Activating

While it is important to recognize information in the meeting's content automatically, using this information to make the meeting fruitful is even more valuable. Previous smart meeting systems tended to record the meeting and then understand the content. Such systems are content-centric but not human-centric. To be human-centric, a smart meeting system should recognize information that activates discussion and facilitates human-human interaction. Before a meeting, the system could search and reorganize knowledge about the participants and build relationships between them based on the main topics that to be discussed. Then, it could recommend interesting topics or useful references appropriate to different situations or social contexts at the meeting. It can also detect potential interactions (e.g., a participate may have some idea about the current topic, and two people could discuss it due to their common background or interest. For this purpose, the system needs to first analyze human semantic interactions (e.g., proposing an idea, giving comments, expressing a positive opinion, requesting information) in order to better understand communicative status [Yu et al. 2008a]. To facilitate decision-making at meetings, Yu et al. [2008c] proposed the Meeting Warming-up system that detects participants' common interests and conflicts before a meeting. With the help of the system, participants can warm-up for discussions around potential outcomes. For instance, the meeting could have an easy and friendly start and commonly liked outcomes; commonly disliked outcomes could be ignored to save time, and

participants could be mentally prepared to discuss the conflicting outcomes fully and carefully.

7.7. Distributed Meetings

Most existing studies on smart meetings were conducted within a single room. Advances in communication and information technologies, specifically teleconferencing, gave rise to distributed meetings [Nijholt et al. 2005]. The new technologies could support collaborative teamwork through multiparty meetings [Cutler et al. 2002; Luo et al. 2004; Han et al. 2007]. A number of challenges exist in building a distributed smart meeting. First, seamless interaction between local and remote participants must be supported (e.g., telepresence, addressing, and floor control). Techniques such as smart cameraman (which automatically selects the most appropriate view according to the activities in the meeting room) and intelligent identity recognition are helpful. Second, the meeting browser should support information navigation, search, and query across multiple meeting rooms. Third, for efficient data sharing between remote meetings, data transfer should adapt to changing situations. For instance, if the result analyzed is not good and the network bandwidth is high, raw data (video and audio) should be delivered. On the other hand, if the network speed is very low but the analytical ability is powerful, the system can offer high-level semantics, such as human interactions and user status (interested or bored). Mechanisms like broadcasting [Defago et al. 2004; Luo et al. 2004] and quality of service (QoS) [Uludag et al. 2007; Wang and Crowcroft 1996] are also required.

7.8. Security and Privacy

Information shared during meetings is highly sensitive [Jaimes and Miyazaki 2005]. Interaction and information exchange between entities must be secure and private. Security includes three main properties: confidentiality, integrity, and availability [Stajano 2002]. For instance, some internal meetings in a company involve business secrets. It is necessary to protect the content from unauthorized access or restrict access to portions of some data. Privacy is the personal claim of attendees for when, how, and to what extent information is recorded. In fact, users are often anxious about themselves when they are in an environment with many sensors, since they do not know what will be captured and stored and how it will be used. Privacy mechanisms in ubiquitous computing such as in Moncrieff et al. [2007] and Srinivasan et al. [2008] could be explored and used in smart meeting systems.

8. CONCLUSION

Smart meeting systems, which constitute a problem-rich research area, have attracted much interest in both industry and academia over the last decade. In this article, we first reviewed the existing work in this field and gave an overview of underlying technologies so that researchers in the smart meeting domain can understand the key design issues of such a system. Then, several open issues (e.g., improvement of current technologies, infrastructure, active capture, context-awareness, real-time feedback, meeting activation, distributed meetings, and security and privacy) were discussed as possible ways to extend the capabilities of current smart meeting systems. We hope that the issues presented here will advance the discussion in the community towards future smart meeting systems.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

AIZAWA, K., ISHIJIMA, K., AND SHIINA, M. 2001. Summarizing wearable video. In *Proceedings of the IEEE ICIP*. IEEE, Los Alamitos, CA, 398–401.

