A REVIEW ON MULTIMODAL VIDEO INDEXING

Cees G.M. Snoek and Marcel Worring

Intelligent Sensory Information Systems, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
{cgmsnoek, worring}@science.uva.nl

ABSTRACT

Efficient and effective handling of video documents depends on the availability of indexes. Manual indexing is unfeasible for large video collections. Efficient, single modality based, video indexing methods have appeared in literature. Effective indexing, however, requires a multimodal approach in which either the most appropriate modality is selected or the different modalities are used in collaborative fashion. In this paper we present a framework for multimodal video indexing, which views a video document from the perspective of its author. The framework serves as a blueprint for a generic and flexible multimodal video indexing system, and generalizes different state-of-the-art video indexing methods. It furthermore forms the basis for categorizing these different methods.

1. INTRODUCTION

For browsing, searching, and manipulating video documents, an index describing the video content is required. It forms the crux for applications like digital libraries storing multimedia data or filtering systems which automatically identify relevant video documents based on a user profile. To cater for these diverse applications, the indexes should be rich and as complete as possible.

Until now, construction of an index is mostly carried out by documentalists who manually assign a limited number of keywords to the video content. The specialist nature of the work makes manual indexing of video documents an expensive and time consuming task. Therefore, automatic video indexing methods are necessary.

Most solutions to video indexing use a unimodal approach, i.e. only the visual, auditory, or textual modality is used. Instead of using only one modality, multimodal video indexing strives to automatically classify (pieces of) a video document based on multimodal analysis. In this paper we put forward a unifying framework for multimodal video indexing. In contrast to others, who view a video document from a data perspective, we view a video document from the perspective of its author. This is important, as the ultimate goal of video indexing is to capture the intentions of the author.

The framework is defined in section 2. This framework forms the basis for structuring the discussion on video document segmentation in section 3. In section 4 the role of multimodal analysis is discussed, and an overview is given of the index types that can be distinguished. Finally, in section 5 we end with the conclusions and a perspective on future research.

2. AN AUTHOR'S PERSPECTIVE

An author of a video document uses visual, auditory, and textual channels to express his or her ideas. Hence, the content of a video is intrinsically multimodal. Let us make this more precise. In [12] multimodality is viewed from the system domain and is defined as "the capacity of a system to communicate with a user along different types of communication channels and to extract and convey meaning automatically". We extend this definition from the system domain to the video domain, by using an authors perspective as:

Definition 1 (Multimodality) *The capacity of an author of the video document to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels.*

We consider the following three information channels or modalities, within a video document:

- *Visual modality*: contains the *mise-en-scène*, i.e. everything that can be seen in the video document;
- *Auditory modality*: contains the speech, music, and environmental sounds that can be heard in the video document;
- *Textual modality*: contains textual resources that describe the content of the video document;

For each of those modalities, definition 1 naturally leads to a semantic perspective, a content perspective, and a layout perspective.

The first perspective expresses the intended semantic meaning of the author, and defines segments on four different levels within a semantic index hierarchy. The first two levels are related to the video document as a whole, and define segments based on consistent appearance of layout or content elements. We define:

- *Genre*: set of video documents sharing similar style;
- *Sub-genre*: a subset of a genre where the video documents share similar content;

The next level of our semantic index hierarchy is related to parts of the content, and is defined as:

- *Logical units*: a continuous part of a video document's content consisting of a set of named events or other logical units which together have a meaning;

Where named event is defined as:

- *Named events*: short segments which can be assigned a meaning that doesn't change in time;

The content perspective relates segments to elements that an author uses to create a video document. The following elements can be distinguished [3]:

- *Setting*: time and place in which the video's story takes place;
- *Objects*: noticeable static or dynamic entities in the video document;
- *People*: human beings appearing in the video document;

Finally, the layout perspective considers the syntactic structure an author uses for the video document. In essence, the syntactic structure for each modality is a temporal sequence of *fundamental units*, which in itself do not have a temporal dimension. Upon the fundamental units an aggregation is imposed, which is an artifact from creation. We refer to this aggregated fundamental units as *sensor shots*, defined as a continuous sequence of fundamental units resulting from an uninterrupted sensor recording. For the visual and auditory modality this leads to *camera shots* and *microphone shots* which are a result of an uninterrupted recording of a camera or microphone. For text, sensor recordings do not exist. In writing, uninterrupted textual expressions can be exposed on different granularity levels, e.g. word level or sentence level, therefore we define *text shots* as an uninterrupted textual expression.

An author of the video document is also responsible for concatenating the different sensor shots into a coherent structured document by using *transition edits*. For the visual modality abrupt cuts, or gradual transitions, like wipes, fades, or dissolves can be selected. For the auditory transitions an author can have a smooth transition using music, or an abrupt change by using silence [3]. To indicate a transition in the textual modality, e.g. closed captions, an author typically uses ">>>", or different colors. They can be viewed as corresponding to abrupt cuts as their use is only to separate shots, not to connect them smoothly. The final component of the layout are the optional visual or auditory *special effects*, used to enhance the impact of the modality, or to add meaning.

