

Portable Meeting Recorder
A multimodal meeting recorder designed by Ricoh

Lawrence Michel
Université de Fribourg
Seminar - DIVA Research group
lawrence.michel@unifr.ch

March 2005

1 Introduction

A typical definition of a meeting could be understood as a process where information is shared between two or more meeting participants. Information transmission is commonly fulfilled within numerous media type, such as physical and electronic media (documents, slideshows), speech, visual activities, and so on. Another important component of a meeting is time. A meeting process is always delimited in time (e.g. there is always a starting and an ending event that will start, respectively end the process at a given time).

A common task that must be performed is a so-called summary of all significative information and events occurred within a meeting. Meeting segmentation, combining events and information together, structuring them logically within time and finally archiving in a sensefull way for future needs are predominant goals in multimodal meeting research. Said in another way, how could we maximise the information capture during an event and formulate their correlation logically?

Several attempts to solve this problematic have been tested. A common method would be to simply record the meeting and post-process it by hand. There is no doubt that it is already an effective way for extracting significant data, but it lacks the ability to capture important events, such as gesture, handnotes, working documents who might be strongly correlated to what is has been spoken at a given time. Multimodal capture has to fulfill multiple requirements, such as video, audio, electronic data extraction and a multimodal

processor has to be able to organize these datas, analyze them and standardize the overall structure in a way that we could access the content efficiently for future purposes. Most existing multimodal recorders are characterized by their strong audio and video capturing devices dependency with high intrusiveness and complexity of use. An important goal to achieve is to design a complete system fitting our requirements under constraints of maximizing ergonomoy and effectiveness.

The proposed solution is a portable meeting recorder that captures omni-directional audio and video during a meeting. The system permits on-the-fly analysis of the data that enables various output format, such as browsable video and audio, metadata structure for the logical structure. A relevant particularity remains in the fact that the designing process of our meeting recording has been done with the idea of portability and compatibility with common hardware in mind.

2 System description

The solution designed by Ricoh's engineers is dedicated to be a complete multimodal recording solution that should fit most of our requirements under constraint of ergonomoy, ease-of-use and portability. The hardware has been designed in a way that all capture components are located in one single device, under strong constraint of minimizing intrusiveness. The software part is handle by a common PC. Data transfer is actually done through a single-wire connectivity (FireWire). A third-part device as a meeting recorder interface is included

in the set that enables the user to gain control to the recording process of the meeting.

2.1 Hardware specification

The capture device is a compound of audio and video sub-devices. The video device is characterized by a 360 degree panoramic video camera and sound capture accomplished by four omni-directional microphones (each located at the square based edges of the compound)(fig.1).



Figure 1: The recording device and its browser interface

Moreover, data recording and processing is accomplished by a standard PC.

2.2 Software specification

The core of the system resides in the software which continuously computes the data streamed from the recording device. Several techniques have been applied to extract a maximum of information from basic audio and video data. We will take a specific look on particular ones in section ??.

3 The recording process

The complete recording process can be subdivided into three distinct steps :

3.1 Capturing data (step 1)

Video data is streamed "as-is" directly to the computer. The particularity of the video capture technique is the unconventional way to permit a 360

degree panoramic capture using a single standard low-cost camera focusing a panoramic mirror located at the top of the capture device (fig.2). Image stretching algorithms will be later used for user viewing conveniency.

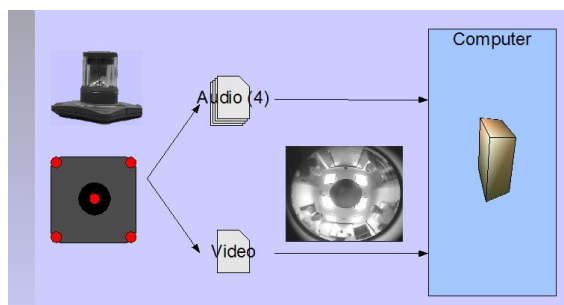


Figure 2: Audio and Video capture

3.2 Processing data (step 2)

Sound Localization algorithm based on input audio files and panoramic mpeg2 compression combined with image stretching algorithms are first computed. Their respective output will be mainly used for other computational process, such as View Selection, Face Extraction, Location Recognition, Motion Analysis and Audio Activity (fig.3).

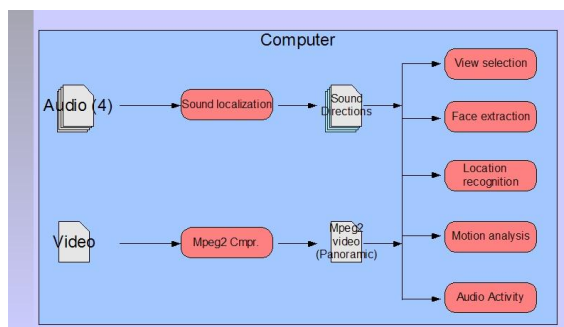


Figure 3: Audio and Video processing

We will discuss about Sound Localization and Meeting Location Recognition in section 4.

