MULTIMODAL INTERACTIONS FOR MULTIMEDIA CONTENT ANALYSIS

Thomas Martin[1,2], *Alain Boucher*[3,1], *Jean-Marc Ogier*[2]

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

L3i - Univ. of La Rochelle[2]
17042 La Rochelle cedex 1
La Rochelle, France

IFI[3]
ngo 42 Ta Quan Buu
Hanoi, Vietnam

thomas.martin@mica.edu.vn, alain.boucher@auf.org, jean-marc.ogier@univ-lr.fr

ABSTRACT

In this paper, we are presenting a model for multimodal content analysis. We are distinguishing between media and modality, which helps us to define and to characterize 3 inter-modal relations. Then we are applying this model for recorded course analysis for e-learning. Different useful relations between modalities are explained and detailed for this application. We are also describing on two other applications: telemonitoring and minute meetings. Then we compare the use of multimodality in these applications with existing inter-modal relations.

1. INTRODUCTION

Nowadays, as the available multimedia content grows every day, the need for automatic content analysis is becoming increasingly important. For example, information retrieval in broadcast news archives requires to index different medias available. Many projects currently focus on these research topics (content analysis, media enrichment...) but most of these works are focused on one sole media, and are unaware of other medias. Because information is not concentrated in one media but distributed among all the medias, such approaches are losing important parts of this information and ignore media interactions. Recently, many research works[1] have focused on the use of multiple modalities 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 this paper, we propose a first attempt to develop such a framework.
In the next section, we will give some definitions, followed by a review of the existing literature in multimodal analysis. Then we will present our model. After that, we will analyze some applications, 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 often a confusion in the literature between the concept of media and the concept of modality. In many papers, the authors use both words refering to the same concept. This does not seems to be exact as we can see the two different concepts in the context of content analysis. We propose 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 structure of the provided information (for example video text *vs* motion). One media can then be divided in multiple modalities, following two criteria: the semantic structuration of the information and the algorithms involved in the analysis process. While the concept of media is independant from the application, the concept of modality is application dependant.
As proposed in [2] we will use generic modalities listed in three main families. First, the audio family includes different modalities in terms of structuration like speech, music or sound. Second, we distinguish between still image and motion (video) in visual family. While both being acquired from a camera, motion contains time structuration and is more rich in term of content than still image. Third, the text family includes printed text and handwritten text.
This split of media into modalities can surely be discussed and different organization can be proposed. We will use this scheme through this paper using several examples taken from some applications to illustrate our choice. We insist on the fact that the information contained in each modality has a different structuration, regarding the algorithms that can be used, the difficulty for content extraction and for the semantic that can be given to it.
Once modality is defined, the next step is to define multimodality. In video indexing context, Snoek and Worring [1] have proposed to define multimodality from the author's point of view: it is "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 chan-

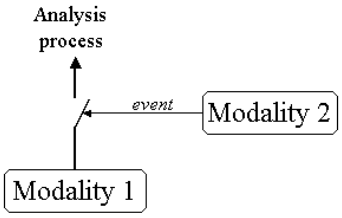


Fig. 1. In the trigger relation, the analysis process for one modality is activated with an event detected in another modality.

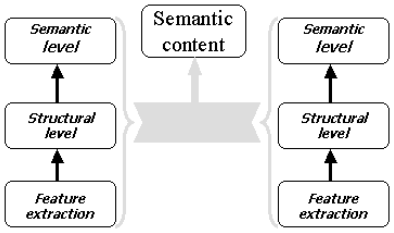


Fig. 2. The integration relation provides higher level information combining two or more modalities. The integration can be done at different levels.

