KNOWLEDGE BASED METHODS FOR VIDEO DATA RETRIEVAL

S.Thanga Ramya [1] P.Rangarajan[2]

RMD Engineering College
R.S.M Nagar
Kavaraipettai, Chennai-601 206
[1] thangaramya @ yahoo.com
2 rangarajan69@gmail.com

ABSTRACT

Large collections of publicly available video data grow day by day, the need to query this data efficiently becomes significant. Consequently, content-based retrieval of video data turns out to be a challenging and important problem. This paper addresses the specific aspect of inferring semantics automatically from raw video data using different knowledge-based methods. In particular, this paper focuses on three techniques namely, rules, Hidden Markov Models (HMMs), and Dynamic Bayesian Networks (DBNs). First, a rule-based approach that supports spatio-temporal formalization of high-level concepts is introduced. Then the focus of this paper is towards stochastic methods and also demonstrates how HMMs and DBNs can be effectively used for content-based video retrieval from multimedia databases.

KEYWORDS

Hidden Markov Models(HMM), Dynamic Bayesian Networks (DBNs), Content-Based Video Indexing and Retrieval(CBVIR),Content Based Video Retrieval(CBVR)

1. INTRODUCTION

The development of various multimedia compression standards in last decade has made the widespread exchange of multimedia information a reality. Due to significant increase in desktop computer performance and a decrease in the cost of storage media, extraordinary growth of multimedia information in private and commercial databases has been seen. Further its ubiquity throughout the World Wide Web, presents new research challenges in computing, data storage, retrieval and multimedia communications. Intuitive handling of this vast multimedia information is the demand of users. Keeping this in mind, multimedia and computer vision researchers are focusing on the development of content based multimedia indexing and retrieval. However, evolution of functional multimedia management system is hindered by the "semantic gap"; a discontinuity between simplicity of content description that can be currently computed automatically and the richness of semantics in user's queries posed for media search and retrieval [1]. The availability of cost effective means for obtaining digital video has led to the easy storage of digital video data, which can be widely distributed over networks or storage media such as CDROM or DVD. Unfortunately, these collections are often not catalogued and are accessible only by the sequential scanning of the sequences. To make the use of large video databases more feasible, user should be able to automatically index, search and retrieve relevant material. Content-Based Video Indexing and Retrieval (CBVIR) has been the focus of the research community during last 15 years. The main idea behind this concept is to access

DOI : 10.5121/ijcsit.2011.3514

information and interact with large collections of videos for referring and interacting with its content, rather than its form. Although there has been a lot of effort put in this research area, the outcomes have not been very encouraging.

As an information recording way, video has been more widely used with the rapid development of multimedia computer technology. Video is the most complex media form with strongest performance, which contains large amount of information and vivid images. But the unstructured data format and the non-transparent content of the performance make video data management and analysis (such as video browsing, retrieval) become rather difficult. Retrieving desired video clips quickly and accurately from the massive video database has become one of the key issues in the development of the video database[1]. The former video information retrieval system is to visit videos based on the keywords. But manual index cost large amounts of time and effort, and text index of the video information is bound to be some omissions, or cause the return of a large number of low-quality matches, so the Content-Based Video Retrieval (CBVR) was put forward and become a research focus. First CBVR analyzes the content of images, video and audio directly, and extracts the characteristics and semantic. It then establishes index and retrieves videos utilizing these content characteristics [3] [4]. This paper deals with organizing and managing video more effectively. In this paper, the video retrieval process based on the content allowing users to obtain the video in need from the platform quickly and accurately and improving the utilization of high-quality video materials has been looked in to.

2. SPATIO - TEMPORAL EVENT RECOGNITION

The spatio-temporal event formalization is used for the description of high-level concepts. To extract these concepts automatically based on features and spatiotemporal reasoning, an object and event grammars that define rules for object and event descriptions based on spatial, temporal, and feature operators has been developed [7]**.** The rules facilitate automatic mapping from features to high-level concepts (objects and events contained in a video). The syntax of rules and an engine that supports their use are described in [7]**.** Query example from the tennis video domain is illustrated

```
SELECT vi.frame-seq FROM video vi WHERE s-contains (vi.frame-seq, event, Playernear-
thenet = ({o1: player, o2: net}, {},{},

{y-distance (o1,02 )< 50}, {duration (this) > 60}),o1.name = 'Sampras')
```

The query retrieves all video segments where Sampras is playing from close to the net for a given period of time. It is formulated using an extended OQL, where s_contains is a function that checks if a sequence contains specific objects or events. The "player-near_the-net" event type is defined in terms of spatio-temporal object interactions. This rule uses a spatial relation ***(distance)*** defined on features and also one temporal relation ***(duration).*** The temporal relation says that this event type should last for a specific period of time. Similarly, other 'spatio-temporal' rules are used to define events like rally, long point, etc [7].

