1st IEEE International Workshop on Managing Data for Emerging Multimedia Applications (EMMA) in conjunction with 21th IEEE Conference on Data Engineering (ICDE 2005), April 9, 2005, Tokyo, Japan)

BUILDING A SMART MEETING ROOM: FROM INFRASTRUCTURE TO THE VIDEO GAP (RESEARCH AND OPEN ISSUES)

Alejandro Jaimes and Jun Miyazaki
{alex.jaimes, jun.miyazaki}@fujixerox.co.jp
FXPAL Japan
Fuji Xerox Co., Ltd.

ABSTRACT

At FXPAL Japan we have built an (experimental) Smart Conference Room (SCR) that contains multiple cameras, microphones, displays, and capture devices. Based on our experience, in this paper we discuss research and open issues in constructing SCRs like the one built at FXPAL for the purpose of automatic content analysis. Our discussion is grounded on a novel conceptual meeting model that consists of *physical* (from layout to cameras), *conceptual* (meeting types, actors), *sensory* (audio-visual capture), and *content* (syntax and semantics) components. We also discuss storage, retrieval, and deployment issues.

1. INTRODUCTION

Meetings are important events in any organization and recently there has been a renewed interest in building smart meeting rooms to capture meetings on video for future viewing. This is due to lower computer and video equipment costs, higher computational power, and because keeping accurate records in companies has become more important than ever (for knowledge, risk management, and compliance, among others). In the United States, for example, the SOX act [21] and recent laws require accurate record keeping to ensure the financial data the CEO and CFO sign off on is auditable. Although recording of meetings is not a requirement, it is possible for meeting videos to play an important role in the future: traditional note-taking is insufficient to store all relevant meeting events, it is subjective, often incomplete, and inaccurate.

Many smart meeting conference room environments [39][61][43] and portable meeting systems [38] have been developed. Most of the focus has been on developing techniques to automatically process the generated audio-visual content (e.g., face detection and action recognition [67]; speech recognition for topic detection [62], and many others [3]). However, little attention has been given to the overall meeting capture framework, the issues around building the *infrastructure* necessary to deploy a real world application, and the impact of such infrastructure on the development of automatic content analysis techniques.

In this paper, we propose a multiple-component conceptual meeting model, and give an overview of the major research issues in building and deploying a smart conference room environment from the perspective of automatic content analysis. We discuss issues ranging from physical room layout and hardware infrastructure to automatic content analysis and metadata.

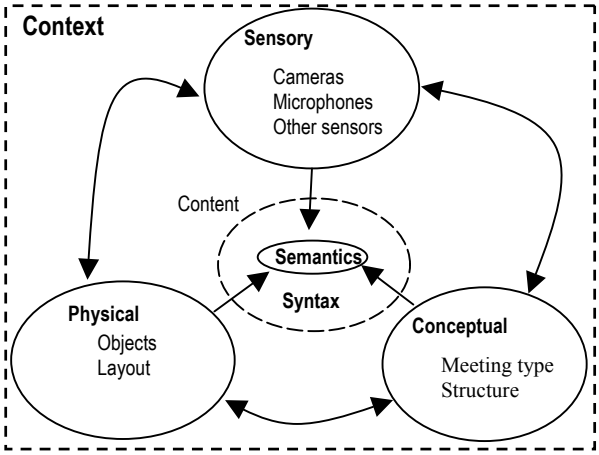

Figure 1. Our meeting model.

Our model (Figure 1) consists of four components: *physical structure, conceptual structure, sensory acquisition*, and *acquired content*[1]. The physical component models the objects and layout of a smart meeting room (e.g., tables). The conceptual component models the structure of the meeting (e.g., meeting type, roles). The sensory component models the capture of the meeting using multiple sensing devices (cameras, microphones, etc.). The four components of our model are directly linked by a *contextual mesh*, which we define as the set of conditions under which the meeting takes place. As the circle in the center indicates, the meeting's acquired content (visual structure of videos, topics discussed, meaning of meeting segments, information from sensors, metadata, etc.) is directly influenced by all of the components—the computational approach depends on all of them.

[1] We will use the term content-based analysis to refer to audio-visual content. Note however, that this component may include information obtained from other types of sensors.

1.1.The FXPAL Japan Smart Conference Room

The Smart Conference Room we have built at FXPAL Japan is depicted in Figure 2. The walls of the meeting room were built so that they contain shelves useful for equipment storage (Figure (a)). Since this is an experimental meeting room, the cameras (b) have been set up so they can be placed in different configurations. As figure (c) shows, the furniture is also highly configurable so that we can experiment with different room layouts. All of the walls in the conference room can be used as black boards or as projection walls (note projector on the table). Figure (d) also shows videoconference equipment and large displays which show the input from the cameras or presentation materials. We discuss further details throughout the paper.

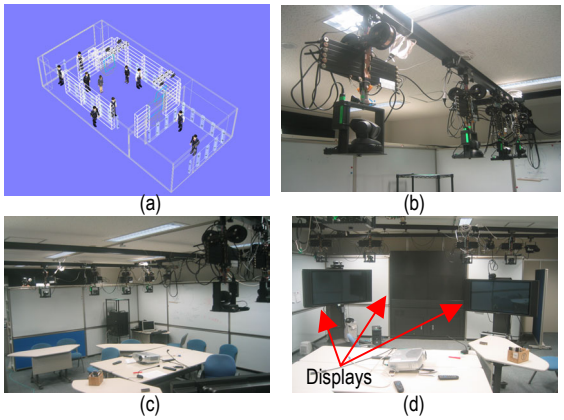

Figure 2. FXPAL Japan Smart Conference Room.

