Cobra: A Content-Based Video Retrieval System

Milan Petković and Willem Jonker

Computer Science Department, University of Twente,
PO BOX 217, 7500 AE, Enschede, The Netherlands
{milan,jonker}@cs.utwente.nl

1 Introduction

An increasing number of large publicly available video libraries results in a demand for techniques that can manipulate the video data based on content. In this paper, we present a content-based video retrieval system called Cobra. The system supports automatic extraction and retrieval of high-level concepts (such as video objects and events) from raw video data. It benefits from using domain knowledge, but at the same time, provides a general framework that can be used in different domains.

The contribution of this work is twofold. Firstly, we demonstrate how different knowledge-based techniques can be used together within a single video database management system to interpret low-level video features into semantic content. The system uses spatio-temporal rules, Hidden Markov Models (HMMs), and Dynamic Bayesian Networks (DBNs) to model and recognize video objects and events. Secondly, we show how these techniques can be effectively used for different application domains. In particular, we validate our approach in the domain of tennis and Formula 1 videos.

2 System Description

The Cobra DBMS is easily extensible, supporting the use of different knowledge-based techniques for identifying the video contents. The content abstractions, which are stored as metadata, are used to organize, index and retrieve the video source (Fig. 1). The meta-data is populated off-line most of the time, but can also be extracted on-line in the case of dynamic feature/semantic extractions in the query time.

At the logical level, the system uses the Moa object algebra [1], enriched with a video data model and several extensions. The algebra accepts all base types of the underlying physical storage system and allows their orthogonal combination using the structure primitives: set, tuple, and object. At the physical level, we use Monet [1] - an extensible parallel database kernel that supports a binary relational model, main memory query execution, extensibility with new abstract data types and index structures, as well as parallelism.

In order to achieve content independence and provide a framework for automatic extraction of semantic content from raw video data, we propose the

C.S. Jensen et al. (Eds.): EDBT 2002, LNCS 2287, pp. 736–738, 2002.

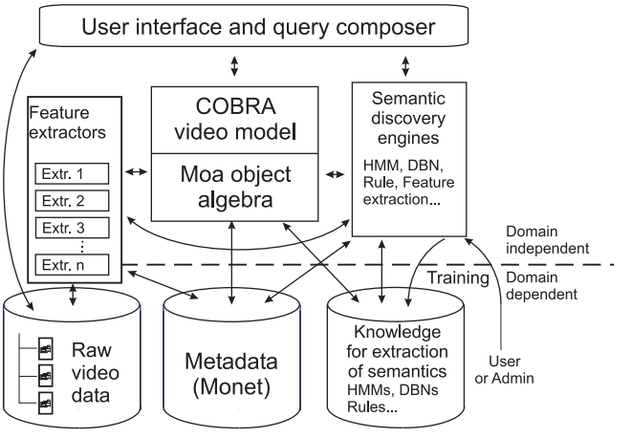

Fig. 1. The architecture of the Cobra system

COntent-Based RetrievAl (COBRA) video data model (a detailed description can be found in [2]). The model is in line with the latest development in MPEG-7, distinguishing four distinct layers within video content: the raw data, the feature, the object and the event layer. The object and event layers are concept layers consisting of entities characterized by prominent spatial and temporal dimensions, respectively. By using this model, we achieved insulation between applications and feature/semantic extraction techniques on one hand, and data on the other hand.

The techniques, which are used to inference high-level semantics from raw video data, are integrated within the system as Moa/Monet extensions. In the current implementation we have three extensions: (1) Video processing and feature extraction, (2) HMM, and (3) rule-based extension.

The video-processing and feature-extraction extension encapsulates operations used for video segmentation, processing and feature extraction. Operations are implemented using Matlab and its image processing toolbox and as such used through a Matlab server directly by the system.

The other two extensions are tightly coupled with the system. The rule-based extension is implemented within the query engine. It is aimed at formalizing descriptions of high-level concepts, as well as their extraction based on features and spatio-temporal reasoning. The HMM extension implements two basic HMM operations: training and evaluation. Here, we exploit the parallelism of our database kernel and implement the parallel evaluation of different HMMs at the physical level. For a more detailed description of these two extensions see [2].

By extending the system at all levels, several knowledge-based techniques have been efficiently integrated within our VDBMS. This is an important advantage over approaches that implement a video extension at the application level, which results in a much slower system.

3 Content-Based Retrieval

Having described our system and the model, this section explains how high-level concepts are extracted from raw video data. We start with segmenting a video into different shots using differences in the color histograms of neighboring frames. Then, shots are classified in different categories. Note that from that step, domain knowledge starts to play a very important role, since the shot classification and event recognition are domain dependent. We continue the process by extracting visual features characterizing color, shape, texture and motion, as well as audio features. In our system we use general, but also domain specific features, such as skeleton features in the tennis domain [2]. Having the features extracted we use domain knowledge coupled with the aforementioned techniques to map them to high-level concepts.

In our experiments, we applied spatio-temporal formalization of video objects and events using rules. In that way, we were able to describe, for example, a tennis player as a video object, and playing close to the net or rallies as video events. Consequently, we were able to interactively query the database for video segments with these semantic concepts. However, this rule-based approach is essentially restricted to the extent of recognizable events, since it might become difficult to formalize complex actions of non-rigid objects using rules. In order to solve this problem, we have exploited the automatic learning capability of HMMs. In our tennis case study, they have been used to recognize different tennis strokes. The results of large experiments we run showed that we were able to recognize six different tennis strokes, namely, forehand, backhand, service, smash, forehand volley and backhand volley, with the accuracy of 88%. The stroke recognition provides our system with the ability to answer even more detailed queries such as: retrieve all video sequences with Sampras approaching the net with the backhand stroke.

On the other hand, to demonstrate the generality of our approach, we have also done some experiments in the domain of Formula 1 videos [3]. Here, we employed dynamic Bayesian networks to find the highlights of Formula 1 programs. We extracted different multi-modal cues (namely text, audio and visual cues) and found that DBNs are very useful for fusing them. The accuracy of about 80% was obtained compared to the human annotation. To the best of our knowledge this is the first time that dynamic Bayesian networks are used for indexing and characterization of TV broadcasting programs.

References

1. P. Boncz, A.N. Wilschut, M.L. Kersten. Flattering an object algebra to provide performance. In *Proc. IEEE Intl. Conf. on Data Engineering*, pages 568-577, 1998.
2. M. Petković, W. Jonker. Content-Based Video Retrieval by Integrating Spatio-Temporal and Stochastic Recognition of Events. In *Proc. IEEE International Workshop on Detection and Recognition of Events in Video*, pages 75-82, 2001.
3. V. Mihajlović, M.Petković. *Automatic Annotation of Formula 1 Races for Content-Based Video Retrieval*, CTIT Technical Report, TR-CTIT-01-41, 2001.