Although spatio-temporal formalization can be used for inferring video semantics from low-level feature representations and extracting events like net-playing and rally, spatio-temporal approach has some drawbacks. Spatio-temporal approach is essentially restricted to the extent of recognizable events, since it might become difficult to formalize complex actions of non-rigid objects. Especially, this is applicable for an ordinary user who is not familiar with video features and spatio-temporal reasoning. An expert can help, but even then for some events the approach will not grant the best results. Spatio-temporal approach also requires that someone, either a user or an expert to create an object and event descriptions, which can be time consuming and error-prone.

In order to overcome the above mentioned drawbacks, stochastic techniques, such as Hidden Markov Models (HMMs) have been proposed. These techniques exploit automatic learning capabilities to derive knowledge and avoid the need for an expert as well as reduce the retrieval time.

3. HIDDEN MARKOV MODELS

The block diagram for a basic content based video retrieval system is shown in Figure 1. The process of database population and 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 metadata and indexes. In the process of the database population, the features, objects, and events are extracted. Indexes and metadata are put in the storage server and videos in the file system. Most queries are resolved directly in the storage server. The left out unresolved queries are extracted by the extractors dynamically.

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. It should also support a basic set of spatio-temporal relations. 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 an attribute and a set of intervals, where it occurs. The basic relations of interval and point temporal, the mapping between them, as well as operations on the interval data type, such as intersect and union, have to be defined.

A well-known technique for modelling temporal processes – Hidden Markov Models has been used. In this approach, semantic features are extracted from multimodal behavior of each action. Video and audio extracted features are used in the framework proposed. By using semantic features, it is possible to recognize high-level semantic actions and to encode more semantic details, which will enable users to find answers on questions easier and faster. In order to achieve this goal, a matching computation mechanism has been proposed. For each shot from the image stream, one key frame along with its image features (colour, motion, edge) and from audio stream such as (cries, scream, engines) has been extracted. The output was then parsed by a HMM process[5].

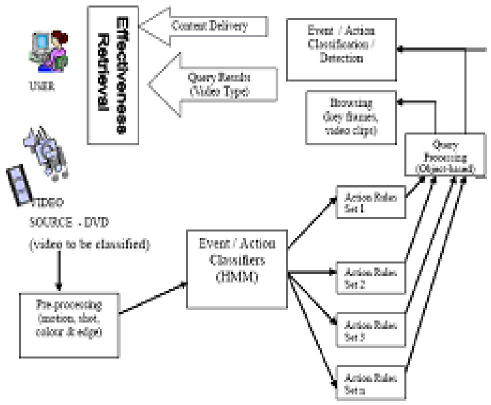
USER
Effectiveness Retrieval
Content Delivery
Query Results (Video Type)
Event / Action Classification / Detection
Browsing (key frames, video clips)
Query Processing (Object-based)
VIDEO SOURCE - DVD (video to be classified)
Pre-processing (motion, shot, colour & edge)
Event / Action Classifiers (HMM)
Action Rules Set 1
Action Rules Set 2
Action Rules Set 3
Action Rules Set n