1.2.Related Work

Other related meeting room projects include [16][17][18][19][7] and [20] (several others are mentioned throughout the paper). Our conceptual model is related to the models presented in [51] and [32], which use layered components, but focus mainly on the conceptual or logical aspects (e.g., for annotation). The model in [4] deals with annotations, and the model in [9] on meeting outcomes. The work in [40] focuses on browsing, [41] describes a technical approach to detect human interaction, [51] deals with a general framework for corpus based multi-modal research, and [23] uses ontologies. While all of these offer useful perspectives on the meeting analysis problem, they generally focus on only one component of meeting video analysis. Our discussion extends to the areas indicated by the model in Figure 1.

1.3.Outline

The rest of the paper is organized as follows. In section 2 we discuss the conceptual component. In section 3 we discuss the physical component. In section 4 we discuss the sensory component, while in section 5 we discuss content analysis. Section 6 deals with storage, metadata, and retrieval. We conclude in section 7.

2. CONCEPTUAL COMPONENT

The conceptual component models the type of meeting, the roles of meeting participants, and the actions and events that occur in the meeting. The basic idea behind this component is that the structure of these elements impact audio-visual structure. For example, conceptual structure has been used to develop algorithms for automatic indexing in news analysis [5], sports [65], among others.

2.1. Meeting Structure

The type of meeting determines the conceptual structure (e.g., who speaks and when, for how long, whether there is an order to the meeting, what actions occur, etc.), the number of people that attend, the size of the meeting room, and the layout (section 3).

In structured meetings there is an *explicit* structure not only in terms of who attends the meeting and the roles, but also on who speaks and when. Agendas or documents such as Robert's Rules of Order, therefore, can help guide the indexing process.

In unstructured meetings, on the contrary, there is no *explicit* structure: anyone may speak at anytime, there may not be an agenda and if there is one it may be fairly general. Examples of meeting types are given in Table 1 (a similar classification was done in [32]).

Table 1. Type s of meetings.

Structured	**Unstructured**
Panel	Brainstorming
Talk	Discussion
Presentation	Decision making
Debate	Coordination
Interview, report, hearing, etc.	

For different types of meetings there are, of course, different types of events, which contain sub-events: a presentation may contain questions and answers, and most likely images of slides (see section 5).

2.2.Meeting Actors & Actions

Individuals at the meeting may have specific roles, which constrain their particular actions. The master of ceremony (manager or meeting leader), for instance, structures the discussions and the meeting. Therefore, he will appear frequently in the meeting recordings so the actions performed by the meeting leader may be more important than those of others.

Table 2. Actor's roles.

Role	**Description**
Sponsor	Owner of the meeting.
Facilitator	Plans and manages the meeting.
Participant	Attends, contributes to the meeting.
Reporter	Produces final meeting report.
Organizational agent	People that find their actions directly affected by the outcome of the meeting.

We can identify five major types of roles played by meeting attendees, or actors [9], whether these roles are *explicit*, *implicit*, *temporary*, or *permanent* (Table 2).

A detailed model of the actions associated with each role can be constructed (e.g., [11]). What is most important here is to consider the differences in the structure and roles of the participants and how that affects the acquired content (e.g., audio-visual).

2.3.Open Issues and Research Directions

There has been a significant amount of research in communications and psychology on how people interact. *Computer Supported Cooperative Work (CSCW)*, for example, has been an active area of research for many years. In spite of this, there is very little work on using models from psychology and communications research in analyzing content captured in smart conference rooms (see an interesting discussion in [48] on socially aware computing), or on analyzing the social impact of these new technologies when deployed in a real setting.

Another area that has not been explored is the use of structured meeting documents for indexing [32]. The agenda, for example, can be effectively used to improve the performance of content classification methods. Additional metadata about the meeting participants can also be of importance, giving contextual information (i.e., to help determine whose comments are important). To our knowledge, actor's roles have not been used in automatic meeting content analysis.

Meetings might also be examined in relation to other meetings: periodic meetings on a specific topic, for example can have the same individual structure and be linked semantically.

3. PHYSICAL COMPONENT

The physical component models the objects in the meeting room and their layout. Our interest in indexing meeting videos is mainly on the actions of the participants, but the physical layout of the meeting room, the number of participants, and the meeting structure, are tightly linked. We consider the physical component separately because it can place strong limitations on the audio-visual capture (section 4).

3.1.Table Layout

The table layout (Figure 3) may be *fixed* or *modular*, allowing tables and chairs to be placed in different arrangements. Although variations within a single type of meeting might be small, any variations in sitting positions or table layout can lead to major differences in the acquired audio-visual content. One important SCR design decision, therefore, is what the basic arrangement should be and whether the tables and chairs can be rearranged.

3.2.Objects

We separate objects into private and public objects which can be moved or are fixed (Table 3). Such classification is useful for automatic content analysis: in [27] a framework was developed for detecting actions involving *fixed* objects (e.g., when someone stands by the board).

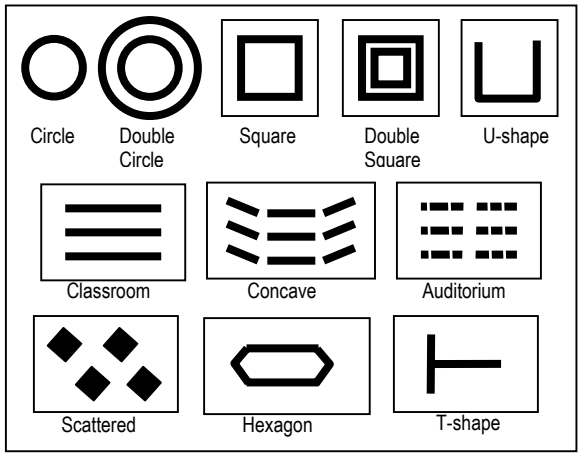

Figure 3. Table layout examples.

Table 3. Types of objects.

