Multimodal Analysis of Recorded Video for E-Learning

Thomas Martin
MICA Center [*], L3i [†]
thomas.martin@mica.edu.vn

Alain Boucher
IFI [‡], MICA Center
alain.boucher@auf.org

Jean-Marc Ogier
L3i
jmogier@univ-lr.fr

ABSTRACT

In this paper, we present a model for multimodal content analysis. We distinguish between media and modality, which helps us to define and characterize three inter-modal relations. Then we apply this model for recorded course analysis for e-learning. Different useful relations between modalities are explained and detailed for this application.

Categories and Subject Descriptors: I.2.10 [Artificial Intelligence]: Vision and Scene Understanding— *Video analysis*

Keywords: E-learning, media understanding, multimodal analysis, video analysis.

1. INTRODUCTION

Nowadays, as the available multimedia content grows every day, the need for automatic content analysis is becoming more and more important. For example, information retrieval in broadcast news archives requires to index the different medias present in these archives. As one look to the litterature, lot of research work has been done on media content analysis, but most of these works are focused on one sole media. But such approaches have some limits. The most important one is that information is not concentrated and limited into one specific media, but is spreaded among all the medias. Recently, many research works have focused on the use of multiple medias (or modalities, as we will use this name in this paper) to increase the potentiality of analysis. However, to our knowledge, there is no existing framework for multimodal analysis, and there is only few serious analysis of the possibilities of interaction between modalities. In the next section, we will give some definitions, followed by our model for multimodal content analysis, and completed with a review of the existing litterature in this domain. Then, we will analyze one application, to enhance and describe the possible interactions that can exist between modalities in different situations. We will conclude with a discussion on inter-modal relations.

2. MULTIMODALITY

There is a confusion in the litterature between concepts of media and modality. In many papers, the authors use both words refering to the same concept. This does not seems to be exact as we can identify two different concepts in the context of content analysis. We propose here to define a modality as a refinement of the media concept. A media is characterized mostly by its nature (for example audio, video, text), while a modality is characterized by both its nature and the physical structuration of the provided information (for example X-Ray image, MRI images). One media can then be divided in multiple modalities, following two criteria: semantic structuration of the information and algorithms involved in the analysis process. While the concept of media is more independant from the application, we see the concept of modality as application dependant.

Many generic modalities can be listed in three main modality families. First, audio family includes different modalities in terms of structuration like speech, music or sound. Second, visual family in which we distinguish between still image and image sequence (video) because while both being acquired from a camera, image sequence contains time information. Third, the text family includes printed text and manuscript text. This split of media into modalities can surely be discussed and different organization can be proposed. But we will use this scheme through this paper using several examples taken from one application to illustrate our choice.

After having defined what is a modality, the next step is to define multimodality. Applied to video indexing, Snoek and Worring [2] have proposed to define multimodality as "the capacity of an author of the video document to express a semantic idea, by combining a layout with a specific content, using at least two information channels". Thus, they locate inter-modal relation at a high level using semantic. However, in the context of speech recognition, Zhi et al. [5] have implemented the multimodal integration at low level, just after the feature extraction phase and an alignment step. Both these definitions are incomplete. We propose a more general definition for multimodality as an interaction process between two or more modalities. Most of the multimodal applications found in the existing litterature uses two modalities, audio and video. Examples of such works include applications for video indexing [4] or emotion recognition [3]. Zhu and Zhou [6] combine audio and video analysis for scene change detection. Zhi et al. [5] use facial analysis (video) to improve speech recognition (audio). Zotkin et al. [7] propose a tracking method based on multiple cameras

[*] MICA Center, C10, Truong Dai Hoc Bach Khoa
1 Dai Co Viet, Hanoi, Vietnam

[†] L3i - Univ. de La Rochelle, Pôle Sciences et Technologie
17042 La Rochelle cedex 1, France

[‡] IFI, ngo 42, Ta Quang Buu, Hanoi, Vietnam

MM'05, November 6–11, 2005, Singapore.
ACM 1-59593-044-2/05/0011.

and a microphone array. Bigün et al. [1] propose biometric authentication using three modalities: fingerprint, face and speech.

We can identify three different types of inter-modal relations described using three different properties: direction, arity and interaction level. First, trigger relation is the most simple relation. It is a monodirectional relation which takes place when an event detected in one modality activates an analysis process in another modality. The interaction level relation is not used for the trigger relation. Second, integration relation is already widely used and is mainly characterized by its interaction level. The analysis processes are done separately for each modality, but followed by a process of integration of their results, as presented in [2] for multimodal video indexing. The third relation is collaboration. It is the strongest multimodal relation, consisting in a close interaction of two modalities during the analysis process. The results of the analysis of one modality are used for analyzing a second one.