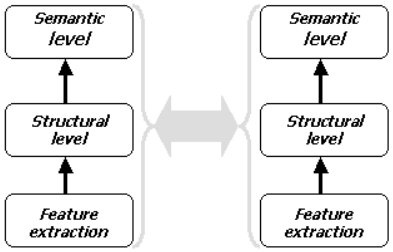


Fig. 3. The collaboration relation improves the analysis process for one modality using the results another one. This relation can be bidirectionnal.

nels". The inter-modal relation is then located at a high level using semantic. On the contrary, in the context of speech recognition, Zhi *et al.* [3] have implemented the multimodal integration just after the feature extraction phase and an alignment step. In this case, multimodal integration takes place at low level. Both these definitions are incomplete. Furthermore, several multimodal applications found in the literature use two modalities, audio and video, and the multimodal part of these application is often limited to a fusion step. Examples of such works include applications for video indexing such as [4] where a high level fusion step is processed after speaker segmentation in audio and shot detection in video. Shao *et al.*[5] process a multimodal summarizing of musical video using both audio and video contents. In the same domain, Zhu *et al.*[6] is performing video text extraction and lyrics structure analysis in karaoke contents using multimodal approaches. Song *et al.*[7] is recognizing emotions using a fusion step just after feature extraction in audio and video. Zhu and Zhou [8] are combining audio and video analysis for scene change detection. They have classified audio shots into semantic types and process shot detection in video They integrate then these results to have robust detection. Murai *et al.*[9] and Zhi *et al.* [3] are using facial analysis (video) to improve speech recognition (audio). [9] is detecting shots in video containing speech whereas [3] is combining lip movements and audio features to process speech recogntion. Zotkin *et al.*[10] is proposing a tracking method based on multiple cameras and a microphone array. Bigün *et al.*[11] is proposing a scheme for multimodal biometric authentication using three modalities: fingerprint, face and speech. Fusion is processed after individual modality recognition.

We propose a more general definition for multimodality as an interaction process between two or more modalities. This process is based on an inter-modal relation. We have identified three different types of inter-modal relations [2]: trigger, integration and collaboration. The trigger relation (see *fig.* 1) is the most simple relation: an event detected in one modality activates an analysis process to start in another modality. The integration relation (see *fig.* 2) 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 (fusion or others) of their results. Look at [1] for a review of existing works using widely the integration relation for the application of multimodal video indexing. The third relation is collaboration, and (see *fig.* 3). it is the strongest multimodal relation, consisting in a close interaction of two modalities during the analysis process itself. 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 the production of tools and techniques for multimedia documents oriented for e-learning. The complete course of a professor is recorded in live. Furthermore, textual sources such as course slides may be available. The recorded material from live courses is analyzed and used to produce interactive e-courses. This can be seen as an application of video analysis to produce rich media content. 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, text or image analysis. The main aim of this project consists in pro-

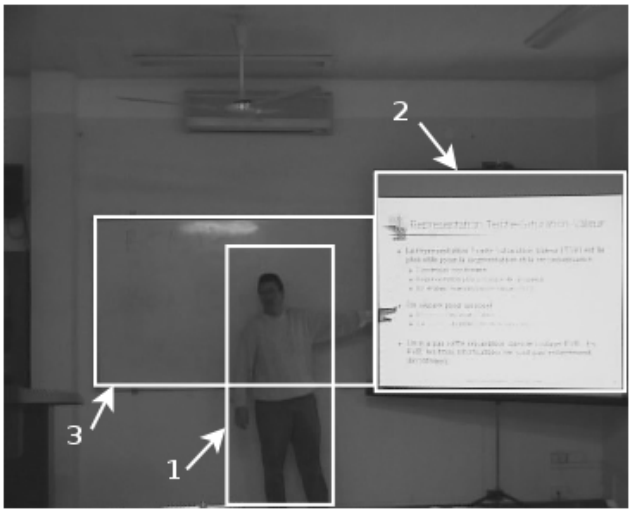
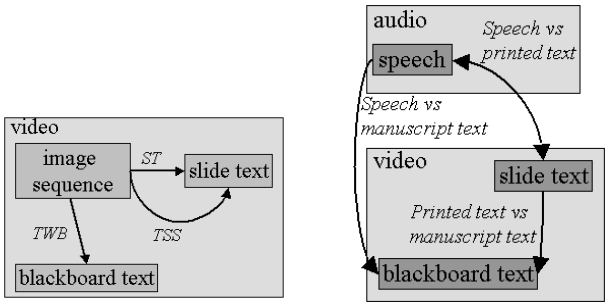


Fig. 4. Frame extracted from a recorded course. White shapes highlight identified actors of the application: the teacher (1), the screen (2) and the whiteboard (3).