	Fixed	**Movable**
Public	Wall, Projector, Teleconference device, Whiteboard, Display	Table, Chair
Private (personal)		Notebook, Laptop, Pen, Documents, Briefcase, Jacket, Umbrella, phone

3.2.Open Issues and Research Directions

The physical infrastructure of the meeting room (tables, layout, number of participants, etc.) has a strong impact on the audio-visual content of the recorded meetings, and therefore on the applicability of automatic content-based algorithms. Although physical layout has been studied widely in architecture, design, and sociology, research in those fields has not been used in automatic content-analysis. Furthermore, in most cases, the assumption is that physical infrastructure construction is a minor part of the project, as most researchers focus on the content of the videos and not on the physical component.

Since our SCR is experimental, we have opted for a modular layout. Tables and chairs can be arranged in different ways and the room can accommodate mid-size groups (around 20 people). Our work has focused, however, on small groups of less than 10 people.

One of the problems we have found with the flexible layout is that since different people use the meeting room, tables and chairs are often rearranged. This includes rearranging of microphones, and has an impact on the audio and visual capture—what the cameras "see" can vary significantly from meeting to meeting. Thus we have found that while the flexible layout is good for experimental purposes, it would make meeting analysis extremely difficult in practice.

The physical layout also has an impact on where presentations are projected, and on the location of personal objects. Even the types of chairs are important. We chose chairs that do not lean back and do not have wheels because they allow mobility but pose some restrictions on posture. This is useful for automatic content analysis.

Another important issue is the room lighting. Techniques developed for video analysis should be robust enough to work in a variety of lighting conditions. Although this is still challenging, building a fixed infrastructure allows us to decide in advance which types of lighting are more suitable for video analysis. Of bigger difficulty might be to find the right compromise between lighting that is suitable for the meeting type (a worry of architects and designers) and lighting that is suitable for video analysis. We experimented with special T.V. studio-type lights, but found that although they are excellent for video capture, they are too bright for holding regular meetings. Therefore, we opted for standard fluorescent lighting.

4. SENSORY COMPONENT

In the sensory component we model the sensors used to capture the meeting and the sensor's parameters (e.g., cameras, microphones, motion sensors, etc.).

4.1.Fixed vs. Portable

There are basically two types of infrastructure for meeting capture. In the *fixed* infrastructure, the bulk of the capture equipment is permanently installed in the meeting room. Most of the early efforts in meeting capture were around building smart conference rooms equipped with cameras, microphones, and other sensors. More recently, there have been efforts in constructing *portable* meeting recorders that one can easily move from one meeting room to another.

Portable meeting capture systems can also be divided into two categories, one is *central* capture and the other one is *individual* capture. Central capture systems typically attempt to capture the meeting from an objective perspective. For example, systems have been developed to be placed on top of a meeting table so images and audio of all participants are captured. Individual systems, on the other hand, allow individuals to capture the meeting according to their own needs, from their own perspective. The Quindi meeting capture system [58], for example, allows users to record the meeting using their own camera and laptop.

Fixed meeting capture systems have the advantage that if properly deployed, they can be used by any group meeting in the SCR. No one has to be in charge of carrying the portable capture system to the meeting, and the recordings can be stored centrally for multiple access, possibly eliminating some of problems with portable capture systems (who is in charge of the data if the goal is not personal use, etc.). We focus on fixed meeting capture systems, although some of the discussions apply to both.

4.2.Video Capture

Factors in video capture include the following:

- *Number of cameras:* how many views are sufficient?
- *Camera parameters:* if cameras can pan, tilt, zoom, the types of lenses (how wide), light sensitivity (aperture), the types of cameras (e.g., infrared), etc..
- *Camera locations:* should the cameras be fixed or not?
- *2D vs. Stereo:* is it necessary to obtain stereo information (e.g., to determine exactly where someone is pointing)?

The goals of the capture system will determine many of the factors above. For instance, it may be useful to have a close-up view of each participant's face to determine what their emotions are during the meeting (if emotion algorithms are to be used).

4.3.Audio

Audio plays a major role in meeting video capture and automatic analysis. On one hand, it is desirable to have high quality recordings in which all participants' utterances are heard clearly. On the other hand, the quality of the audio has a very strong impact on automatic analysis.

In most projects, microphones are either placed on the ceiling [7], placed on the tables [16], or they are worn by meeting attendees [33]. In general, wearable lapel microphones are the most accurate (see documents related to [33]), but most intrusive. In the non-wearable case obvious choices include the number of microphones, types, and locations.

The type of setup chosen depends on the goal of the project and how the meeting contents will be used (e.g., speaker identification, segmentation of meetings into speech and non-speech segments, affective analysis, speech recognition, etc.). For example, the quality of the audio obtained from a single microphone may be sufficient for human users, but insufficient for speech recognition. It is important to note here that even manual transcription of meeting contents is difficult in most cases (see [2] and discussions in section 5 on different levels of analysis).

4.5. Open Issues and Research Directions

The sensory component and the physical infrastructure are tightly linked, since decisions on room layout as well as sensor placement have a big impact on the audio-visual content. Some of the open issues include the following:

- What is the camera setup that yields scenes optimal for automatic content analysis? (e.g., determining a minimum face area in number of pixels, for face detection algorithms)
- Is it desirable to fully specify an SCR setup so that it may be replicated to function well with automatic

content analysis algorithms? Or should the algorithms adapt to different room setups?

In our SCR we have opted for a primary configuration of 8 pan tilt cameras with zoom lenses. The cameras are placed on rails and dollies to allow us to place them in any configuration (Figure 2b). We also have several additional cameras that are used within the scope of the project.

The cameras in the basic configuration have proven to be sufficient for accurately capturing all of the actions and events in group meetings of around 5 people. We have performed experiments using a face detector to determine reasonable camera locations.

The number of sensors (streams in the case of video) is also an important issue. As the cost of cameras decreases, the problem is not so much the cameras, but the cost (computational and otherwise) of recording multiple streams simultaneously (see section 6).

