Content-based video retrieval

M. Petković
Centre for Telematics and Information Technology, University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
Phone: +31 53 4893725, Fax: +31 53 4894524
E-mail: milan@cs.utwente.nl

1. Introduction

Recent advances in multimedia technologies allow the capture and storage of video data with relatively inexpensive computers. Furthermore, the new possibilities offered by the information highways have made a large amount of video data publicly available. However, without appropriate search techniques all these data are hardly usable. Users are not satisfied with the video retrieval systems that provide analogue VCR functionality. They want to query the content instead of the raw video data. For example, a user analysing a soccer video will ask for specific events such as goals. Content-based search and retrieval of video 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 paper 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 architecture of a content-based video retrieval system is presented in the fourth section. The fifth section draws conclusion.

2. The state of the art overview

As we already mentioned, modelling the video content is one of the most important tasks in video retrieval. In the literature, video content is approached at different levels: raw data, low-level visual content and semantic content. The raw video data consists of elementary video units together with some general video attributes such as format, frame rate etc. Low-level visual content is characterised by visual features such as colour, shapes, textures etc. Semantic content contains high-level concepts such as objects and events. The semantic content can be presented through many different visual presentations using different sets of raw data. It is obvious that requirements for the extraction of these contents are different. The process of extracting the semantic content is the most complex, because it requires domain knowledge or user interaction, while extraction of visual features can be often done automatically and it 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 [1]. These approaches are mainly based on similarity measurement. Examples include VisualSEEk [2], Photobook [3], Blobworld [4], as well as Virage video engine [5], CueVideo [6] and VideoQ [7] in the field of video. The image retrieval systems allow a user to make queries based on visual image content – properties such as colour percentages, colour layout and textures occurring in the images usually by using instances of prior matches (query by example). Some of these systems use the spatial information and allow the user to make queries by sketching the layout of colour regions, or by providing the URL of a seed image. First approaches in video retrieval just added the functionality for segmentation and key frame extraction to existing image retrieval systems. After key-frame extraction, they just apply similarity measurement on them based on low-level features. This is not satisfactory because video is temporal media, so sequencing of individual frames creates new semantics which may not be present in any of the individual shots. Furthermore, choosing the key-frames is still a challenging problem.

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. 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 far 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 the problem even more complex.

The simplest way to model the video content is by using free text manual annotation. An example is 'stratification' approach [8] with a few extensions [9, 10]. Some other approaches [11, 12] 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.

Another way to model the video entities assumes using spatio-temporal relations. The concept of video object can be associated to the sub-frame region that conveys useful information, while spatio-temporal relations among these objects can be defined as events. Modelling of these high-level concepts (objects and events) 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 retrieval. Let's take as an example a soccer game, humans usually remember goals, interesting actions, red cards etc. A few attempts to include these high-level concepts into video model are made in [13, 14, 15].

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 annotation is tedious, subjective and time consuming. One of the major limitations of this approach is that search process is based mainly on the predefined attribute information, which are associated by video segments manually by human or (semi)automatically in the process of annotation. 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 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 the raw data. There are two types of video features: static features characterising a still image (frame), such as shapes, textures, colour histogram, etc., and dynamic features characterising frame sequences, such as temporality, motion, etc. The features are assigned to regions. We define a region as a contiguous set of pixels that is homogeneous in one or more features. As we can see in [16] a region could be automatically extracted and tracked. The concept layer is on the top. It consists of logical concepts that are subject of interest to users or applications. Automatic mapping from the raw video data layer to the feature layer is already achieved, but automatic mapping from the feature to the concept layer is still a challenging problem. We simplify this problem by dividing the concept layer into the object and event layer.