Figure 1. Block diagram of Content based Video Retrieval System

As a result, each shot is assigned to a classifier. In this study, when a user specifies a clip (sub-sequence) as a query example *Q*, clips whose motion trajectories are similar to *Q*'s are retrieved from the given motion collection. Since the frame number in the collection is large and each frame contains many feature values, searching similar clips at an interactive rate is not a trivial task. The matching value is very important to assist in increasing the detection process. In this recognition process, the input motion data should be quantified to some value, which represents the degree how the input motion satisfies the conditions of the assigned action. This value is termed the *matching value*, and is calculated in each condition of the action with various methods. Matching values have a value between 0.0 and 32, and are used in an evaluation function and as a flag. The evaluation function calculates the matching value based on how closely the pose or motion of the body region satisfies the conditions of a given action, while the flag is used to deal with the sequential features of an action, The system stores the time the action turning left and turning right as the time that the flag is set. The system calculates the matching value for the sequential feature using these times. This method obtains multiple matching values. After calculating the matching values for all features, the matching values are multiplied to create the recognition output. The output represents how well the input matches to the assigned action. If the product of all the features is larger than some threshold, the assigned action is recognized. The number of the features and the threshold of the final product are specific to each action.

Table 1. Range of χ (matching value of motion)

Action	Range of (Matching value of action)
Sitting	$0 \leq \chi < 3.9$
Standing	$3.9 \leq \chi < 7.1$
Walking	$7.1 \leq \chi\chi < 10.7$
Punching	$10.7 \leq \chi < 17.1$
Kicking	$17.1 \leq \chi < 32$
Running	$32 \leq \chi$

4. DYNAMIC BAYESIAN NETWORKS

4.1 Case Study

The aim of the case study is twofold. Firstly, investigating the effectiveness of Bayesian and dynamic Bayesian networks for content-based video retrieval. Secondly, analyzing the applicability of these networks for the fusion of multimodalities in the retrieval process. In particular, main focus is on fusing the evidence obtained from [2].

4.2 Information Sources

In the process of extraction of multi-modal cues, three different media components can be used for the TV broadcasting program: audio, video, and text. Audio plays a significant role in the detection and recognition of events in video. In this domain, the importance of the audio signal is even bigger, since it encapsulates the announcer's comment, which can be considered as a kind of the on-line human annotation.

Furthermore, whenever something important happens the announcer raises his voice due to his excitement, which is a good indication for the highlights. Based on a few experiments, four audio features will be used for speech endpoint detection and extraction of excited speech. Short Time Energy (STE), pitch, Mel-Frequency Cepstral Coefficients (MFCCs), and pause rate has been chosen for the purpose. A description of methods for excited speech and speech endpoint detection can be found in [4]. For the recognition of specific keywords in announcer's speech, a keyword-spotting tool may be used.

For visual analysis, color, shape, and motion features can be used. First, the video is segmented into shots. Then, the amount of motion will be calculated and semaphore, dust, sand, and replay detectors will be applied in order to characterize passing, start, and fly-out events, as well as to find replay scenes (for a description of these detectors see [2]).

The third information source used is the text that is superimposed on the screen. This is another type of online annotation done by the TV program producer, which brings some additional information with intention to help viewers to better understand the video content. In order to speed up the detection and recognition of the superimposed text, the existing technique was modified by considering the properties of race videos [2].

4.3 Probabilistic Fusion

As the majority of techniques for event detection, which relay solely on the one-media cues, showed to have robustness problems, This paper focuses the analysis on the fusion of the evidence obtained from the aforementioned information sources. In order to find the most appropriate technique, numerous experiments has been performed and compared Bayesian Networks (BNs) versus Dynamic Bayesian Networks (DBNs), different network structures, temporal dependences, and learning algorithms. For learning, the Expectation Maximization algorithm is applied. Inference process uses the modified Boyen-Koller algorithm for approximate inference. For the descriptions of both algorithms see [4].

Based on literature, three races, have been digitized namely, the German, Belgian, and USA Grand Prix (GP). Feature values, extracted from the audio and video signals, are represented as probabilistic values in range from zero to one. The features extracted form a video are: keywords (f,), pause rate *(f2).* average values of STE fi), dynamic range of STE *v4)* maximum values of STE *V;),* average values of pitch W6), dynamic range of pitch (f,), maximum values of pitch (fs), average values of MFCCs *G),*maximum values of MFCCs *No),*p art of the race *(f l l)* , replay *VIZ)*color difference *(f l j)* , semaphore Nd), dust Nd, sand (f,6), and motion *N7).*By developing different structures of BNs and corresponding DBN structures. The intention was to explore how different network structures can influence the inference step in this type of networks. The structures of BNs, which are also used for one time slice of DBNs, are depicted in Figure 2.