AKITA, Y., TRONCOSO, C., AND KAWAHARA, T. 2006. Automatic transcription of meetings using topic-oriented language model adaptation. In *Proceedings of the Western Pacific Acoustics Conference (WESPAC)*.

AMI. 2007. AMI project homepage, http://www.amiproject.org/.

BARON, D., SHRIBERG, E., AND STOLCKE, A. 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *Proceedings of the International Conference on Spoken Language Processing*. 949–952.

BELL, G. AND GEMMELL, J. 2007. A digital life. *Scientific American*.

BETT, M., GROSS, R., YU, H., ZHU, X., PAN, Y., YANG, J., AND WAIBEL, A. 2000. Multimodal meeting tracker. In *Proceedings of the International Conference on Content-Based Multimedia Information Access (RIAO)*. 32–45.

BOUNIF, H., DRUTSKYY, O., JOUANOT, F., AND SPACCAPIETRA, S. 2004. A multimodal database framework for multimedia meeting annotations. In *Proceedings of the International Conference on Multi-Media Modeling (MMM)*. 17–25.

BUDNIKOV, D., CHIKALOV, I., KOZINTSEV, I., AND LIENHART, R. 2004. Distributed array of synchronized sensors and actuators. In *Proceedings of the EUSIPCO*.

BUSSO, C., HERNANZ, S., CHU, C. W., KWON, S., LEE, S., GEORGIOU, P. G., COHEN, I., AND NARAYANAN, S. 2005. Smart room: Participant and speaker localization and identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 2. IEEE, Los Alamitos, CA, 1117–1120.

CHEN, X. AND YANG, J. 2002. Towards monitoring human activities using an omnidirectional camera. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*. IEEE, Los Alamitos, CA, 423–428.

CHIU, P., KAPUSKAR, A., REITMEIER, S., AND WILCOX, L. 2000. Room with a rear view: Meeting capture in a multimedia conference room. *IEEE Multimedia 7*, 4, 48–54.

COLBATH, S. AND KUBALA, F. 1998. Rough'n'ready: A meeting recorder and browser. In *Proceedings of the Perceptual User Interface Conference*. 220–223.

CUTLER, R., RUI, Y., GUPTA, A., CADIZ, J. J., TASHEV, I., HE, L., COLBURN, A., ZHANG, Z., LIU, Z., AND SILVERBERG, S. 2002. Distributed meetings: A meeting capture and broadcasting system. In *Proceedings of the 10th ACM Conference on Multimedia*. ACM, New York, 503–512.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv. 40*, 2, Article 5.

DEFAGO, X., SCHIPER, A., AND URBON, P. 2004. Total order broadcast and multicast algorithms: Taxonomy and Survey. *ACM Comput. Surv. 36*, 4, 372–421.

DEY, A. K., SALBER, D., ABOWD, G. D., AND FUTAKAWA, M. 1999. The conference assistant: Combining context-awareness with wearable computing. In *Proceedings of the 3rd International Symposium on Wearable Computers (ISWC)*. 21–28.

DIELMANN, A. AND RENALS, S. 2004. Dynamic Bayesian networks for meeting structuring. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Los Alamitos, CA, 629–632.

DIMICCO, J. M., HOLLENBACH, K. J., AND BENDER, W. 2006. Using visualizations to review a group's interaction dynamics. Extended abstracts. In *Proceedings of the CHI*. 706–711.

FOOTE, J. AND KIMBER, D. 2000. FlyCam: Practical panoramic video and automatic camera control. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Los Alamitos, CA, 1419–1422.

GATICA-PEREZ, D., MCCOWAN, I., ZHANG, D., AND BENGIO, S. 2005. Detecting group interest-level in meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 1. IEEE, Los Alamitos, CA, 489–492.

GATICA-PEREZ, D., MCCOWAN, I., BARNARD, M., BENGIO, S., AND BOURLARD, H. 2003. On automatic annotation of meeting databases. In *Proceedings of the International Conference on Image Processing (ICIP)*. Vol. 3. 629–632.