The object layer consists of entities (logical concepts) that can be assigned to one or more regions. We define a video object as a collection of video regions, which have been grouped together under some criteria defined by the domain knowledge. These objects should also satisfy some conditions such that they should be semantically consistent, representing one real-world object, and subject of interest to users or applications. Some examples of video objects are a specific player or the ball in a soccer game or a specific car in a car-race video. As we can see in the literature [17, 18, 19] automatic detection of video objects (sub-frame entities) in a known domain is feasible. For this purpose, we proposed the 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, the colour is white and it is moving, that object is a ball".

The event layer is the highest layer in our model. It consists of events, which describe object interactions in the spatio-temporal manner. For the automatic mapping from the object layer to the event layer, we propose the event grammar that consists of rules for describing event types. The event type could be described using object types, audio segment types, spatio-temporal and real-world relations. 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 that it provides a framework for 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 audio component is integrated in the model to provide additional information that can be critical to the perception and understanding of video content.

The formal definition of the proposed video model, as well as a few modelling examples can be found in [20].

4. The architecture of content-based video retrieval system

The proposed architecture for a content-based video retrieval system is shown in Fig. 1. The process of database population is shown with dashed lines, while querying is shown with solid lines. The raw video data is stored in the file system, while the storage server is used to store video content meta data and indexes. In the process of the database population, the features, objects, and events that are specified by system administrator, are extracted. Indexes and meta-data are put in the storage server and videos in the file system. Most queries are resolved directly in the storage server, but if the query comprises something that has not been already extracted the extractors do that dynamically.

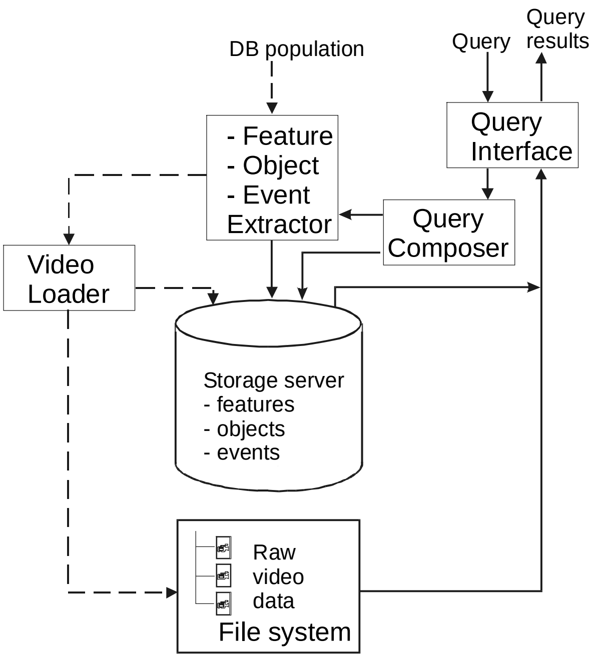

Fig. 1. The system architecture

The implementation platform for the storage server should be chosen very carefully. In addition to storage, it should support efficient management and homogeneous querying of features, objects and events. For example, the storage server should be capable to deal with distance functions in feature spaces to perform similarity measurements. In order to enable modelling and querying objects and events using the grammars, it should support a basic set of spatio-temporal relations like ones defined in [21, 22]. As far as temporal relations are concerned, point and interval data type should be supported to represent frames and frame sequences respectively. Each object and event has, as an attribute, a set of intervals, i.e. frame sequences, where it occurs. The basic relations of interval and point temporal algebra, the mapping between them, as well as operations on the interval data type, such as intersect and union, have to be defined.

Having these requirements in mind, we chose the Moa/Monet platform for the prototype of the video retrieval system that is currently under development. At the physical level, we use Monet [23] – an extensible parallel database kernel developed at the CWI and the University of Amsterdam. The advantages of Monet, that have influence on our decision to chose it for physical level, are its main memory query execution, extensibility and parallelism. At the logical level, we use the Moa object data model and algebra [24] developed at the University of Twente. The Moa data model is a structural object data model. It accepts all base types of the underlying physical storage system and allows their orthogonal combination using the structure primitives set, tuple and object. The chosen platform is extensible and we believe that our goal, the video database system that allows retrieval based on the video content, can be reached easier.

