Content-based Video Retrieval

M. Petković
Centre for Telematics and Information Technology, University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
Email: milan@cs.utwente.nl

1. Introduction

With technology advances in multimedia, digital TV and information highways, a large amount of video data is now publicly available. However, without appropriate search technique all these data are nearly not usable. Users are not satisfied with the video retrieval systems that provide analogue VCR functionality. They want to query the content instead of raw video data. For example, a user will ask for specific part of video, which contain some semantic information. Content-based search and retrieval of these data becomes a challenging and important problem. Therefore, the need for tools that can manipulate the video content in the same way as traditional databases manage numeric and textual data is significant.

This extended abstract presents our approach for content-based video retrieval. It is organised as follows. In the next section, we give an overview of related work. The third section describes our approach with emphasis on the video modelling as one of the most critical processes in video retrieval. The fourth section draws conclusion.

2. State of the art

Video content can been grouped into two types: low-level visual content and semantic content. Low-level visual content is characterised by visual features such as colour, shapes, textures etc. On the other hand, semantic content contains high-level concepts such as objects and events. The semantic content can be presented through many different visual presentations. The main distinction between these two types of content is different requirements for extracting each of these contents. The process of extracting the semantic content is more complex, because it requires domain knowledge or user interaction, while extraction of visual features is usually domain independent.

Extensive research efforts have been made with regard to the retrieval of video and image data based on their visual content such as colour distribution, texture and shape. These approaches fall into two categories: query by example and visual sketches. Both of these are based on similarity measurement. Examples include IBM's Query by Image Content (QBIC) [1], VisualSEEk [2], Photobook [3], Blobworld [4], as well as Virage video engine [5], CueVideo [6] and VideoQ [7] in the field of video. Query by example approaches are suitable if a user has a similar image at hand, but they would not perform well if the image is taken from a different angle or has a different scale. The naive user is interested in querying at the semantic level rather then having to use features to describe his concepts. Sometimes it is difficult to express concepts by sketching. Nevertheless, good match in terms of the feature metrics may yield poor results (multiple domain recall, e.g. a query for 60% of green and 40% of blue may return an image of a grass and sky, a green board on a blue wall or a blue car parked in front of a park, as well as many others).

Modelling the semantic content is more difficult then modelling the low-level visual content of a video. At the physical level video is a temporal sequence of pixel regions without direct relation to its semantic content. Therefore, it is very difficult to explore semantic content from the raw video data. In addition to that, if we consider multiple semantic meaning such as metaphorical, associative, hidden or suppressed meaning, which the same video content may have, we make a problem more complex.

The simplest way to model the video content is by using free text manual annotation. Some approaches [8, 9] introduce additional video entities, such as objects and events, as well as their relations, that should be annotated, because they are subjects of interests in video. One of the major limitations of these approaches is that search process is based mainly on the attribute information, which are associated by video segment manually by human or (semi)automatically in the process of annotation. These approaches are very limited in terms of spatial relations among sub-frame entities. Spatio-temporal data models overcome these limitations by associating the concept of video object to the sub-

frame region that conveys useful information, and by defining events that include spatio-temporal relations among objects. Modelling of these high-level concepts gives the possibility to describe objects in space and time and capture movements of objects. As humans think in term of events and remember different events and objects after watching video, these high-level concepts are the most important cues in content-based video retrieval. A few attempts to include these high-level concepts into video model are made in [10, 11].

The distinction, we made regarding modelling the video content, makes clear two important things. On the one hand, feature-based models use automatically extracted features to represent the content of a video, but they do not provide semantics that describes high-level concepts of video, such as objects and events. On the other hand semantic models usually use free text/attribute/keywords annotation to represent the high-level concepts of the video content that results in many lacks. The main one is that manual annotation is tedious, subjective and time consuming. Obviously, an integrated approach, that will provide automatic mapping from features to high-level concepts, is the challenging solution.

3. The third way: Concept inferencing

In order to overcome the problem of mapping from features to high level concepts we propose a layered video data model that has the following structure. The raw video data layer is at the bottom. This layer consists of a sequence of frames, as well as some video attributes, such as compression format, frame rate, number of bits per pixel, colour model, duration, etc. The next layer is the feature layer that consists of domain-independent features that can be automatically extracted from raw data. Examples are shapes, textures, colour histogram, as well as dynamic features characterising frame sequences, such as temporality, motion, etc. The concept layer is on the top. It consists of logical concepts that are subject of interest of users or applications. Automatic mapping from raw video data layer to feature layer is already achieved, but automatic mapping from feature to concept layer is still a challenging problem. We simplify this problem by dividing the concept layer into object and event layer.

We define a region, as a contiguous set of pixels that is homogeneous in the features such as texture, colour, shape and motion. As we already mentioned a region could be automatically extracted and tracked. Then, we define a video object as a collection of video regions, which have been grouped together under some criteria defined by the domain knowledge. As we can see in the literature [12, 13, 14] automatic detection of video objects (sub-frame entities) in a known domain are feasible. For this purpose, we proposed an object grammar that consists of rules for object extractions. A simplified example of an object rule in the soccer domain could be “if the shape of a region is round, and the colour is white, and it is moving, that object is a ball”. For the second part of the problem - automatic mapping from this object layer to event layer, we propose the event grammar that consists of rules for describing event types in terms of spatio-temporal object interactions. The event types can be primitive and compound. The primitive event type could be described using object types, spatio-temporal and real-world relations among object types, as well as audio segment types and temporal relations among them. Nevertheless, predefined event types, their temporal relations, as well as real-world and spatial relations among their objects can together be a part of compound event type description. For example, in the soccer domain, if the ball object type is inside the goalpost object type for a while and this is followed by very loud shouting and a long whistle, that might indicate that someone has scored a goal, which should be recognised as a goal event.