The query node is Excited Announcer (EA), which is used to determine if the announcer raises his voice due to an interesting event that is taking place in the race. The shaded nodes represent evidence nodes, which receive their values based on features extracted from the audio signal of the video. The temporal dependencies between nodes from two consecutive time slices of DBNs. were defined as in Figure 3.

The BN parameters on a sequence of 300s, consisting of 3000 evidence values, extracted from the audio signal is used. For the DBNs, the same video sequence of 300s, which was divided into 12 segments with *25s* duration each is used. The inference was performed on audio evidence extracted from the German GP. For each network structure precision and recall is computed.

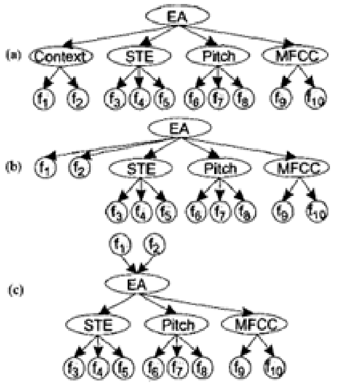

Figure 2. Different structures of processing of audio features

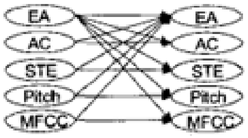

Figure 3. Temporal dependencies for the DBNs

By comparing different BN structures, there is no significant difference in precision and recall obtained from them (around *55%)*. The corresponding DBNs perform similarly, except for the DBN that corresponds to the BN depicted in Figure 2a. It gives much better results than the other BNiDBN networks (more than 80%). Conclusions from experiments performed are twofold. The conclusion is that the DBN learning and inference procedures depend a lot on the selected DBN structure for one time slice. This is not the case when inference and learning are performed with BNs. Secondly, these experiments showed the advantages of the DBN structure depicted in Figure 2a over the other BNDBN networks. The audio DBN can only extract the segments of the I race where the announcer raises his voice. Other interesting segments (highlights), which were missed by the announcer, could not he extracted. Therefore, the employment of the audio DBN for highlight extraction would lead to high precision, but low recall (about 50%).

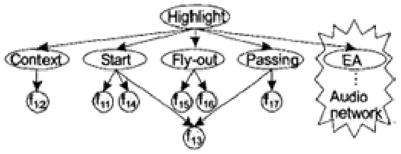

Figure 4. Audio-Visual DBN for one time slice

To improve the results obtained solely from audio cues audio-visual DBN was developed for highlight detection. The structure that represents one time slice of this network is depicted in Figure 4. The Highlight node was chosen to be the main query node, while querying nodes: Start, Fly Out, and Passing, in the experiments. With this network the precision and recall rate of 80% **is** achieved in average for the three races.

5. CONCLUSION

The three approaches, which have been identified and used for bridging the semantic gap, have shown that each of them is advantageous for a specific kind of problems. The spatio-temporal formalization is beneficial for the characterization of events that have spatio-temporal nature and can be described by transitions of object positions and their spatio-temporal relations (e.g. net playing event). HMMs and DBNs are more appropriate for events that have more 'stochastic' nature. Furthermore, by using these two techniques, benefited thing is their automatic learning capabilities. Although this work has not compared the two stochastic approaches between each other, an intuitive conclusion is that the DBN approach is more suitable for fusing multimodalities in retrieval. Based on this conclusion on the property of DBNs that each feature can influence the decision with a specific probability. In HMM approach the process of quantization, which leads to discrete HMMs, has treated all features equally. However, operations with HMMs are less time-consuming than with DBNs. By integrating the work presented in this paper within the content-based video retrieval system presented in [7], the flexibility is improved, exploring the properties of databases as general purpose systems. Therefore, the necessary adjustments of the system, when the application domain changes are minimized[8-16].

6. REFERENCES

[1] J. Calic , N. Campbell , A. Calway , M. Mirmehdi , B. T. Thomas, T. Burghardt , S. Hannuna, C. Kong , S. Porter , N. Canagarajah , D. Bull, "Towards Intelligent Content Based Retrieval of Wildlife Videos", Proc. to the 6th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS'2005, EPFL, Montreux, Switzerland, April 2005.

[2] V. Mihajlovic, M. Petkovic, Automatic Annotation of Formula I Races for Content-Based Video Refrieval, TRCTIT- OI-41,2001, 37 pp.