Although we are not aware of any SCR projects that use other types of sensors (e.g., motion, RFID, etc.), these should also be considered in this component of the model.

5. CONTENT ANALYSIS

Our main focus is on the audio-visual content of meeting videos. We base our discussion on the pyramid structure described in [28], which classifies visual content features into ten levels: four *syntactic* levels and six *semantic* levels. Features at the syntactic levels describe only visual appearance, while features at the semantic levels are related to meaning. We use this structure to frame the discussion because it has been shown that any visual content descriptor can be classified into one of the ten levels (see other alternatives in [25]). It gives us an overall perspective of the different levels of indexing that are possible in meeting video analysis and therefore contributes to the design of indexing algorithms.

In the descriptions below we use the terms feature and descriptor interchangeably. For example, the color of an object is a descriptor, as is the name of a person in the video, or a description of what someone is doing (e.g., a raise hand action).

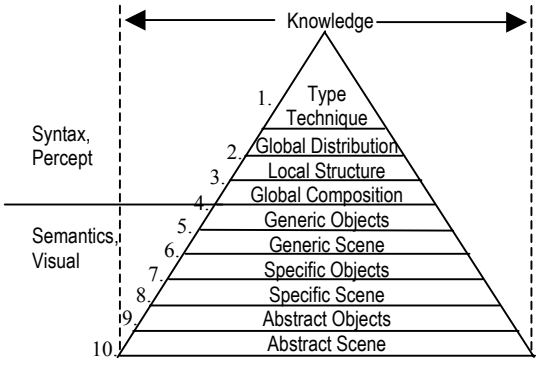

Figure 4. Multi-level indexing pyramid from [26].

5.1.Syntax

Type (color, b/w, etc.): type is characterized by the capture device and content: different types of content are produced by infrared cameras, standard cameras, and omnidirectional cameras. Infrared cameras in a meeting room may be used, for instance, to detect hand motions such as done in [37] and in similar projects.

In terms of audio, the type attribute is also determined by the kind of input device: (e.g., mono, or stereo).

Since meeting video capture is highly controlled and a small number of sources is used, type classification is not a major issue and would probably not be necessary except in special cases for pre-processing. For example, in a video stream that captures only the presentation slides, it might be useful to classify each slide into "photograph," "text," "graphic," types or combinations of these categories.

Global Distribution (global features)**:** in video (visual domain), it is necessary to make a distinction between global *intra-frame* (space) features and global *inter-frame* (time) features. The basic characteristic of features at this level is that they are taken globally (over an entire frame or over a range of frames).

Global features can been used to structure multiple stream meeting videos for visualization or to automatically find highlights. In [29], for example, global motion and lighting changes are used with a combination of face and skin detection algorithms to compute an activity measure. Global activity (over multiple frames) is then used to structure the video for browsing. In [67] the authors use global features to recognize actions in meeting videos.

In audio it makes little sense to compute a feature at a specific instant *t*, thus the distinction of local vs. global will depend on the application: a global feature may be computed over the entire meeting, while a local one only for a small segment. Features are usually computed locally at some granularity—for a segment of length *t.*

In meeting videos the general color and texture distributions are similar within the same meeting. Although some researchers have used global features for meeting video analysis, in general, the use of global features seems to be limited: our interest in meetings is on specific actions or events, or on global ones (group actions). The same argument applies to audio: features computed over an entire video are probably of little use. One exception might be the use of features for data mining (i.e., studying global feature variations across videos within a large video collection).

Local Structure (local features)**:** local structure features are usually extracted to obtain other higher-level features. In the visual domain, this includes the extraction of lines, motion blobs, and other low-level syntactic features (e.g., the author of [42] uses local features for gesture labeling).

In audio, local structure features can be used to improve the extraction of semantic features. The authors of [49], for example, detect speech activity. The output of the

detector is used by a speech recognizer to improve recognition rate.

Since local features are often used to compute higher level semantic features, this remains an important area of research, particularly in the meeting domain. Accurately finding the location of items in 3D space, for instance, could have a strong impact on the extraction of important semantic information: which part of the slide was person A pointing at during the discussion? Where exactly was the person looking?

Global Composition (arrangement of syntactic elements)**:** in the visual domain, global composition has been shown to be useful for detecting actions when fixed cameras are used. The fact that visual scenes of different meetings in the same room are structurally very similar can be an advantage. In [26] composition is used to detect actions and events involving fixed objects. In other domains scene composition can be used to significantly distinguish types of content (e.g., in news video, anchor shot vs. outdoor shot), or for automatic editing [34][46][47][53].

In the audio domain techniques have been developed to detect turn taking, speech vs. non-speech, etc.. A global (composition) representation of this kind of information in the audio signal may give a general idea of the content of the video (where are the most active speech segments, or the silence gaps?).

5.2.Semantics

Generic Object (everyday objects)**:** generic semantic features in the meeting domain include persons [8], and faces. It also includes generic actions or events such as individual and group actions (e.g., standing up, raising hand, entering room).

In audio, algorithms at the generic object level classify the signal into categories such as speech, laughter [35], silence, and male vs. female.

Generic Scene (type of scene): at this level the scene is classified into types (e.g., head and shoulder close-up, table view, etc.). Attributes here can be very useful for automatic video editing [47][53].

Specific Object (individually named objects): features at this level identify specific people (face recognition), and specific actions performed by them. Other features may include identifying specific documents or objects in the meeting room, and in terms of audio identifying speakers.

Specific Scene (individually named scenes): in portable meeting recording systems, it is possible to use algorithms to automatically recognize particular scenes in specific rooms (e.g., if I use the portable recorder in room X it can recognize that it is room X). We're not aware of any work to do this, however, and specific scene recognition in fixed camera meeting rooms is not necessary.

