Multimodal Summarization of Meeting Recordings

Berna Erol, Dar-Shyang Lee, and Jonathan Hull
Ricoh California Research Center
2882 Sand Hill Rd. Suite 115, Menlo Park, California, USA
{berna_erol, dsl, hull }@rii.ricoh.com

***Abstract* — Recorded meetings are useful only if people can find, access, and browse them easily. Key-frames and video skims are useful representations that can enable quick previewing of the content without actually watching a meeting recording from beginning to end. This paper proposes a new method for creating meeting video skims based on audio and visual activity analysis together with text analysis. Audio activity analysis is performed by analyzing sound directions — indicating different speakers— and audio amplitude. Detection of important visual events in a meeting is achieved by analyzing the localized luminance variations in consideration with the omni-directional property of the video captured by our meeting recording system. Text analysis is based on the Term Frequency–Inverse Document Frequency measure. The resulting video skims better capture the important meeting content compared to the skims obtained by uniform sampling.**

1. INTRODUCTION

Capturing the content of meetings is useful for many applications. Audio-visual recordings of meetings provide the capability for reviewing and sharing meetings, clarifying miscommunications, and thus increasing efficiency. Recognizing this need, several meeting recorder systems have been developed in the recent years [1][2].

Enabling efficient access to captured meeting recordings is essential to benefit from this content. Searching and browsing audiovisual information can be a time consuming task. The two most common approaches for overcoming this problem are key-frame based representations and summarization with video skims. Key-frame based representations are useful for video browsing [3], as they give a quick overview of the multimedia content. On the other hand, video skims are content-rich summaries that contain both audio and video. Efficiently constructed video skims can be used like movie trailers to communicate the essential content of a video sequence. In [4], Waibel et al. propose summarization of meeting content using video skims with a user-determined length. The skims are generated based on relevance ranking and topic segmentation using speech transcripts. A summarization technique for educational videos based on shot boundary detection, followed by word frequency analysis of speech transcripts, is suggested in [5] by Taskiran et al. The comparison of the efficiency of their skims with randomly generated skims does not reveal any significant performance difference. They conclude that this is possibly due to their evaluation method. In [6], He et al. present a method for summarizing audio-video presentations using slide transitions and/or pitch activity. They compare the efficiency of these summaries with the efficiency of audiovisual summaries created by the author of the presentation. The efficiency of summaries is measured by performance improvements on quizzes given before and after watching the video skims. As can be expected, author-generated summaries were more efficient for communicating content, but automatically generated summaries also resulted in substantial quiz score improvement. Furthermore, as a part of the Informedia™ project, in [7] Christel et al. compared video skimming techniques that used: (1) audio analysis based on audio amplitude and term frequency-inverse document frequency (TF-IDF) analysis, (2) audio analysis combined with image analysis based on face/text detection and camera motion, and (3) uniform sampling of video sequences. They reported that audio analysis combined with visual analysis yield significantly better results than the skims obtained purely by audio analysis and uniform sampling.

Speech content and natural language analysis techniques are commonly used for meeting summarization. However, language analysis-based abstraction techniques may not be sufficient to capture significant visual and audio events in a meeting, such as a person entering the room to join the meeting or an emotional discussion. In this paper, we describe a multimodal meeting summarization technique, which considers the audio and visual content along with the speech transcriptions. We propose a novel measure for finding the important audio segments in a meeting recording. Our measure is based on sound localization output and the magnitude of the audio signal, and is for detecting segments with a high degree of interaction between the participants and the presence of loud speech. We also propose a novel measure for detecting the important visual events in a meeting. Meeting sequences generally contain a very small amount of visual activity. Also, because our meeting recorder captures omni-directional video, there is no camera motion and therefore no scene breaks. Considering these properties of our meeting video collection, the proposed visual activity measure is based on localized differences of luminance values in the compressed domain. Lastly, language analysis is performed using the well-known Term Frequency-Inverse Document Frequency (TF-IDF) measure, which indicates the relative importance of the words in a document. The important segments identified by using these three analysis techniques are then combined to create meeting video skims.

In the next two sections, the proposed audio and visual activity measures are described. Section 4 explains the language analysis technique used to create meeting skims. In Section 5, we describe the multimodal scheme used for meeting summary generation. Experimental results are given in Section 6 and conclusions and future work are discussed in Section 7.