[3] Fu Shiguang. Research and realization of the search engine based on the theme [D]. Beijing: *Beijing Jiaotong University*, 2007.

[4] Ji Chun, Key frame extraction technology in the video retrieval based on content [J].*Intelligence Journal,* NO 11, 2006.

[5]. John, S., Boreczky and D. Lynn, 1998. A hidden markov model framework for video segmentation using audio an image features. In: Proceedings of IEEE International conference on Acoustics, Speech and Signal Processing, May 12-15, 6:3741-3744. Doi:10.1109/ICASSP . 1998 . 679697

[7] M. Petkovic, Confent-Based Video Retrieval Supported by Database Technology, PhD Thesis, Enschede, the Netherlands, 2003.

[8] P. Turaga, A. Veeraraghavan, and R. Chellappa, "From videos to verbs: Mining videos for activities using a cascade of dynamical systems," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2007, pp. 1–8.

[9] R.Hamid, S.Maddi,A. Bobick, and M. Essa, "Structure from statistics— Unsupervised activity analysis using suffix trees," in *Proc. IEEE Int. Conf Comput. Vis.*, Oct., 2007, pp. 1–8.

[10] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 39, no. 5, pp. 489–504, Sep. 2009.

[11] J. Tang, X. S. Hua, M. Wang, Z. Gu, G. J. Qi, and X. Wu, "Correlative linear neighborhood propagation for video annotation," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 39, no. 2, pp. 409–416, Apr. 2009.

[12] X. Chen, C. Zhang, S. C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," *IEEE Trans. Syst, Man, Cybern., C: Appl. Rev.*, vol. 39, no. 2, pp. 228–233,
Mar. 2009.

[13] Y. Song, X.-S. Hua, L. Dai, and Wang, “Semi-automatic video annotation based on active learning with multiple complementary predictors,” in *Proc. ACM Int. Workshop Multimedia Inf. Retrieval*, Singapore, 2005, pp. 97–104.

[14] Y. Song, X.-S. Hua, G.-J. Qi, L.-R. Dai, M. Wang, and H.-J. Zhang, “Efficient semantic annotation method for indexing large personal video database,” in *Proc. ACM Int. Workshop Multimedia Inf. Retrieval*, Santa Barbara, CA, 2006, pp. 289–296.

[15] S. L. Feng, R. Manmatha, andV. Lavrenko, “Multiple Bernoulli relevance
models for image and video annotation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun./Jul. 2004, vol. 2, pp. 1002–1009

[16] Weiming Hu, *Senior Member, IEEE*, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank “A Survey on Visual Content-Based Video Indexing and Retrieval” in IEEE transactions on systems, man, and cybernetics—part c: applications and reviews

Authors

S. Thanga Ramya obtained her Bachelor’s degree in Computer Science and Engineering in the year 1999 from Manonmanium Sundaranar University, Tirunelveli. She has been awarded M.S (by Research) degree from Anna University in 2008.

She has published papers in two National Conferences. Her research interests include Database Management System and object oriented model. She is a life member of Indian Society for Technical Education.

She worked as Lecturer in P.S.N.A. College of Engineering and Technology, Dindigul and National Engineering College, Kovilpatti. Presently she is working as an Assistant Professor in R.M.D. Engineering College, Kavaraipettai.

Dr.P.Rangarajan passed **B.E** in Electrical and Electronics Engineering from Coimbatore Institute of Technology , Coimbatore in 1990 and **M.E** in Power Electronics from College of Engineering , Guindy , Anna University, Chennai in 1995. He completed his **Ph.D** in **VLSI and Signal Processing** from College of Engineering , Guindy , Anna University , Chennai in 2004. He is in teaching profession for the past **twenty one years**. He has published fifteen papers in International conferences and journals. He has been the Principal coordinator in NSTDB, **DST**, New Delhi for the project in TEDP on “Opportunity Analysis in Real Time Embedded Systems”..He was also the consultant in Kranium Technologies , Chennai. He has guided more than ten M.E thesis .He is currently guiding six Ph.D Scholars in the area of VLSI Signal Processing ,Image Processing, Digital Communication, Datawarehouse and mining.

e-mail : rangarajan69@gmail.com