Abstract Object (interpretation of objects)**:** features at this level include detecting and classifying particular actions for affective content. This includes emotion recognition from facial expressions, or from the audio signal.

Although this is a very interesting area, we're not aware of any work in the meeting domain to automatically detect affect in individual meeting participants. Instead, work has focused on obtaining segments of interest at the scene level. One of the difficulties is that most approaches to recognize emotions require close-up, frontal views of the face. This is often not possible with current camera systems in an unobtrusive meeting application.

Abstract Scene (interpretation of scenes)**:** like in abstract object, the goal here is to find interesting segments. However, this is done at the scene level. In [13], for example, the authors try to find "scenes of interest" using a combination of simple visual (e.g., skin blobs, global person motion, etc.) and audio (energy, pitch and speaking rate) features. The authors of [30], on the other hand, study the affective content of meetings using low-level features at the scene level (to determine arousal and valence).

In [36], the pitch of the audio signal is used to find interesting segments. If the audio signal can be attributed to a particular object (e.g., using audio source separation), it is said to be an abstract object measure.

5.3.Actions and Events

We divide meeting events into two categories, individual actions, and meeting events. Meeting events (e.g., presentation, discussion, etc.) affect the conceptual structure of the meeting as a whole, while individual meeting actions may not (e.g. a person taking notes).

Actions may be directed or undirected. The action "stand" for example involves only the person standing. The action point, on the other hand can involve an object or another person (pointed at). We also distinguish between visual, audiovisual, and audio actions in Table 4.

Each of the actions in Table 4 can be classified at any of the levels of the pyramid of Figure 4. A nod, for example, may be detected (generic object), and interpreted to have a particular meaning (abstract object), or associated with a particular person (specific object).

Table 4. Individual actions in meetings. Directed actions are marked with (D).

Visual Only	Visual + Audible	Audible Only
Smile, frown, etc.	Laugh	Audio played
Raise hand	Speak (D)	
Stand, sit down	Type	
Write, Point (D)	Applause (D)	
Gaze (D),	Cough	
Scratch, etc.		
Turn light off/on		
Posture, Nod (D)		

As we discuss below, detecting individual actions is of great importance in meeting video analysis because it can lead to automatically detecting important or interesting meeting segments.

5.4.Open Issues and Research Directions

Most of the work in content analysis deals with low and mid-level features, which in turn support the extraction of high-level features. Although there has been significant work in automatic extraction of visual features at the generic object level (e.g., face, hand tracking), there are still many open issues. In face detection one of the problems is the difficulty in detecting non-frontal faces. Person identification is considered the most important feature at the specific object level. As with face detection, and the detection and tracking of human parts (e.g., hand, arms) there are difficulties with lighting, and occlusion.

With respect to audio in the SCR setting, one of the biggest problems is noise, which results in low performance. Problems in speaker identification include the existence of cross-talk and the need for training data. In multi-language scenarios and with non-native speakers speech recognition is also a major challenge (even in manual transcriptions).

6. STORAGE AND RETRIEVAL

6.1.Storage

Although there has been a very large amount of research on content analysis in the last 10 years, there has been little progress on developing effective database systems for multimedia data. Most approaches have focused on single stream videos, rather than on synchronized, multiple stream content.

Some of the main research issues and their implications are described below:

- Random access of multiple streams by multiple users: implications on synchronization and security, among others.
- Integration of multiple data types into cohesive information units: meeting videos are often associated with large amounts of diverse metadata.
- Storage in multiple formats: meeting videos at multiple resolution may be required for different applications.
- Compression: impact on viewing quality and the applicability of automatic analysis algorithms.

We're not aware of any frameworks specifically for supporting multiple-stream meeting contents. However, a database schema for meeting recording annotations was presented in [4]. The authors of [44] discuss a general video database framework in the context of TREC video retrieval.

6.2.Metadata

Large amounts of metadata are generated before, during, and after meetings. From the perspective of video analysis, metadata includes basically everything except the audio-visual content itself.

The following are example sources of metadata (see also Table 5):

- E-mails & documents,
- Sensor data (e.g., participant locations and other information),
- Room name, layout,
- Participants names and profiles,
- Participant roles (e.g., manager, coordinator, etc.).

In standards such as MPEG-7 and other initiatives there has been a strong emphasis on using XML compatible formats to describe metadata. Two of the key issues, therefore, are compatibility for data exchange, and storage of the metadata in *conjunction* with the video data. Many of the documents generated are created in different applications in a variety of formats and it is necessary to decide how they will be integrated.

Table 5. Types of metadata generated before, during, and after a meeting.

Pre-meeting	During Meeting	After meeting
Agenda	Participant notes	Annotations
Presentation slides	Group notes	Links to related material
List of participants	Diagrams	Summary
Documents	Summary	Links to other meetings
Videos	Sensory information	
Organizers	Digital photos	
Meeting room information	Activities of a participant	
Links to other meetings	Position of a participant	

Another important issue concerns textual annotations generated manually or automatically from content analysis (e.g., [4]). Unlike in traditional databases, it is likely that annotations will include spatial location data (within the meeting room and for particular views). This may include selected regions in the videos, highlighting the need to properly represent video regions, scenes, frames, and their associated metadata (e.g., MPEG-4 and MPEG-7).

6.3.Retrieval

In other domains (e.g., sports, news, movies) video content is usually structured around shots or scenes. This approach is not feasible for the continuous streams obtained in SCRs because it is difficult to define scene and shot boundaries.

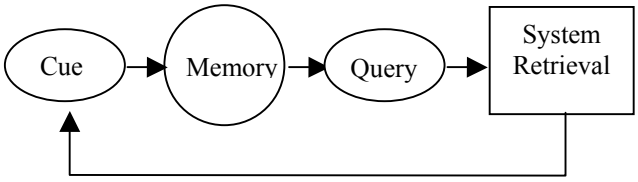

Figure 5. Memory-cue retrieval.