(a) (b)

Fig. 5. (a) Scenario extraction. The event detection in the *image sequence* triggers differents analysis process in both *printed text* and *manuscript text*. Events involved are "Teacher Points at Screen" (TPS), "Teacher Writes on Whiteboard" (TWB) and "Slide Transition" (ST). (b) Content indexing. Three modalities are callaborating in the MARVEL application: *speech*, *printed text* and *manuscript text*.

viding semi-automatic tools to produce e-learning courses from recorded live normal courses.

In this project, three different medias are available: audio, video and lecture material (essentially the slides). Following the model proposed in section 2, we have identified five different modalities: *printed text* which contains the text of the slides and, if available, from other external textual sources. This modality is present in both video and lecture material media; *handwritten text* which represents the text written on the whiteboard; *graphics* which include all the graphics and images present in the slides. *motion* which contains the motion content of the video media; *speech* which contains the teacher's explanations. To simplify the explanations in this paper, we will not take into account the *graphic* modality and we consider only the textual parts of the slides. We are making a difference between *handwritten text* and *printed text* for two reasons. First, as presented in section 2, the nature of both modalities is different (handwritten text *vs* printed text). The second reason is specific to this application: the two modalities do not contain the same data. Even if the contents of both modalities are related to the course, one (*printed text* is more structured than the other. The *printed text* modality is avalaible in two different medias: video and text. It is a good example to illustrate our distinction between media and modality (section 2). Even if it is available into 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 (see *fig.* 4) is analyzed to extract the course scenario (explaining the current slide, writing on whiteboard, 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 whiteboard are detected. Detection of slide changes or new writing on the whiteboard are events that will be used. The content indexing of available media has to be done using the speech given by the teacher, the printed text on the slides and the handwritten text on the whiteboard. 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), 3 trigger relations (see *fig.* 5(a)) are involved. These relations are directly related to the actors who interact in a course: teacher, whiteboard and screen. The trigger source is the *motion* modality. First, the "slide transition" event triggers the *printed text* detection and recognition. Second, the "teacher points at screen" event triggers the point of interest search. Third, similar to the first, the "teacher writes on whiteboard" event triggers the *handwritten text* recognition process.

The second part of the application (content indexing) contains most of the inter-modal relations (see *fig.* 5(b)). 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, *motion-speech* interaction can be also useful [9, 3]. Recognition of *handwritten text* is a difficult task, especially in video. We propose to help recognition of *handwritten text*

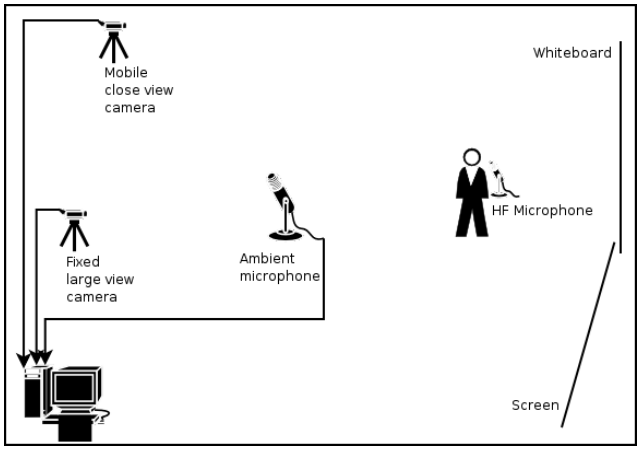


Fig. 6. The teacher gives the lecture. Two cameras and two microphones record it.

using both *speech* and *printed text* modalities. Both relations, *speech-handwritten text* and *speech-printed text*, are bimodal and unidirectional.

4. EXPERIMENTS

Our model has to be validated through experiments on real material. The experiment protocol comprises two steps. Firstly, the recording of a course according to a beforehand given protocol; secondly, the analysis through several experiments on collected data.

We have defined a protocol to control course recording. Two cameras are used in order to have both a large view on the scene and a close view on the screen (slides) and the whiteboard zones. The complete acquisition schema is presented in *fig.* 6. The first camera records a fixed large view of the lecture and give access to the general course of the lecture such as teacher behavior, while the second records a close view of both the screen (see *fig.* 7) and the whiteboard (see *fig.* 8) alternatively, providing lecture content with sufficient quality for further processing.