2. AUDIO ACTIVITY ANALYSIS

Analysis of the audio signal is useful in finding segments of recordings containing speaker transitions, emotional arguments, and topic changes, etc. In [8], Dagtas et al. find the important segments of a sports video based on audio magnitude. In [6] and

2003 IEEE Int. Conf. on Multimedia and Expo (ICME),July 6-9, Baltimore, MD.

[7], pitch and audio power are employed to find topic changes and significant audio events, respectively.

Our goal is to find segments containing arguments and discussions among meeting participants. Even though high audio amplitude provides a good indication of someone raising their voice and the presence of emotion, our experiments showed that amplitude by itself is not sufficient to capture such segments. Our solution is to combine this information with the sound localization output of our meeting recorder system. This is motivated by the fact that sounds coming from different directions in a short time window is potentially an indication of a discussion between several speakers. We define a speaker activity measure, S_a, as follows.

$$S_a(t) = \sum_{n=-W/2}^{n=W/2} G(n)C(t+n),$$

where t is time in seconds, $G(n)$ is the smoothing filter coefficient, W is the length of the filter, $C(t)$ is the number of changes in the sound direction at t. Audio activity measure, U_a, is defined as

$$U_a(t) = S_a(t) \times \sum_{k=-f/2}^{f/2} \left| X(f\,t+k) \right|,$$

where $S_a(t)$ is speaker activity at t, f is the audio sampling frequency, and $X(n)$ is the audio amplitude of the n^{th} sample.

Figure 1 shows the sound localization and the audio activity outputs, as well as the transcription corresponding to the peaks in the audio activity measure, for a staff meeting. It is challenging to evaluate the performance of the audio activity measure. Nevertheless, our initial experiments with several recordings of staff meetings, presentations, and brain storming sessions, showed that the peaks in the proposed audio activity measure correspond to the audio segments with a high degree of meeting participant interactions and very few silent periods.

3. VISUAL ACTIVITY ANALYSIS

Motion content in video can be used for efficiently searching and browsing particular events in a video sequence as demonstrated in various applications [9][10]. In meeting sequences, most of the time there is minimal motion. High motion segments usually correspond to significant events such as a participant getting up to make a presentation, or someone joining the meeting.

Several motion activity descriptors exist in the literature. Some of these descriptors are based on the magnitudes and directions of the motion vectors in the MPEG bitstream [11]. The visual activity measure we employ uses the local luminance changes in a video sequence. A large luminance difference between two consecutive frames is generally an indication of a significant content change, such as a person getting up and moving around. However, other events, such as dimming the lights or all the participants moving slightly, may result in a large luminance difference between two frames. In order to eliminate such events, we define the visual activity as the luminance changes in a small window rather than luminance change in a whole frame.

The luminance changes are found by computing the luminance difference between consecutive intra coded (I) frames in the MPEG-2 stream. We employ I-frames because the luminance values in I-frames are coded without prediction from the other frames, and they are therefore independently decodable [12]. We compute luminance differences on the average values of 8×8 pixel blocks obtained from the DC coefficients, which are extracted from the MPEG bit stream without full decompression.

Because the video in our system is doughnut shaped (omni-directional), the pixels in the outer parts of the video contain less object information (i.e. more pixels per object). Therefore, when computing the frame differences, the pixel values are weighted according to their location to compensate for this. The assignment of weights is done by considering the parabolic properties of the mirror as follows:

$$w(r) = 1/\cos^{-1}\left[\frac{1-4(r/R_{\max})^2}{1+4(r/R_{\max})^2}\right],$$

where r is the radius of the DC coefficient location in frame centered polar coordinates and R_{max} is the maximum radius of the doughnut image. The coefficients that do not contain any information (locations that are outside of the mirror area) are weighed zero.

We employ a window size of 9×9 DC coefficients, which corresponds to a 72×72 pixel area. The weighted luminance difference is computed for every possible location of this window in a video frame. The local visual activity, V_a, is defined as the maximum of these differences as follows:

$$V_a = \max\{ \sum_{n=-L/2}^{L/2} \sum_{m=-L/2}^{L/2} (\omega(\sqrt{(x+n)^2+(y+m)^2})\, A_{x+n,\,y+m})\},$$

$$\forall x = [-W/2 + L/2 \ldots W/2 - L/2], \forall y = [-H/2 + L/2 \ldots H/2 - L/2].$$

where W and H are the width and height of the video frame (in number of DC blocks), L is the size of the small local activity frame (in number of DC blocks), ω(r) is the weight of the DC block at location r (in polar coordinates), and A_{ij} is the luminance difference between two blocks at location (i×8,j×8) in two consecutive I frames.