3.3 Processing and storing Meta-data (step 3)

The result of all enumerated algorithms are combined and structured into an XML metadata. The

choice for such standard is explained by its high portability capacity and compatibility with other multimodal systems, such as browsers. Metadata, audio and video files are stored into a database for future use (fig.4).

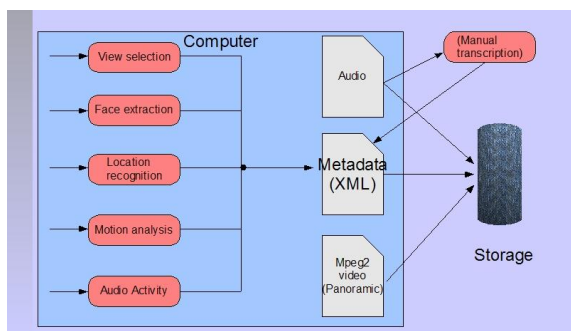


Figure 4: Metadata processing and final storage

4 Processing algorithms

In normal meetings, most interest is given in its speech and, if it is the case, in physical document content. In multimodal meeting, details are critical. Because every event during a meeting can be more or less strongly correlated to other ones, there is the need to capture as much as possible significant events that may help the overall metadata creation process.

The Portable Meeting Recorder system discussed here has been designed in a way to detect those details. An important challenge that the designers had to face is to implement efficient algorithms that must require few parameters as input (four audio signals and a single mpeg2 video stream in this case). We will keep our interest at two specific algorithms that have been selected.

4.1 The Sound Localization algorithm

360-degree sound localization is calculated as follows. For each pair of microphones on the diagonal of the device, an angle between 0 and 180 degrees is calculated. Phase difference calculation has been used to "guess" the midpoint from each sound source. Since locutors are located somewhere in a 3D space, the system needs to detect the

elevation of the source. Multiple techniques have been used to maximize the accuracy of the result. However, in trying to estimate both azimuth and elevation, the solution remained unstable and very sensitive to multiple factor (random noise, sound quantization, distance, and so on).

After computing these data, output is then used within other algorithms, such as Face detection, View Selection, Audio Activity. Because the effectiveness of those algorithms mostly relies on the data computed within the primary one, and after taking in consideration that processing through our first algorithm remains highly parameter-sensitive, the whole system won't be able to guarantee its reliability.

Some remarks :

- This method is processed at real-time (requires 30-40% CPU load in a 933MHz PC).
- Accuracy remains highly dependent on several factors, such as room specification (e.g. reflective surfaces that leads to high signal reverberation), amplitude of the source signals, case of speech overlap, particular source angles.
- Hardware dependency : Accuracy is strongly dependent to signal sampling rate, sensitivity of the capture devices.

4.2 The Meeting Location Recognition algorithm

This algorithm is applied on the video stream. The problematic to be solved is to make the system be able to detect where the meeting takes place. Since it is unable to guess accurately the location, several pattern matching algorithms are used. Because we are only focusing on location characteristics, we need to remove each moving segments of the sample image. This process is as follows :

1. Analyzing frame by comparing its histogram with a template
2. Applying foreground extraction
3. Resulting background image will be set as the newest template

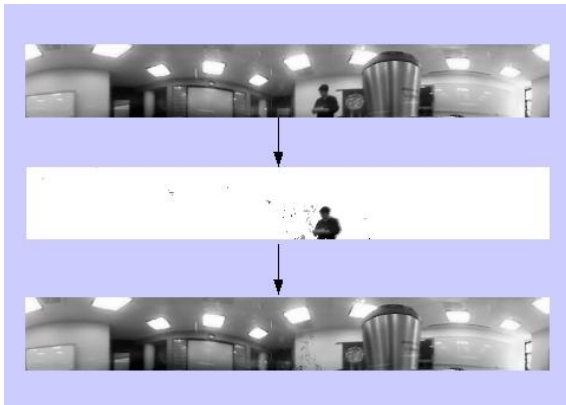


Figure 5: Pre-processing image

Figure 5 illustrates this process.

After having extracted all foreground items from the sample video, the basic structure of the image is compared with predefined patterns from a database.