With regard to the audio part, we use a high-frequency microphone for recording the teacher as well as an ambient microphone for the students (see *fig.* 6).
As we just want to have video captures of the screen and the whiteboard, an efficient real time tracking algorithm is not necessary to control the camera. We rather decided to move the camera focus from one zone to the other following two simple rules. Firstly, the camera switches from the screen to the board if the teacher writes or points something on the whiteboard. Secondly, the camera switches from the whiteboard to the screen if a slide transition occurs or if the teacher points something at the screen. These events can be detected in the large view using relatively simple image

Fig. 7. Screen view extracted from mobile camera record. A close view of the screen is needed to match it with lecture material and also try to identify parts of the slide the teacher is pointing

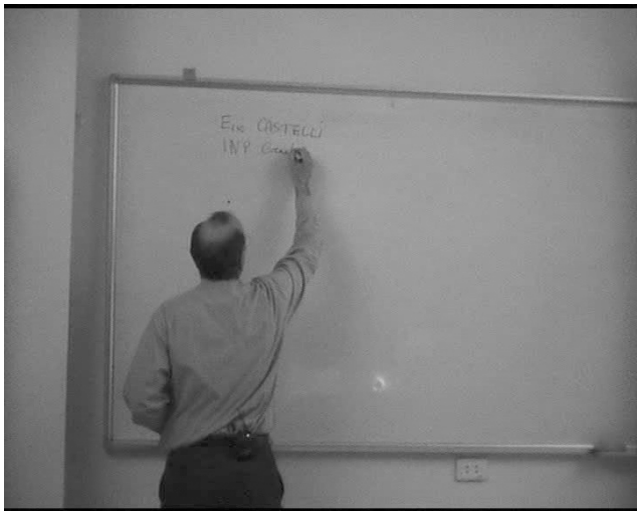


Fig. 8. Whiteboard view extracted from mobile camera record. A close view of the whiteboard is usefull for further handwritten text recognition.

processing techniques.
The validation of the model requires a representative set of situations occuring in real lectures. Various situations of interest have been defined and are recorded such as:

- the teacher is writing on the whiteboard;
- the teacher is pointing something at the whiteboard or on the screen;
- the teacher is moving across the classroom;
- the teacher is using gestures to support the speech;
- the teacher is reading word by word the shown slide;
- the teacher is explaining orally the content of the slide without using the same structure (word ordering);
- teacher-student interactions (such as question/answer session;
- the use of speech emphasis to stress some words;
- the use of demonstrative expressions to indicate a figure or a zone on the screen or the whiteboard;
- the use of the slide words in the oral speech;
- the use of synonymous words than the slide words in the oral speech.

From all the collected data, we are currently analysing different interactions in order to validate our model.
We are currently mainly focusing on the *printed text-speech* relation. After a recognition step, we are performing multiple experiments such as:

- temporally link the words in both modalities
- follow the speech on the slide text, like a karaoke.

This has to be done to measure the relation between *printed text* and *speech*. As second step, we will attempt to introduce the result of the *printed text* recognition in the speech recognition process and reciprocally. As both recognition processes make use of language models, we plan to achieve the collaboration between both modalities at this level.
The validation of the trigger relation will be done by implementing the tracking method presented above to switch between the screen and the whiteboard with the close camera. In a further step, other interactions will be studied, such as focusing on the words pointed by the teacher for the language model (for speech recognition).