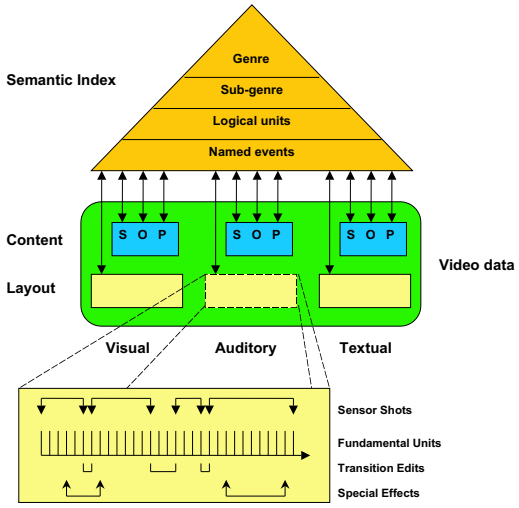

Figure 1: *A framework for multimodal video indexing. The letters S, O, P stand for setting, objects, and people. An example layout of the auditory modality is highlighted.*

Based on the discussion in this section we come to a unifying multimodal video indexing framework based on the perspective of an author. This framework is visualized in figure 1. It forms the basis for our discussion on video document segmentation techniques in the next section.

3. VIDEO DOCUMENT SEGMENTATION

For analysis purposes the process of authoring should be reversed. To that end, first a segmentation should be performed that decomposes a video document into its layout and content elements.

3.1. Layout reconstruction

For reconstruction of the visual layout, several techniques already exist to segment a video document on the camera shot level. For an extensive overview of different cut detection methods and detection of transition edits we refer to the survey of Brunelli in [4] and the references therein.

Detection of abrupt cuts in the auditory layout can be achieved by detection of silences and transition points, i.e. locations where the category of the underlying signal changes. In literature different methods are proposed for their detection, for example [8].

Detection of text shots can be achieved in different ways, if we are only interested in single words we can use the occurrence of white space as the main clue. When more context is taken into account one can reconstruct sentences from the textual layout by detection of periods [10]. Transitions are typically found by searching for predefined patterns.

3.2. Content segmentation

For the reconstruction of content elements different approaches can be followed. People can be detected by means of their faces or other body parts, speech, and appearance of names. Specific objects can be detected by means of specialized visual detectors, motion, sounds, and appearance in the textual modality. Setting can be detected by visual detectors, setting sounds, and geographic information. For a detailed review of the different detection algorithms we refer to [15].

4. MULTIMODAL ANALYSIS

After reconstruction of the layout and content elements through video segmentation, the next step in the inverse analysis process is integrated analysis of the layout and content to extract the semantic index, see figure 2.

Before integrating the different layout and content elements, it is useful to apply modality conversion of some elements into more appropriate form. For analysis, conversion of elements of visual and auditory modalities to text is most appropriate. A typical component we want to convert from the visual modality is overlayed text, see e.g. [9]. From the auditory modality one typically wants to convert the uttered speech into transcripts [7].

To achieve the goal of multimodal integration, several approaches can be followed. We categorize those approaches by their distinctive properties with respect to the processing cycle, the segmentation results, and the classification method used. The processing cycle of the integration method can be iterated, allowing for incremental use of context, or non-iterated. The segmentation results can be exploited by using the different modalities in a symmetric, i.e. simultaneous, or asymmetric, i.e. ordered, fashion. Finally, for the classification one can choose between a statistical, i.e. data-driven, or knowledge-based approach. An overview of the different integration methods found in literature is in table 1.

Most integration methods reported are symmetric and non-iterated. Some follow a knowledge-based approach for classification of the data into classes of the semantic index hierarchy [5, 13]. Many methods in literature follow a statistical approach [1, 6, 7, 11, 14]. An example of a symmetric, non-iterated statistical integration method is the Name-It system presented in [14]. The system associates detected faces and names, by calculating a co-occurrence factor that combines the analysis results of face detection and recognition, name extraction, and caption recognition.

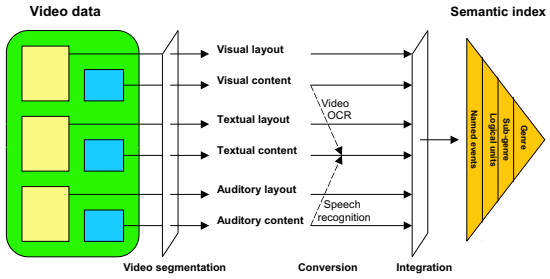

Figure 2: *Multimodal video document analysis framework.*

Hidden Markov Models (HMM) are frequently used as a statistical classification method for multimodal integration [1, 6]. A clear advantage of this framework is that it is not only capable to integrate multimodal features, but is also capable to include sequential features. Moreover, an HMM can also be used as a classifier combination method.