5 Searching and browsing with Visual and Audio Content

The Meeting Recording System is, in ideal circumstances, able to extract relevant information from captured audio and video. We have seen the basic overall process and took a specific look on two important algorithms. The recording system finally stores all computed data within a structured XML metafile. Audio and video streams are at the same kept within a storage server. We now have a consistent meeting database for future purposes, such as meeting Browsing.

Due to its complex informational structure, searching and browsing such audiovisual information is a high time and effort consuming task. Designing a user-friendly, efficient, meeting browser is actually a challenge for most research labs. The Meeting Recording System is yet still unable to automatically transcript audio speech. This fact is basically due to lack of accuracy of speech recognition systems. Furthermore, assuming that accurate speech recognition system has been designed, their efficiency might be highly decreased in case of speech overlap (which occurs quite often during a typical

meeting).

Ricoh's designers followed another approach. The basic idea is to implement a browsing technique basically based on visual and audio activity. The user is guided through the meeting by graphs and tags within time axe that may point him to relevant events (fig. ??).

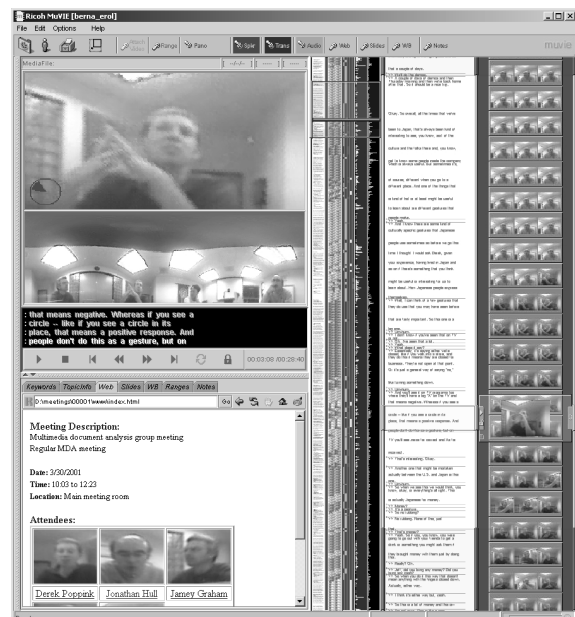


Figure 6: Visual and Audio Content

The meeting segmentation process relies actually exclusively on audio and video streams. Fig.7 gives us a closer look at the browser.

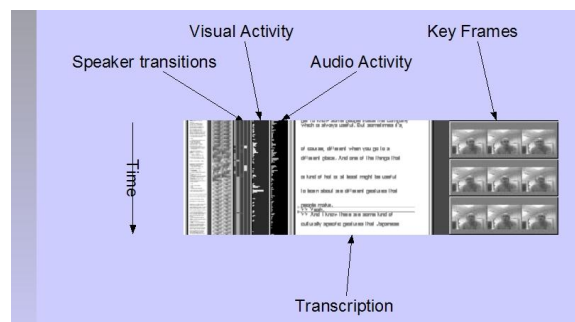


Figure 7: Visual and Audio Content (2)

5.1 Visual Activity Analysis

In normal cases, a meeting video sequence won't contain much relevant motions. In order to segment video input into a sequence of variable length segments, motion analysis algorithm is applied such that it must detect high motion segments from the stream and mark them as significant. Marks are then distributed within a time axis according to the video stream.

5.2 Audio Activity Analysis

Segmenting audio is quite a similar task. In order to segment audio stream according to time axis in a relevant way, sound level evolution detection algorithms are then used: someone starting talking, or might increase volume of its speech would be enough to fit the segment marking condition. On the other hand, the system loses efficiency when audio is badly sampled, contains too much noise, and so on.

6 Conclusion

Nowadays, it is actually no doubt that the demand for multimodal recording system increases. Congresses, business meetings, educational purposes are good examples of where relevant information is created and shared. Capturing them and organizing them for future purpose is an important task to be achieved. Ricoh's solution is an attempt to satisfy this demand. The designers designed the system under constraints of simplicity of use, ergonomics and compactness. However, effectiveness of the system is strongly correlated to external (and internal) factors.

Innovation also took place in the way they designed the meeting browser. Having control to the most information captured during a meeting is done using segmented audio and video events. It's no doubt that there is still room for future improvements.

References

- [1] Dar-Shyang Lee, Berna Erol, Jamey Graham, Jonathan J. Hull and Norihiko Murata, *Portable Meeting Recorder*, ACM Multimedia 2002, pp. 493-502, Juan Les Pins, France, 2002.
- [2] Jamey Graham and Jonathan J. Hull, *A Paper-Based Interface for Video Browsing and Retrieval*, IEEE Int. Conf. on Multimedia and Expo (ICME), Baltimore, July 6-9, 2002.