Figure 2 shows a plot of the local visual activity measure for a meeting video. As shown in the figure, most peaks in the visual activity score correspond to significant visual events, for example, a person taking his place at the table (Figure 2.a), another person leaving the meeting room (Figure 2.c), entering the room (Figure 2.d and Figure 2.e), etc. On the other hand, the video segment shown in Figure 2.b does not have a visual significance. This segment has a large activity value because the person moved close to the camera and appeared as a large moving object because of the perspective. Exclusion of such segments from the important visual events is possible only if we compensate for the distance of the objects from the camera.

4. TEXT ANALYSIS

Language analysis techniques are commonly used to summarize documents and audio transcriptions. Here, we compute the well-known Term Frequency-Inverse Document Frequency (TF-IDF) [13] on meeting transcriptions in order to find segments that contain important keywords. TF-IDF is defined as *TF-IDF=tf/df*, where *tf* is the frequency of a word in a document and *df* is the frequency of the same word in collection of documents. This measure is employed in our summarization system as follows. First, words with the highest TF-IDF score are defined as keywords. In order to find when a given keyword most frequently occurs, we divide audio transcriptions into 10-second segments and compute a document occurrence score as follows

$$DO_k(i) = \sum_{n=-W/2}^{n=W/2} G(n)\, O_k(i+n),$$

where i is the audio segment number, $G(n)$ is the smoothing filter coefficient, W is the length of the smoothing filter, $O_k(i)$ is the number of occurrences of the keyword k in a 10 second segment. The audio segment with the highest DO_k value is defined as the *keyword segment* for keyword k. These segments are then used in skim generation as described in the next section.

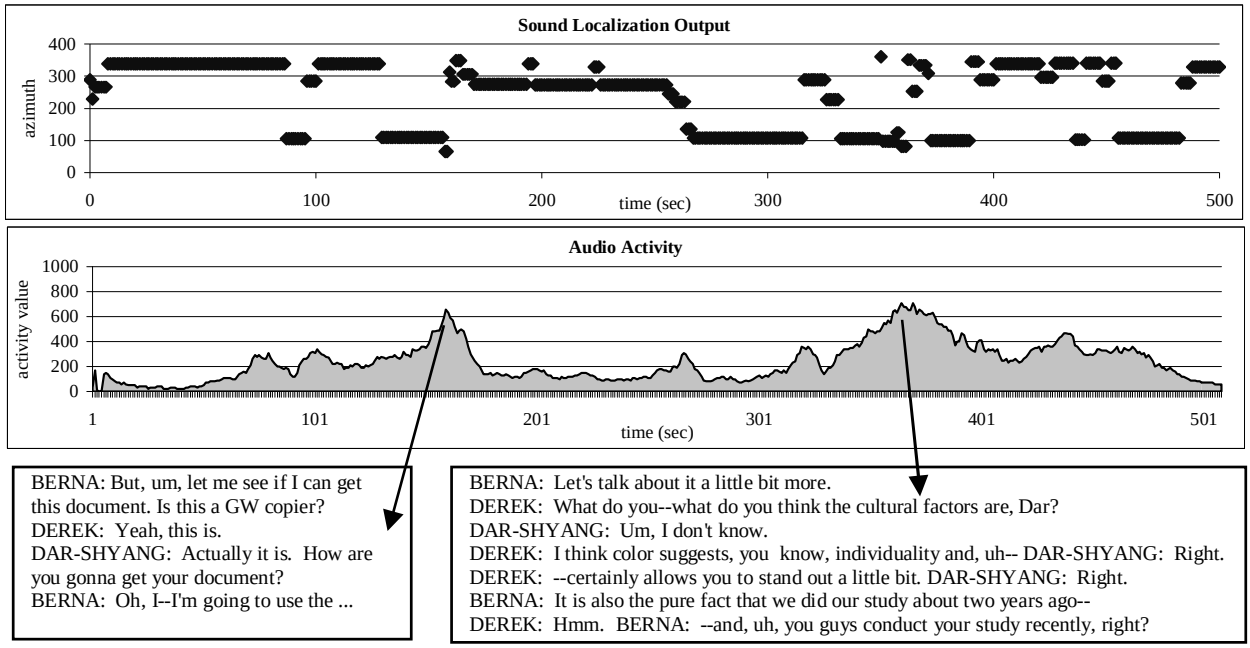

Figure 1. Sound localization and audio activity output for a staff meeting recording.