5. OTHER APPLICATIONS

We plan to apply our model to a few other applications. One application, like MARVEL, cannot contain all possible inter-modal interactions. Then we are interested at two other different applications. First, medical telemonitoring aims, by transforming home in a smart environment, to allow elderly or ill people to keep autonomy in their life, thus to live at home. A fast detection of emergency situations allows a fast intervention of medical staff. Due to the variety of sensors – video cameras, microphones, localization sensors, specific medical equipments (such as electrocardiogram) in such application – used in a smart environment, the telemonitoring application is multimodal: *Motion* contains all the video information; *Sound* contains all the audio information; separation between speech and other audio information is useless in this application. *Sensors* modality gather the available signals (motion sensors, ECG,...). These three modalities are respectively parts of the video, the audio and the signal media. Due to the nature of this application, we are paying attention to only one person, the monitored person. This person is evolving in a smart environment, going from one room to another and doing tasks. Sometimes, the person can be in a crisis situation. To detect this situation, the telemonitoring system needs to detect abnormal behavior. However, detection is not sufficient: a behavior understanding process is necessary to identify real emergency situations through false alarms. Audio-video monitoring can be used for behavior understanding, by scenario extraction. This scenario extraction is the result of an integration relation between both *sound* and *motion* modalities. This step is preceded by a bidirectionnal collaboration relation between these two modalities. In fact, video elements can help sound recognition and results of this sound recognition can also help video recognition. The other available sensors can also be involved in audio-video scenario extraction. For example, localization sensors can provide usefull information for audio-video recognition. The second interesting application is for minute meetings, which consist in producing automatic multimedia minutes of meetings, using audio-video records. This application is similar to MARVEL project (for e-learning, see section 3). The main difference relies in the number of people. In MARVEL, the focus in only on the teacher, while for minute meetings, every person appearing in the audio-video streams is important. Modality relations are the same than in the MARVEL application, except for one more relation added between *speech* and *motion*. As one person is speaking, we can detect lip movements on video. Then, lip movements can help speaker segmentation in the speech modality. Generally, a meeting schedule is available, more or less detailed. This program is directly related to the audio-video content recorded during the meeting, and represent a first draft of

relation	application		
	MARVEL	MT	MM
speech/printed text	ci		ci
speech/handwritten text	ci		
sound/motion		ci	
speech/motion			c
printed text/motion	t		ct
handwritten text/motion	t		
sensors/sound		c	
sensors/motion		c	

Fig. 9. This table summarizes the different inter-modal relations identified in each application. MT stands for "medical telemonitoring" and MM for " minute meeting"; 'c', 'i' and 't' stand respectively for collaboration, integration and trigger relations.

scenario. The course of the meeting follows this schedule more or less accurately.

6. CONCLUSION

In this paper, we have analyzed multimodality and identified three different types of relations between the modalities. The first relation is trigger. This relation needs synchronized modalities. The second relation is integration. This relation 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. This relation cannot be used between all the modalities but is really useful in case of modalities with different structurations. 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. Another point regarding the relations between modalities is that they are not exclusive. There can be a collaboration relation between two modalities followed by an integration relation. Collaboration will improve content extraction from the separate modalities whereas integration will process a fusion of these results. For example, in the e-learning application, speech recognition results will be used to improve the text recognition process and, after that, results of both processes will be integrated for indexing purpose. We have applied this model to three different applications: e-learning, medical telemonitoring and minute meeting (see *fig.* 9). Relations are application dependant but some relations are common to several applications. We are currently experimenting these relations on real material to validate and improve our model. We particularly focus on collaboration relations to show the utility of this relation.

7. REFERENCES

[1] C. G. M. Snoek and M. Worring, "Multimodal video indexing: a review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[2] T. Martin, A. Boucher, and J.-M. Ogier, "Multimodal analysis of recorded video for e-learning," in *Proc. of the 13th ACM Multimedia Conference*. 2005, pp. 1043–1044, ACM Press.

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

[4] S. Tsekeridou and I. Pitas, "Audio-visual content analysis for content-based video indexing," in *Proc. of ICMCS*. 1999, vol. 1, pp. 667–672, IEEE Computer Society.

[5] X. Shao, C. Xu, and M. S. Kankanhalli, "Automatically generating summaries for musical video.," in *Proc. of ICIP*, 2003, vol. 2, pp. 547–550.

[6] Y. Zhu, K. Chen, and Q. Sun, "Multimodal content-based structure analysis of karaoke music.," in *Proc. of the 13th ACM Multimedia Conference*. 2005, pp. 638–647, ACM Press.

[7] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition - a new approach," in *Proc. of CVPR*. 2004, vol. 2, pp. 1020–1025, IEEE Computer Society.

[8] Y. Zhu and D. Zhou, "Scene change detection based on audio and video content analysis," in *Proc. of ICCIMA*, 2003, p. 229.

[9] K. Murai, K. Kumatani, and S. Nakamura, "Speech detection by facial image for multimodal speech recognition," in *Proc. of ICME*. 2001, p. 149, IEEE Computer Society.

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

[11] 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*. 2003, pp. 2–11, IEEE Computer Society.