GEYER, W., RICHTER, H., AND ABOWD, G. D. 2005. Towards a smarter meeting record – Capture and access of meetings revisited. *Multimedia Tools Appl. 27*, 3, 393–410.

GEYER, W., RICHTER, H., AND ABOWD, G.D. 2003. Making multimedia meeting records more meaningful. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. Vol. 2. IEEE, Los Alamitos, CA, 669–672.

GROSS, R., YANG, J., AND WAIBEL, A. 2000. Face recognition in a meeting room. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, Los Alamitos, CA, 294–299.

HAN, S., KIM, N., CHOI, K., AND KIM, J. W. 2007. Design of multi-party meeting system for interactive collaboration. In *Proceedings of COMSWARE*. 1–8.

HILLARD, D., OSTENDORF, M., AND SHRIBERG, E. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference (HLT-NAACL)*. 34–36.

HORSTER, E. AND LIENHART, R. 2006. Approximating optimal visual sensor placement. In *Proceedings of the IEEE ICME*. IEEE, Los Alamitos, CA.

JAIMES, A. AND MIYAZAKI, J. 2005. Building a smart meeting room: from infrastructure to the video gap (research and open issues). In *Proceedings of the 21st International Conference on Data Engineering Workshops (ICDEW)*. 1173–1182.

JAIMES, A., OMURA, K., NAGAMINE, T., AND HIRATA, K. 2004. Memory cues for meeting video retrieval. In *Proceedings of the First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE)*. ACM, New York, 74–85.

JAIN, R. 2003. Experiential computing. *Comm. ACM. 46*, 7, 48–54.

JAIN, R., KIM, P., AND LI, Z. 2003. Experiential meeting systems. In *Proceedings of the ACM Workshop on Experiential TelePresence*. ACM, New York, 1–12.

JANIN, A., ANG, J., BHAGAT, S., DHILLON, R., EDWARDS, J., MACIAS-GUARASA, J., MORGAN, N., PESKIN, B., SHRIBERG, E., STOLCKE, A., WOOTERS, C., AND WREDE, B. 2004. The ICSI meeting project: Resources and research. In *Proceedings of the NIST ICASSP Meeting Recognition Workshop*.

KAPLAN, J. 2005. Next-generation conference rooms. In *Proceedings of the UBICOMP Workshop on Ubiquitous Computing in Next Generation Conference Rooms*.

KENNEDY, L. AND ELLIS, D. 2004. Laughter detection in meetings. In *Proceedings of the NIST ICASSP Meeting Recognition Workshop*. 118–121.

KENNEDY, L. AND ELLIS, D. 2003. Pitch-based emphasis detection for characterization of meeting recordings. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, Los Alamitos, CA, 243–248.

KERN, N., SCHIELE, B., JUNKER, H., LUKOWICZ, P., AND TROSTER, G. 2003. Wearable sensing to annotate meeting recordings. *Person. Ubiquit. Comput. 7*, 5, 263–274.

KULYK, O., WANG, C., AND TERKEN, J. 2005. Real-time feedback based on nonverbal behaviour to enhance social dynamics in small group meetings. In *Proceedings of the Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*. 150–161.

LEE, D., EROL, B., GRAHAM, J., HULL, J. J., AND MURATA, N. 2002. Portable meeting recorder. In *Proceedings of the 10th ACM Conference on Multimedia*. ACM, New York, 493–502.

LIU, Z., ZHANG, Z., HE, L.W., AND CHOU, P. 2007. Energy-based sound source localization and gain normalization for ad hoc microphone arrays. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE. Los Alamitos, CA.

LIWICKI, M., SCHLAPBACH, A., BUNKE, H., BENGIO, S., MARIETHOZ, J., AND RICHIARDI, J. 2006. Writer identification for smart meeting room systems. In *Proceedings of the 7th IAPR Workshop on Document Analysis Systems*. 186–195.