One option is to develop alternative mechanisms that do not rely on scenes or shots. The authors of [32] use events as the main retrieval units. The work in [31] is based on retrieval using human memory. The basic idea behind that approach is to help users remember what they are looking for. This is shown in Figure 5: the user remembers a retrieval cue and performs a query. After seeing the results of the query, the user remembers a new retrieval cue and performs a new query. The cycle repeats until the user finds the desired content. The GUI [57] (Figure 6) makes use of many data and metadata sources (e.g., room layout information, person location, etc.), and makes strong use of spatial information. In other words, retrieval is based on objects (people), events, and the locations of the events.

One desirable alternative for meeting video retrieval is to rely heavily on speech recognition (e.g., [63]). This can work extremely well, but it is problematic in most cases as speech recognition can be highly inaccurate, particularly in multi-language scenarios.

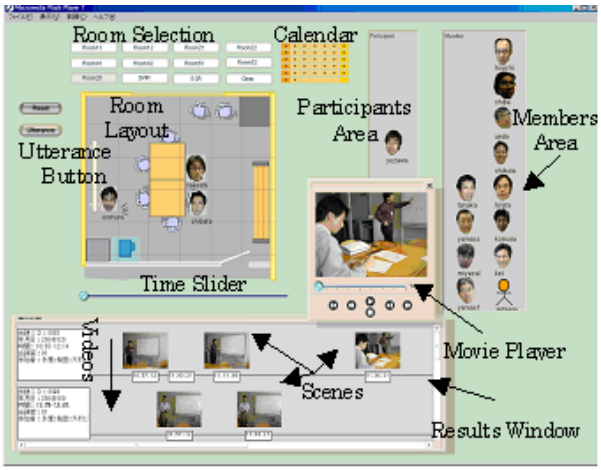

Figure 6. Memory interface.

6.4. The Video Gap: Open Issues and Research Directions

Meeting recording is quickly becoming an important area, in part, because of the large range of potential applications of keeping accurate records of one of the most essential modes of communication in any organization.

Before the videos from an SCR application are widely used, however, several issues must be solved, in the technical (infrastructure and data engineering) and human areas (how the contents are used). For example, in terms of infrastructure and data engineering, the following factors must be considered:

- *Security and privacy*: often, information shared during meetings is highly sensitive.
- *Scalability and efficiency:* in large organizations, the amount of data to be stored and accessed is very large. Effective mechanisms must be in place to handle an appropriate number of simultaneous users of the data. As the size of the meeting databases grow, it is essential to have efficient access mechanisms in place.
- *Network synchronization and delivery*: synchronization of multiple streams and devices is not trivial, but very important.
- *Data mining*: efficient methods must be developed to maximize the value of the data stored. Data mining is one option.
- *Data engineering:* all aspects of this field play an important role given the large amounts of data of multiple types and from multiple sources.

In terms of search, traditional database query paradigms may not suffice in this domain. Open issues include the development of new query languages suitable for audio-visual data (including metadata) and that consider different types of objects beyond what may be represented at the shot level.

One of the biggest problems in building a meeting video capture and analysis system is what we term the "video gap:" who the users will be and how the system will actually be used. Although many questions remain, it is our belief that user studies are necessary in both cases. In building the interface of Figure 6, for example, we conducted studies (reported in [31]) to determine what kinds of items people remember well and what kinds they do not. In [31] we also reported on a survey to determine possible uses for video, and the reasons video is not used often. We found that there is a *video gap* in three senses:

- Most places do not have an infrastructure for meeting recording, so it is difficult for people to record meetings and view them.
- Video has not been used traditionally in the meeting context, so most people do not know how to use it or why it would be useful (*building the infrastructure is not enough!*).
- Does the use of the video system have any impact on the workflow of the group using it?

Our current efforts are focusing on evaluating the retrieval methods to determine if the system we have constructed is indeed effective for retrieval (see [63] for a proposed methodology for evaluating browsers, also an important area).

7. CONCLUSIONS AND FUTURE WORK

We have discussed research and open issues in constructing SCRs like the one built at FXPAL Japan for

the purpose of automatic content analysis. Our discussion was grounded on a novel conceptual meeting model that consists of *physical* (from layout to cameras), *conceptual* (meeting types, actors), *sensory* (audio-visual capture), and *acquired content* (syntax and semantics) components. We discussed issues ranging from physical room layout and hardware infrastructure to technical issues related to automatic content analysis and metadata. Finally, we discussed human and social issues and the implications of the adaptation of Smart Conference Room (SCR) technologies.

An SCR project has many components. Our model is merely a starting point and further work is needed in modeling each of the sub-components of our framework so that they can be applied computationally. Future work includes addressing several of the open issues described throughout the paper. Of particular interest is the integration of metadata with automatic audio-visual analysis.

REFERENCES

[1] J. Ajmera, I. McCowan, and H. Bourlard, "Robust Speaker Change Detection," *IDIAP RR 02-39*, 2002.

[2] J. Ang, Y. Liu and E. Shriberg, "Automatic Dialog Act Segmentation and Classification in Multiparty Meetings," *IEEE ICASSP 2005*, Philadelphia, PA.

[3] S. Bengio and H. Bourlard, (eds.) *Machine Learning for Multimodal Interaction, First International Workshop, MLMI 2004*, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers. Springer 2004.

[4] H. Bounif, O. Drutskyy, F. Jouanot, S. Spaccapietra, "A Multimodal Database Framework for Multimedia Meeting Annotations," *10th International Multimedia Modelling Conference*, Brisbane, Australia, Jan. 2004.