The main advantage of the proposed layered video data model is automatic mapping from features to concepts. This approach bridges the gap between domain independent features, such as colour histograms, shapes, textures and domain dependent high-level concepts such as objects and events. The proposed event grammar formalises the description of spatio-temporal object interactions. However, metaphorical, associative, hidden or suppressed meaning of the video content is not covered by this grammar. Although we proposed traditional annotation approach for this kind of content, this could be a direction of our future work.

4. Conclusion

We proposed a layered video data model that integrates audio and video primitives. Four layers structure of our video model makes easier a process of translating raw video data into efficient internal representation that captures video semantics. Our model allows dynamic (ad-hoc) definition of video

objects and events that can be used in process of content-based retrieval. This enables a user to dynamically define a new event, insert a new index for it and query the database, all by one query. Easy description of video content is supported by robust object and event grammars that can be used for specifying even very complex objects and events. With the proposed event grammar, we try to go one step further in video content description. We put effort into formalising events as descriptions of objects' (inter)actions. This results in easier capturing of high-level concepts of video content and queries are closer to user way of thinking (users' cognitive maps of a video). The corresponding query language enables users to specify wide range of queries using audio, video and image media types. The layered model structure allows dynamic logical segmentation of video data during querying.

A prototype of video database system based on proposed model and query language is under development. We use MOA object algebra [15] developed at the University of Twente and MONET database management system [16] developed at CWI and University of Amsterdam as implementation platform.

References

[1] M. Flinker, H. Samhey, W. Niblack et al., "Query by Image and Video Content: The QBIC System", IEEE Computer, 28, (Sept. 1995), pp. 23-32.

[2] J. R. Smith, S-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System", ACM Multimedia Conference, Boston, MA, November 1996.

[3] A. Pentland, R. W. Picard, S. Sclaroff, "Photobook: Content-Based Manipulation of Image Databases", Int. J. Computer Vision, 18 (3), pp. 233-254.

[4] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, J. Malik, "Blobworld: A System for Region-Based Image Indexing and Retrieval", Third Int. Conf. On Visual Information and Information Systems, Amsterdam, 1999, pp. 509-516.

[5] A. Hampapur, A. Gupta, B. Horowitz, C-F. Shu, C. Fuller, J. Bach, M. Gorkani, R. Jain, "Virage Video Engine", SPIE Vol. 3022, 1997.

[6] D. Ponceleon, S. Srinivasan, A. Amir, D. Petkovic, D. Diklic, "Key to Effective Video Retrieval: Effevtive Cataloging and Browsing", ACM Multimedia, '98, pp. 99-107.

[7] S-F. Chang, W. Chen, H. Meng, H. Sundaram, D. Zhong, "A Fully Automated Content Based Video Search Engine Supporting Spatio-Temporal Queries", IEEE Transaction on Circuits and Systems for Video Tecnology, Vol. 8, No. 5, Sept., 1998.

[8] S. Adali, K. S. Candan, S-S. Chen, K. Erol, V. S. Subrahmanian, "Advanced Video Information System: Data Structure and Query Processing", Multimedia System Vol. 4, No. 4, Aug. 1996, pp. 172-86.

[9] C. Decleir, M-S. Hacid, J. Kouloumdjian, "A Database Approach for Modelling and Querying Video data", LTCS-Report 99-03, 1999.

[10] H. Jiang, A. Elmagarmid, "Spatial and temporal content-based access to hypervideo databases" VLDB Journal, 1998, No. 7, pp. 226-238.

[11] J. Z. Li, M. T. Ozsu, D. Szafron, "Modeling of Video Spatial Relationships in an Object Database Management System", Proc. of Int. Workshop on Multi-media Database Management Systems, 1996, pp. 124-132.

[12] Y. Gong, L. T. Sin, C. H. Chuan, H-J. Zhang, M. Sakauchi, "Automatic Parsing of TV Soccer Programs", IEEE International Conference on Multimedia Computing and Systems, Washington D. C., 1995, pp. 167-174.

[13] S. Intille, A. Bobick, "Visual Tracking Using Closed-Worlds", M.I.T. Media Laboratory, Technical Report No. 294, Nov. 1994.

[14] G. P. Pingali, Y. Jean I. Carlbom, "LucentVision: A System for Enhanced Sports Viewing", Proc. of Visual'99, Amsterdam, 1999, pp. 689-696.

[15] P. Boncz, A.N. Wilschut, M.L. Kersten, "Flattering an object algebra to provide performance", Proceedings of the 14th IEEE International Conference on Data Engineering, Orlando, Florida, 1998, pp. 568-577.

[16] P. Boncz, M.L. Kersten, "Monet: An Impressionist Sketch of an Advanced Database System", Proceedings Basque International Workshop on Information Technology, San Sebastian, Spain, July 1995.