LUO, C., LI, J., AND LI, S. 2004. DigiMetro - An application-level multicast system for multiparty video conferencing. In *Proceedings of Globecom*. 982–987.

MBOSS. 2007. Mboss homepage. http://www.mboss.force9.co.uk/.

MCCOWAN, I., GATICA-PEREZ, D., BENGIO, S., LATHOUD, G., BARNARD, M., AND ZHANG, D. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 3, 305–317.

MIKIC, I., HUANG, K., AND TRIVEDI, M. 2000. Activity monitoring and summarization for an intelligent meeting room. In *Proceedings of the IEEE Workshop on Human Motion*. IEEE, Los Alamitos, CA, 107–112.

MIMIO. 2007. Mimio homepage. http://www.mimio.com/.

MONCRIEFF, S., VENKATESH, S., AND WEST, G. 2007. Privacy and the access of information in a smart house environment. In *Proceedings of the 15th Multimedia Conference*. ACM, New York, 671–680.

MOORE, D. AND MCCOWAN, I. 2003. Microphone array speech recognition: Experiments on overlapping speech in meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 5, 497–500.

MURRAY, G., RENALS, S., AND CARLETTA, J. 2005. Extractive summarization of meeting recordings. In *Proceedings of the Ninth European Conference on Speech*. 593–596.

NAIT-CHARIF, H. AND MCKENNA, S. J. 2003. Head tracking and action recognition in a smart meeting room. In *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. IEEE, Los Alamitos, CA, 24–31.

NIJHOLT, A., ZWIERS, J., AND PECIVA, J. 2005. The distributed virtual meeting room exercise. In *Proceedings of the ICMI Workshop on Multimodal Multiparty Meeting Processing*. 93–99.

OZEKI, M., MAEDA, S., OBATA, K., AND NAKAMURA, Y. 2008. Virtual assistant: An artificial agent for enhancing content acquisition. In *Proceedings of the 1st ACM International Workshop on Semantic Ambient Media Experience*. ACM, New York.

PIANESI, F., ZANCANARO, M., NOT, E., LEONARDI, C., FALCON, V., AND LEPRI, B. 2008. Multimodal support to group dynamics. *Person. Ubiquit. Comput. 12*, 2.

PIMENTEL, M. C., CATTELAN, R. G., AND BALDOCHI, L. A. 2007. Prototyping applications to document human experiences. *IEEE Pervasive Comput. 6*, 2, 93–100.

RAM, S., RAMAKRISHNAN, K. R., ATREY, P. K., SINGH, V. K., AND KANKANHALLI, M. S. 2006. A design methodology for selection and placement of sensors in multimedia surveillance systems. In *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks (VSSN)*. ACM, New York, 121–130.

RAYKAR, V. C., KOZINTSEV, I. V., AND LIENHART, R. 2005. Position calibration of microphones and loudspeakers in distributed computing platforms. *IEEE Trans. Speech Audio Process. 13*, 1, 70–83.

REIDSMA, D., RIENKS, R., AND JOVANOVIC, N. 2004. Meeting modelling in the context of multimodal research. In *Proceedings of the 1st International Workshop on Machine Learning for Multimodal Interaction*. 22–35.

REITER, S. AND RIGOLL, G. 2005. Meeting event recognition using a parallel recurrent neural net approach. In *Proceedings of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*.

RENALS, S. AND ELLIS, D. 2003. Audio information access from meeting rooms. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 4. IEEE, Los Alamitos, CA, 744–747.

RICHTER, H., ABOWD, G. D., GEYER, W., FUCHS, L., DAIJAVAD, S., AND POLTROCK, S. 2001. Integrating meeting capture within a collaborative team environment. In *Proceedings of the International Conference on Ubiquitous Computing (UbiComp'01)*. Lecture Notes in Computer Science, Vol. 2201, Springer, Berlin, 123–138.

RIJSBERGEN, C. J. 1979. *Information Retrieval* 2nd Ed., Butterworths.