[5] L. Chaisorn, T.-S. Chua, C.-H. Lee and Q. Tian, "A Hierarchical Approach to Story Segmentation of Large Broadcast News Video Corpus." *IEEE ICME'2004*. Taipei, Taiwan, June 2004.

[6] M. Chen, "Achieving Effective Floor Control with a Low-Bandwidth Gesture-Sensitive Videoconferencing System," Proc. *ACM Multimedia 2002*, Juan Les Pines, France, 2002.

[7] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox, "Room with a Rear View: Meeting Capture in a Multimedia Conference Room," *IEEE Multimedia*, Vol. 7, No. 4, pp. 48-54., Oct-Dec 2000.

[8] J. Connell, A.W. Senior, A. Hampapur, Y-L Tian, L. Brown, and S. Pankanti, "Detection and Tracking in the IBM PeopleVision System," *IEEE ICME 2004*, June 2004.

[9] C. Costa, P. Antunes, and J. Dias, "A Model for Organizational Integration of Meeting Outcomes," in *Contemporary Trends in Systems Development*, M.K. Sein, B.-E. Munk-vold, T.U. Ørvik, W. Wojtkowski, W.G. Wojtkowski, J. Zupancic, and S. Wrycza, Eds. Kluwer Plenum, 2001.

[10] R. Cutler, et. al., "Distributed Meetings: A Meeting Capture and Broadcasting System," Proc. *ACM Multimedia 2002*, Juan Les Pines, France, 2002.

[11] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting Recorder Project: Dialog Act Labeling Guide," *ICSI Technical Report TR-04-002*, 2004.

[12] D. Ellis and J. Liu, "Speaker turn segmentation based on between-channel differences," at *NIST Meeting Recognition Workshop at ICASSP 2004*, Montreal, May 2004.

[13] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting Group Interest-Level in Meetings," *IDIAP Research Report 04-51*, September 2004.

[14] W. Geyer, H. Ritcher, and G. Abowd, "Making Multimedia Meeting Records More Meaningful," *IEEE ICME 2003*, Baltimore, MD, July 2003.

[15] E. Goffman, *Behavior in Public Spaces*. Free Press, New York, 1963.

[16] http://www.amiproject.org/.

[17] http://www.is.cs.cmu.edu/meeting room/.

[18] http://www.m4project.org/.

[19] http://www.nist.gov/speech/test beds/mr proj/.

[20] http://www.icsi.berkeley.edu/Speech/mr/.

[21] http://www.sarbanes-oxley.com/

[22] N. Kern, B. Schiele, H. Junker, P. Lukowicz, G. Tröster, "Wearable Sensing to Annotate Meeting Recordings," In *Personal and Ubiquitous Computing: Selected papers from the ISWC2002 Conference*, 2003.

[23] A. Hakeem, M. Shah, "Ontology and Taxonomy Collaborated Framework for Meeting Classification", in proc. *ICPR 2004,* Cambridge, UK, August 2004.

[24] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection Of Agreement vs. Disagreement In Meetings: Training With Unlabeled Data," Proc. *HLT-NAACL Conference*, Edmonton, Canada, May 2003

[25] L. Hollink, A.Th. Schreiber, B. Wielinga, M. Worring, "Classification of User Image Descriptions," *International Journal of Human Computer Studies* 61/5, pp. 601-626, 2004.

[26] A. Jaimes. *Conceptual Structures and Computational Methods for Indexing and Organization of Visual Information*, Ph.D. Thesis, Department of Electrical Engineering, Columbia University, February 2003.

[27] A. Jaimes, Q. Wang, N. Kato, H. Ikeda, and J. Miyazaki, "Visual Trigger Templates for Knowledge-Based Indexing," in *proc. Fifth Pacific Rim Conf. On Multimedia (PCM '04) 2004*, Tokyo, Japan, Nov. 30-Dec. 3rd, 2004.

[28] A. Jaimes and S.-F. Chang, "A Conceptual Framework for Indexing Visual Information at Multiple Levels", in *Internet Imaging 2000, IS&T/SPIE*, San Jose, CA, January 2000.

[29] A. Jaimes, N. Yoshida, K. Murai, K. Hirata, and J. Miyazaki, "Interactive Visualization of Multi-Stream Meeting Videos Based on Automatic Visual Content Analysis" in *IEEE International Workshop on Multimedia Signal Processing (MMSP '04)*, Siena, Italy, Sept. 2004.

[30] A. Jaimes, T. Nagamine, J. Liu, K. Omura, and N. Sebe, "Affective Meeting Video Analysis," in proc. *IEEE ICME 2005*, Amsterdam, NL, July 2005.

[31] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata, "Memory Cues for Meeting Video Retrieval," *1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04)*, New York, NY, USA, October 2004.

[32] R. Jain, P. Kim, and Z. Li, "Experiential Meeting System," in *ACM Multimedia Workshop in Experiential Telepresence (ETP 2003),* Berkeley, CA, Nov. 2003.

[33] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, B. Wrede. "The ICSI Meeting Project: Resources and Research," *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, May 2004.

[34] Y. Kameda, Y. Atarashi, S. Nishiguchi, M. Minoh, "Reduction Of Camera Motion Adjustments Under A Planned Video Composition With Pan-Tilt Camera," *in proc. of Asian Conference on Computer Vision 2004 (ACCV 2004)*, Vol.1, pp.216-221, Jeju Island, Korea, 2004.

[35] L. Kennedy and D. Ellis, "Laughter Detection in Meetings," *NIST Meeting Recognition Workshop at ICASSP 2004*, Montreal, May 2004.

[36] L. Kennedy and D. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Automatic Speech Recognition and Understanding Workhop IEEE ASRU 2003*, St. Thomas, December 2003.

[37] H. Koike, S. Nagashima, Y. Nakanishi and Y. Sato, "EnhancedTable: Supporting a Small Meeting in Ubiquitous and Augmented Environment," in *proc. PCM 2004*, Tokyo, Japan.