At the conceptual level, the video data model and a query language that is especially intended to support video retrieval are defined. This level provides easy to use object and event descriptions. A user can define his concepts at this level. They will be automatically translated to the Moa object algebra at the logical level and then again automatically to the Monet's algebra at the physical level.

The basic video model should be accompanied by a domain model. It means that fundamental classes, which deal with the video structure, as well as with meta content (video object types, event types, etc.), are defined. When the domain for the video retrieval is chosen, classes that model the domain itself have to be added. In this way, we keep a general approach and avoid the problem of modelling the whole world.

5. Conclusion

The proposed approach integrates feature-based and annotation-based retrieval approaches taking their best characteristics and avoiding their limitations and drawbacks. The main advantage of the model is that it provides a framework for automatic mapping from features to semantic concepts, integrating audio and video primitives. 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 four-layer structure of our video model guides the process of translating raw video data into efficient internal representation that captures video semantics. Easier description of video content is supported by the robust object and event grammars that can be used for specifying even very complex objects and events. The formalisation of an event as a spatio-temporal description of object interactions results in easier capturing of high-level concepts of video content, and allows queries that are closer to the user way of thinking (users' cognitive maps of a video).

The proposed model has formal foundation and layered structure that enables using of different techniques at different layers as well as combining different techniques at the same layer. This allows concurrent use of for example a MPEG-4 object extractor, other object extractors based on object recognition, and the object grammar at the object layer. The model also supports flexible video segmentation using high-level concepts. This allows a user to make different logical segmentation of the same raw data dynamically, building different hierarchies.

References

[1] P. Aigrain, H. Zhang, D. Petkovic, Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review, *Multimedia Tools and Applications*, Kluwer Academic Publishers, 3(3), 1996, 179-202.

[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 Technology, Vol. 8, No. 5, Sept., 1998.

[8] T. G. Aguierre Smith, G. Davenport, The Stratification System: A Design Environment for Random Access Video, *Workshop on Network and operating system*, La Jolla, CA, 1992.

[9] R. Weiss, A. Duda, D. K. Gifford, Content-based Access to Algebraic Video, *Int. Conf. on Multimedia Computing and Systems*, IEEE Press, 140-151.

[10] E. Oomoto, K. Tanaka, OVID: Design and Implementation of a Video-Object Database System, *IEEE Trans Knowl Data Eng*, 5(4), 1993, 629-643.

[11] 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.

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

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

[14] 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.

[15] A. Woudstra, D.D. Velthausz, H.J.G. de Poot, F. Moelaert El-Hadidy, W. Jonker, M.A.W. Houtsma, R.G. Heller, J.N.H. Heemskerk, Modelling and Retrieving Audiovisual Information - A Soccer Video Retrieval System -, *4th International Workshop on Multimedia Information Systems*; Istanbul, Turkey, September 1998.

[16] D. Zhong, S-F.Chang, Video Object Model and Segmentation for Content-Based Video Indexing, *IEEE Circuits and Systems*, Hong Kong 1997.

[17] 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.

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

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

[20] M. Petkovic, W. Jonker, "A Framework for Video Modelling", 18th IASTED Conference on Applied Informatics, Innsbruck, Austria, 2000.

[21] D. Papadias, Y. Theodoridis, T. Sellis, and M. Egenhofer , "Topological Relations in the World of Minimum Bounding Rectangles: A Study with R-Trees", *SIGMOD '95*, San Jose, CA, M. Carey and D. Schneider (eds.), *SIGMOD RECORD* 24 (2): 92-103, May 1995.

[22] J. F. Allen, Maintaining knowledge about temporal intervals, *Communications of ACM*, 26(11), 1983, 832-843.

[23] 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.

[24] 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.