RUI, Y. 2002. System and process for locating a speaker using 360 degree sound source localization, U.S. Patent.

RUI, Y., GUPTA, A., AND CADIZ, J. J. 2001. Viewing meetings captured by an omnidirectional camera. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 450–457.

RUI, Y., GUPTA, A., GRUDIIN, J., AND HE, L. 2004. Automating lecture capture and broadcast: technology and videography. *Multimedia Syst. 10*, 1, 3–15.

SCHULTZ, T., WAIBEL, A., BETT, M., METZE, F., PAN, Y., RIES, K., SCHAAF, T., SOLTAU, H., WESTPHAL, M., YU, H., AND ZECHNER, K. 2001. The ISL meeting room system. In *Proceedings of the Workshop on Hands-Free Speech Communication (HSC)*.

SINGH, V. K., ATREY, P. K., AND KANKANHALLI, M. S. 2007. Coopetitive multi-camera surveillance using model predictive control. *Machine Vision Appl.* DOI: 10.1007/s00138-007-0082-2.

SRINIVASAN, V., STANKOVIC, J., AND WHITEHOUSE, K. 2008. Protecting your daily in-home activity information from a wireless snooping attack. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp)*. 202–211.

STAJANO, F. 2002. *Security for Ubiquitous Computing*. Wiley, New York.

STANFORD, V., GAROFOLO, J., GALIBERT, O., MICHEL, M., AND LAPRUN, C. 2003. The NIST smart space and meeting room projects: Signals, acquisition, annotation and metrics. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 4. IEEE, Los Alamitos, CA, 6–10.

STIEFELHAGEN, R. 2002. Tracking focus of attention in meetings. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI)*, IEEE, Los Alamitos, CA, 273–280.

STIEFELHAGEN, R., YANG, J., AND WAIBEL, A. 2002. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. Neural Netw. 13*, 4, 928–938.

STIEFELHAGEN, R. AND ZHU, J. 2002. Head orientation and gaze direction in meetings. In *Proceedings of the Conference on Extended Abstracts on Human Factors in Computing Systems* (*Extended Abstracts*) *(CHI'02)*. ACM, New York, 858–859.

STIEFELHAGEN, R., YANG, J., AND WAIBEL, A. 2001. Estimating focus of attention based on gaze and sound. In *Proceedings of the Workshop on Perceptive User Interfaces (PUI)*. 1–9.

STIEFELHAGEN, R., YANG, J., AND WAIBEL, A. 1999. Modeling focus of attention for meeting indexing. In *Proceedings of the 7th ACM International Conference on Multimedia (Part 1)*. ACM, New York, 3–10.

TOMOBE, H. AND NAGAO, K. 2006. Discussion ontology: Knowledge discovery from human activities in meetings. In *New Frontiers in Artificial Intelligence*, Lecture Notes in Computer Science. Vol. 4384, Springer, Berlin, 33–41.

TRIVEDI, M., MIKIC, I., AND BHONSLE, S. 2000. Active camera networks and semantic event databases for intelligent environments. In *Proceedings of the IEEE Workshop on Human Modeling, Analysis and Synthesis (in conjunction with CVPR)*. IEEE, Los Alamitos, CA.

TRUONG, K. N. AND ABOWD, G. D. 2004. INCA: A software infrastructure to facilitate the construction and evolution of ubiquitous capture and access applications. In *Pervasive Computing*. Lecture Notes in Computer Science. Vol. 3001, Springer, Berlin, 140–157.

TUCKER, S. AND WHITTAKER, S. 2004. Accessing multimodal meeting data: Systems, problems and possibilities. In *Proceedings of the Workshop on Multimodal Interaction and Related Machine Learning Algorithms*. 1–11.

TURK, M. AND PENTLAND, A. 1991. Eigenfaces for recognition. *J. Cognitive Neurosci. 3*, 1, 71–86.