[38] D.-S. Lee, B. Erol, J. Graham, H.J. Hull, and N. Murata, "Portable Meeting Recorder," in *proc. ACM Multimedia 2002*, Juan Les Pines, France, 2002.

[39] Q. Liu, D. Kimber, J. Foote, L. Wylcox, and J. Boreczky, "FLYSPEC: A Multi-User Video Camera System with Hybrid Human and Automatic Control", in *proc. ACM Multimedia 2002*, Juan Les Pines, France, 2002.

[40] S. Marchand-Maillet, "Meeting Record Modelling for Enhanced Browsing," *Technical Report Computer Vision and Multimedia Laboratory*, Computing Centre, University of Geneva, March, 2003.

[41] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP*, 2003.

[42] D. McNeil, *Hand and Mind: What Gestures Reveal About Thought.* U. of Chicago Press, 1995.

[43] I. Mikic, K. Huang, and M. Trivedi, "Activity Monitoring and Summarization for an Intelligent Meeting Room," in *proc. IEEE Workshop on Human Motion*, Austin, Texas, Dec. 2000.

[44] N. Moënne-Loccoz, B. Janvier, S. Marchand-Maillet and E. Bruno, "Managing Video Collection at Large," *First Intl. Wsp on Computer Vision meets Databases (CVDB 2004)*, Paris, France, June 13, 2004.

[45] S. Mukhopadhyay, and B. Smith, "Passive Capture and Structuring of Lectures," in *proc. ACM Multimedia '99*, Orlando, FL, 1999.

[46] R.Ogata, Y.Nakamura, Y.Ohta, "Computational Video Editing Model based on Optimization with Constraint-Satisfaction," *proc. Fourth Pacific-Rim Conference on Multimedia*, 2003.

[47] M.Ozeki, Y.Nakamura, Y.Ohta, "Automated Camerawork For Capturing Desktop Presentations – Camerawork Design And Evaluation In Virtual And Real Scenes," *Proc. 1st European Conference on Visual Media Production*, 2004.

[48] A. Pentland, "Socially Aware Computation and Communication," *IEEE Computer Vol 38,No.3*, March 2005.

[49] T. Pfau, D.P.W. Ellis, A. Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder," in *ASRU-01*, Italy, December 2001.

[50] F. Quek, D. McNeill, R. Bryll, C. Kirbas H. Arslan, K.E. McCullough, N. Furuyama, R. Ansari, "Gesture, Speech, and Gaze Cues for Discourse Segmentation," *in proc. of CVPR 2000*, Hilton Head Island, South Carolina, USA, 2000.

[51] D. Reidsma, R.J. Rienks and N. Jovanovic, "Meeting Modelling in the Context of Multimodal Research," in *Proceedings of MLMI '04*, LNCS, volume 3361, Springer Verlag, Martigny, 2005.

[52] Y. Rui, A. Gupta and J.J. Cadiz, "Viewing Meetings Captured by an Omni-Directional Carema," in *Proc. of ACM CHI 2001*, Seattle, WA, March, 2001.

[53] Y. Rui, A. Gupta, J. Grudin and L. He, "Automating lecture capture and broadcast: technology and videography," *ACM Multimedia Systems Journal*, 3-15, Springer V., 2004.

[54] A.W. Senior, S. Pankanti, A. Hampapur, L. Brown, Y-L Tian, and A. Ekin, "Blinkering Surveillance: Enabling Video Privacy through Computer Vision," *IBM Technical Report RC22886*, 2003.

[55] X. Sun, and B.S. Manjunath, "Panoramic Capturing and Recognition of Human Activity," in *proc. of IEEE ICIP 2002*, Rochester, NY, USA, September 2002

[56] T. Strothotte, H. Wagener: *Computational Visualization: Graphics, Abstraction, and Interactivity*. Springer 1999.

[57] T. Nagamine, A. Jaimes, K. Omura, and K. Hirata, "A Visuospatial Memory Cue System for Meeting Video Retrieval," *ACM Multimedia 2004*, New York, NY, USA, October 2004.

[58] http://www.quindi.com/

[59] M. Trivedi, I. Mikic, S. Bhonsle: "Active Camera Networks and Semantic Event Databases for Intelligent Environents", *IEEE Wsp on Human Modeling, Analysis and Synthesis (in conj. with CVPR)*, Hilton Head, South Carolina, June 2000

[60] Uchihashi S, "Improvising Camera Control for Capturing Meeting Activities using a Floor Plan," in *proceedings of ACM Multimedia 2001*, pp. 12-18, 2001.

[61] A. Waibel, et. al., "SmaRT: The Smart Meeting Room Task at ISL," in *IEEE ICASSP 2003*.

[62] A. Waibel et. Al., "Advances in Automatic Meeting Recording and Access," in *ICASSP 2001*, SLC, UT, 2001.

[63] P. Wellner, M. Flynn, S. Tucker, and S. Whittaker, "A Meeting Browser Evaluation Test," *IDIAP-RR 04-53*, 2004.

[64] B. Wrede and E. Shriberg, "Spotting Hot Spots in Meetings: Human Judgements and Prosodic Cues," in *EUROSPEECH 2003*, Geneva, September 2003.

[65] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models," *Pattern Recognition Letters*, Vol. 25, Issue 7, pp. 767-775, May 2004

[66] N. Yoshida, J. Miyazaki, and A. Wakita, "CandyTop Interface: A Visualization Method with Positive Attention for Growing Multimedia Documents," in *IEEE IV '03*.

[67] Zobl, M, Wallhoff, F and Rigoll, G, "Action Recognition in Meeting Scenarios Using Global Motion Features." Proc. *IEEE Intl. Wkshp on Perf. Eval. of Tracking and Surveillance (PETS-CCVS) Graz, Austria, March* 2003.