3. VIDEO ANALYSIS FOR E-LEARNING

Our main application for multimodality is e-learning through the MARVEL project. The goal of MARVEL (Multimodal Analysis of Recorded Video for E-Learning) is to produce tools and techniques for creation of multimedia documents oriented for e-learning in developing countries. The complete course of a professor is recorded in live. The recorded material from live courses – and other optional textual sources – is analyzed and used to produce interactive e-courses. The slides used by the professor in the class can be automatically replaced by an appropriate file in the e-course, being synchronized with the professor explanations. The course given by the professor is indexed using various markers, from speech or image analysis. The main aim of this project consists in providing semi-automatic tools to produce e-learning courses from recorded live normal courses.

In this project, three different medias are available: audio, video and text. Following the model proposed in section 2, we have identified four different modalities: *printed text* which contains the text of the slides and, if available, from other external textual sources; *manuscript text* which represents the text written on the blackboard; *image sequence* which contains the motion content of the video media; *speech* which contains the teacher's explanations. The *printed text* modality is avalaible in two different medias: video and text. This is a good example to illustrate our distinction between media and modality (section 2). Even if available in two different medias, the *printed text* still contains the same information, with the same structuration. Once detected and extracted from the video media, the analysis processes involved are similar whatever the media.

The application is divided into two distinct parts: scenario extraction and content indexing. The scenario is given mainly by the video. The teacher's behavior has to be analyzed to extract the course scenario (explaining the current slide, writing on blackboard, talking to the class, ...). This will be used later as a layout during the e-course production. Other regions of interest such as the screen or the blackboard have to be detected. Detection of slide changes or new writing on the blackboard are events that will be used. The content indexing part of available media has to be done using the speech given by the teacher, the printed text on the slides and the manuscript text on the blackboard. These three sources are complementary to show all the content of the course.

Different inter-modal interactions are identified here. During the first part of the application (scenario extraction), three trigger relations are involved. These relations are directly related to the actors who interact in a course: teacher, blackboard and screen. The trigger source is the *image sequence* modality. First, the "slide transition" event triggers the *printed text* detection and recognition. Second, the "teacher shows screen" event triggers the point of interest search. Third, similar to the first, the "teacher writes on blackboard" event triggers the *manuscript text* recognition process. The second part of the application (content indexing) contains most of the inter-modal relations. First, the *speech-printed text* interaction. This is a bimodal and bidirectional collaboration interaction, with its main direction from *speech* to *printed text*. This is particularly true if the *printed text* modality is only available in the video media. In case of noisy environment, cross-recognition of both *speech* and *printed text* is possible and useful. In this case, *visual-speech* interaction can be also useful [5]. Recognition of *manuscript text* is a hard task, especially in video. We propose to help recognition of *manuscript text* using both *speech* and *printed text* modalities. Both relations, *speech-manuscript text* and *speech-printed text*, are bimodal and unidirectional.

4. CONCLUSION

In this paper, we have analyzed multimodality and identified three different types of inter-modal relations. The first relation is trigger which needs synchronized modalities. The second relation is integration which needs data with a same structuration level as input. In fact, this relation takes the analysis results of two or more modalities to provide higher level information. The third relation is collaboration which cannot be used between all the modalities and is really useful in case of modalities with different structuration. For example, speech recognition can help text recognition in video but the opposite is not obvious, except maybe for noisy environment. In that case, where the speech recognition rates quickly decrease, the use of video can improve recognition.

5. REFERENCES

[1] J. Bigün, J. Fiérrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Multimodal biometric authentication using quality signals in mobile communications. In *Proc. of ICIAP*, pages 2–11. IEEE Computer Society, 2003.

[2] C. G. M. Snoek and M. Worring. Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[3] M. Song, J. Bu, C. Chen, and N. Li. Audio-visual based emotion recognition - a new approach. In *CVPR (2)*, pages 1020–1025. IEEE Computer Society, 2004.

[4] S. Tsekeridou and I. Pitas. Audio-visual content analysis for content-based video indexing. In *ICMCS, Vol. 1*, pages 667–672. IEEE Computer Society, 1999.

[5] Q. Zhi, M. Kaynak, K. Sengupta, A. D. Cheok, and C. C. Ko. Hmm modeling for audio-visual speech recognition. In *Proc. of ICME*, pages 201–204. IEEE Computer Society, 2001.

[6] Y. Zhu and D. Zhou. Scene change detection based on audio and video content analysis. In *Proc. of ICCIMA*, page 229, 2003.

[7] D. Zotkin, R. Duraiswami, and L. S. Davis. Multimodal 3-d tracking and event detection via the particle filter. In *Proc. of Event*, 2001.