ULUDAG, S., LUI, K. S., NAHRSTEDT, K. AND BREWSTER, G. 2007. Analysis of topology aggregation techniques for QoS routing. *ACM Comput. Surv. 39*, 3, Article 7.

VALIN, J. M., MICHAUD, F., ROUAT, J., AND LETOURNEAU, D. 2003. Robust sound source localization using a microphone array on a mobile robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Los Alamitos, CA, 1228–1233.

WAIBEL, A., BETT, M., METZE, F., RIES, K., SCHAAF, T., SCHULTZ, T., SOLTAU, H., YU, H., AND ZECHNER, K. 2001. Advances in automatic meeting record creation and access. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Los Alamitos, CA, 597–600.

WAIBEL, A., BETT, M., AND FINKE, M. 1998. Meeting browser: Tracking and summarizing meetings. In *Proceedings of the Broadcast News Transcription and Understanding Workshop*. 281–286.

WANG, Z. AND CROWCROFT, J. 1996. Quality-of-service routing for supporting multimedia applications. *IEEE J. Select. Areas Comm. 14*, 7, 1228–1234.

WELLNER, P., FLYNN, M., AND GUILLEMOT, M. 2004. Browsing recorded meetings with Ferret. In *Proceedings of the First International Workshop on Machine Learning for Multimodal Interaction (MLMI)*. 12–21.

WREDE, B. AND SHRIBERG, E. 2003a. Spotting 'hot spots' in meetings: Human judgments and prosodic cues. In *Proceedings of the European Conference on Speech Communication and Technology*. 2805–2808.

WREDE, B. AND SHRIBERG, E. 2003b. The relationship between dialogue acts and hot spots in meetings. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, Los Alamitos, CA, 180–185.

YANG, J., ZHU, X., GROSS, R., KOMINEK, J., PAN, Y., AND WAIBEL, A. 1999. Multimodal people ID for a multimedia meeting browser. In *Proceedings of ACM Multimedia*. ACM, New York, 159–168.

YU, H., FINKE, M., AND WAIBEL, A. 1999. Progress in automatic meeting transcription. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech)*. Vol. 2, 695–698.

YU, Z., ZHOU, X., ZHANG, D., CHIN, C. Y., WANG, X., AND MEN, J. 2006. Supporting context-aware media recommendations for smart phones. *IEEE Pervasive Comput. 5*, 3, 68–75.

YU, Z., OZEKI, M., FUJII, Y., AND NAKAMURA, Y. 2007. Towards smart meeting: Enabling technologies and a real-world application. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI)*. 86–93.

YU, Z., AOYAMA, H., OZEKI, M., AND NAKAMURA, Y. 2008a. Collaborative capturing and detection of human interactions in meetings. In *Adjunct Proceedings of the 6th International Conference on Pervasive Computing (Pervasive)*. 65–69.

YU, Z., AOYAMA, H., OZEKI, M., AND NAKAMURA, Y. 2008b. Social interaction detection and browsing in meetings. In *Adjunct Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp)*. 40–41.

YU, Z., ZHOU, X., AND NAKAMURA, Y. 2008c. Meeting warming-up: Detecting common interests and conflicts before a meeting. In *Proceedings of the ACM Computer Supported Cooperative Work Conference Supplement (CSCW)*.

ZHANG, D., GATICA-PEREZ, D., BENGIO, S., MCCOWAN, I., AND LATHOUD, G. 2004. Modeling individual and group actions in meetings: A two-layer HMM framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Event Mining in Video (CVPR-EVENT)*. Vol. 7. IEEE, Los Alamitos, CA, 117–124.

ZHAO, J. AND CHEUNG, S.-C.S. 2007. Multi-camera surveillance with visual tagging and generic camera placement. In *Proceedings of the First ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*. 259–266.

ZOBL, M., WALLHOFF, F., AND RIGOLL, G. 2003. Action recognition in meeting scenarios using global motion features. In *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-CCVS)*. 32–36.

Received February 2008; revised August 2008; accepted October 2008