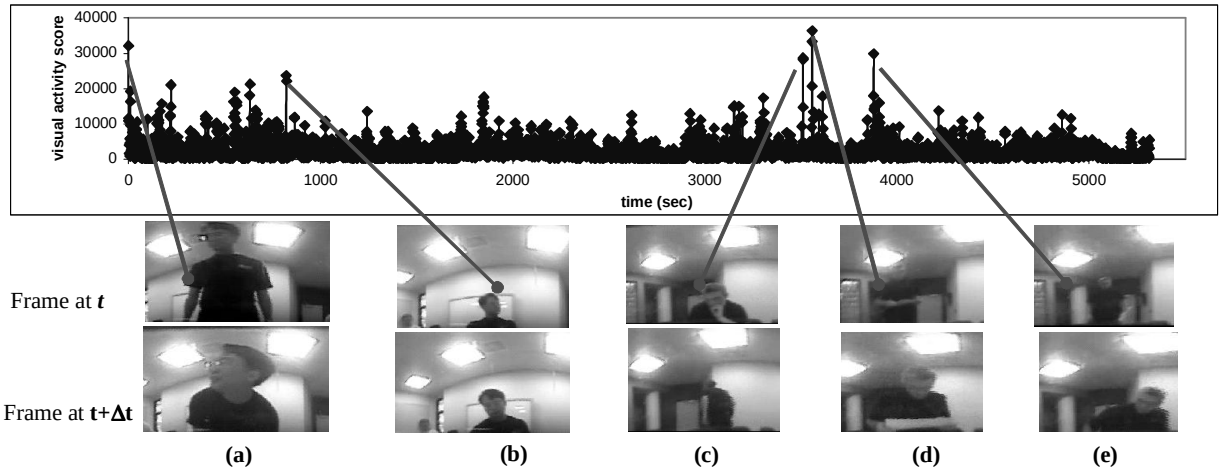

Figure 2. Examples of high visual activity scores corresponding to significant visual events.

5. COMBINING AUDIO, VISUAL, AND TEXT ANALYSIS FOR SKIM GENERATION

Our goal here is to find meeting segments that communicate the salient content of a meeting most efficiently. However, it is difficult to determine how much of audio and visual events, and spoken content relate to the important meeting segments. This is a challenging task requiring large datasets and a good understanding of how much each modality contributes to the final result. In this paper, we leave this as a future research topic and employ a scheme that gives equal importance to each modality. The audio and visual events are first sorted according to the activity scores (V_a and U_a's, respectively), and keyword segments (with the highest DO_k's) are sorted according to the TF-IDF of the keywords. Then, we include the first N important audio, visual, and keyword segments in the meeting skim in the time order. Padding is also applied to keep the meeting segments more comprehensible. If any two segments are close, they are merged to be one segment for the same reason.

Our initial experiments showed that using transition effects between different video segments in a video skim is useful for better comprehension of the skim. A similar conclusion is also reported in [7]. The video skims used in our study are generated with wipe transitions (right to left) between video segments.

6. RESULTS

In this section, we compare the effectiveness of the video skims that are generated by our method to that of the video skims obtained by uniform sampling of a meeting recording. In our experiments, we use two meeting recordings. The first one, Toyossan, is a 40 minute meeting in the form of a brainstorming session where the participants discuss a future product line for a company. The second one, NAB, is a 20 minute meeting where one of the participants reports to his group about a technology show he attended. In both meetings, a whiteboard is used and a participant joined after the meeting started. These two meetings are

summarized by using both our proposed method and uniform sampling. The skim length is 4 min for the Toyossan meeting and 2 minutes for the NAB meeting, which correspond to 10% of the meeting duration.

It is widely accepted that evaluating summary generation algorithms is a difficult task. Here, we employ the quiz method, where subjects are asked questions about a recording after watching a summary. There have been problems associated with this evaluation technique reported in the literature [5], however a better summary evaluation method is currently not available.

The effectiveness of the two summary methods are evaluated as follows. Several participants of the Toyossan and NAB meetings prepared a multiple-choice quiz (without seeing any of the skims) and two groups were asked to take the quiz after watching the meeting skims. We used two groups with 8 people of similar educational levels. One group was shown only the meeting skim generated by our method (*AVT*) and the other group was shown only the skim generated with uniform sampling (*USP*). The quiz scores are presented in Table 1. As can be seen from the table, the skims generated using the proposed method obtained higher scores in both videos. The score difference was much more significant in the case of NAB meeting, however it is minimal in the case of Toyossan meeting. Later, we studied the information captured by the two summarization methods by looking at how many of the answers to the quizzes were actually present in the skims. In both the Toyossan and NAB meetings, a significantly larger number of quiz answers were captured in the summary obtained by the proposed method. However, this difference is not accurately reflected by the quiz results. This is probably because of the small size of our groups and/or the fact that people cannot efficiently observe a large amount of information presented in a short time.