When modalities are independent, they can easily be included in a product HMM. In [6] such a classifier is used to train two modalities separately, which are then combined symmetrically, by computing the product of the observation probabilities. It is shown that this results in significant improvement over a unimodal approach.

In contrast to the product HMM method, a neural network-based approach doesn't assume features are independent. Another approach presented in [6], trains an HMM for each modality and category. A three layer perceptron is then used to combine the outputs from each HMM in a symmetric and non-iterated fashion.

Another advanced statistical classifier for multimodal integration was recently proposed in [11]. A probabilistic framework for semantic indexing of video documents based on so called multijects and multinets is presented. The multijects model content elements which are integrated

	Symmetric	Statistical	Iterated
[1]	✓	✓	
[2]			✓
[5]	✓		
[6]	✓	✓	
[6]		✓	
[7]	✓	✓	
[11]	✓	✓	✓
[13]	✓		
[14]	✓	✓	

Table 1: *An overview of different integration methods.*

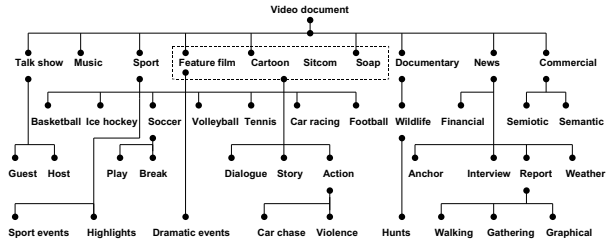

Figure 3: *Semantic index hierarchy with instances as found in literature. From top to bottom instances from genre, subgenre, logical units, and named events. The dashed box is used to group similar nodes.*

in the multinets to model the relations between objects, allowing for symmetric use of modalities. For the integration in the multinet the authors propose a Bayesian belief network. Significant improvements of detection performance is demonstrated. Moreover, the framework supports detection based on iteration.

In contrast to the above symmetric methods, an asymmetric approach is presented in [6]. A two-stage HMM is proposed which first separates the input video document into three broad categories based on the auditory modality, in the second stage another HMM is used to split those categories based on the visual modality. A drawback of this method is its application dependency, which may result in less effectiveness in other classification tasks.

An asymmetric knowledge-based integration method, supporting iteration, was proposed in [2]. First, the visual and textual modality are combined to generate semantic index results. Those form the input for a post-processing stage that uses those indexes to search the visual modality for the specific time of occurrence of the semantic event.

The methodologies described in this section have been applied to extract a variety of the different video indexes described in section 2. In [6] for example, (sub)genres like news reports, commercials, basketball, and football games are distinguished. Logical units such as dialogues are detected in [1, 13]. Named sport events are detected in [2]. Figure 3 presents an overview of all semantic indexes that we found in literature (covering a total of hundred references). For an extensive overview of all those methods, including the low level information from which they are derived, we again refer to [15].

5. CONCLUSION

Viewing a video document from the perspective of its author, enabled us to present a framework for multimodal video indexing. This framework forms the blueprint for a generic and flexible multimodal video indexing system. Moreover it allows for generalization of different state-of-the art video indexing methods. Our future research efforts will be geared towards the development of a multimodal video indexing system, based on the framework presented.

Acknowledgements

This research is sponsored by the ICES/KIS Multimedia Information Analysis project and TNO-TPD.

6. REFERENCES

[1] A. Alatan, A. Akansu, and W. Wolf. Multi-modal dialogue scene detection using hidden markov models for content-based multimedia indexing. *Multimedia Tools and Applications*, 14(2):137–151, 2001.

[2] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*, 4(1):68–75, 2002.

[3] J. Boggs and D. Petrie. *The art of watching films*. Mayfield Publishing Co., Mountain View, CA, 2000.

[4] R. Brunelli, O. Mich, and C. Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78–112, 1999.

[5] S. Fischer, R. Lienhart, and W. Effelsberg. Automatic recognition of film genres. In *ACM Multimedia 1995*, pages 295–304, 1995.

[6] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. Wong. Integration of multimodal features for video scene classification based on HMM. In *IEEE Workshop on Multimedia Signal Processing*, 1999.

[7] P. Jang and A. Hauptmann. Learning to recognize speech by watching television. *IEEE Intelligent Systems*, 14(5):51–58, 1999.

[8] D. Li, I. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.

[9] H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video. *IEEE Trans. on Image Processing*, 9(1):147–156, 2000.

[10] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[11] M. Naphade and T. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. on Multimedia*, 3(1):141–151, 2001.

[12] L. Nigay and J. Coutaz. A design space for multimodal systems: concurrent processing and data fusion. In *INTERCHI'93 Proceedings*, 1993.

[13] S. Pfeiffer, R. Lienhart, and W. Effelsberg. Scene determination based on video and audio features. *Multimedia Tools and Applications*, 15(1):59–81, 2001.

[14] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35, 1999.

[15] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. Technical Report 20, Intelligent Sensory Information Systems, University of Amsterdam, 2001.