Another observation of our study was that uniform sampling gives a better sense of the span of the events in meetings. For example, if three out of five video segments in a meeting skim showed a participant presenting slides, one can easily imagine that the participant gave a presentation in approximately 60% of the meeting. On the other hand, it is not possible to make such conclusions by looking at the skims generated by the proposed method, as all of the three video segments could have been taken from a small time frame. One solution to this problem is to display the actual time of a video segment in the meeting recording during the playback. Alternatively, the time stamps can be inserted between video segments when wipe transitions occur.

	NAB meeting		Toyosan meeting	
	AVT	USP	AVT	USP
Standard dev.	0.05	0.09	0.05	0.1
Average score	80%	51%	65%	61%

Table 1. Quiz scores for the NAB and Toyosan skims using the proposed (ATV) and uniform sampling (USP) methods.

7. Conclusions and Future Work

Frequently, meetings last hours and have many lengthy boring parts. Often people are interested in seeing only the interesting and important segments of a meeting instead of watching an entire recording. In this paper, we described a solution for automatically creating such meeting skims. We proposed methods for visual and audio event detection, and combined these with a text analysis method in a multimodal scheme for skim generation. Meeting summaries obtained using our method resulted in better quiz scores than those obtained with uniform sampling. In addition to skim generation, the proposed method can be employed for efficient browsing of meeting videos. Important visual, audio, and keyword segments can be highlighted in an interface as guidance for navigation of long meetings. In our study, we assumed that the presentation slides, whiteboard and document capture were not available. In the future, we plan to utilize these as additional modalities for skim generation.

Video skim generation is a new research topic, and many of the issues remain to be understood. Selection of the video segment lengths included in skims, the relative importance of visual and audio events and transcription analysis, presentation of video skims, and methods for summary evaluation, are subjects for further research.

8. References

[1] Foote, J. and Kimber, D., “FlyCam: Practical panoramic video and automatic camera control,” Proceedings of International Conference on Multimedia & Expo, vol.3, pp.1419-1422, 2000.

[2] Gross, R., Bett, M. Yu, H., Zhu, X., Pan, Y., Yang, J., Waibel, A., “Towards a multimodal meeting record,” Proceedings of International Conference on Multimedia and Expo, pp. 1593-1596, New York, 2000.

[3] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video Manga: Generating Semantically Meaningful Video Summaries.",ACM Multimedia, (Orlando, FL) ACM Press, pp. 383-392, 1999.

[4] A. Waibel, M. Bett, et al., "Advances in Automatic Meeting Record Creation and Access," Proceedings of ICASSP, 2001.

[5] C. Taskiran, A. Amir, D. Ponceleon, and E. J. Delp, "Automated Video Summarization Using Speech Transcripts," SPIE Conf. on St. and Ret. for Media Databases, pp. 371-382, 2002.

[6] He, L., Sanocki, E., Gupta, A., and Grudin, J., "Auto-summarization of audio-video presentations," In Proc. ACM Multimedia, 1999.

[7] Christel, M., Smith, M., Taylor, C.R., and Winkler, D. “Evolving Video Skims into Useful Multimedia Abstractions,” Proc. of the ACM CHI, pp. 171-178, 1998.

[8] S. Dagtas, M. Abdel-Mottaleb, "Extraction of TV Highlights using Multimedia Features", Proc. of MMSP, pp. 91-96, 2001.

[9] Pingali, G. S., Opalach, A., Carlbom, I., "Multimedia retrieval through spatio-temporal activity maps", ACM Multimedia, pp. 129-136, 2001.

[10] Divakaran, A., Vetro, A., Asai, K., Nishikawa, H., "Video browsing system based on compressed domain feature extraction", IEEE Transactions on Consumer Electronics, vol. 46, pp. 637 - 644, 2000.

[11] Dorai, C., Kobla, V., “Perceived visual motion descriptors from MPEG-2 for content-based HDTV annotation and retrieval”, IEEE 3rd Workshop on Multimedia Signal Processing, pp. 147-152, 1999.

[12] ISO/IEC, “Information technology - generic coding of moving pictures and associated audio information: Video,” 13818-2, 1995.

[13] G. Salton, Automatic Text Processing, Addison-Wesley, 1989.