Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Robust Semantic Concept Detection in Large Video Collections

Jialie Shen
Singapore Management University
Singapore

Dacheng Tao
Nanyang Technological University
Singapore

Xuelong Li
The University of London
United Kingdom

***Abstract*—With explosive amounts of video data emerging from the Internet, automatic video concept detection is becoming very important and has been received great attention. However, reported approaches mainly suffer from low identification accuracy and poor robustness over different concepts. One of the main reason is that the existing approaches typically isolate the video signature generation from the process of classifier training. Also, very few approaches consider effects of multiple video features. The paper describes a novel approach fusing different information from diverse knowledge sources to facilitate effective video concept detection. The system is designed based on CM*F scheme [7], [5] and its basic architecture contains two core components including 1) CM*F based video signature generation scheme and 2) CM*F based video concept detector. To evaluate the approach proposed, an extensive experimental study on two large video databases has been carried out. The results demonstrate the superiority of the method in terms of effectiveness and robustness.**

***Index Terms*—Video concept, Information retrieval, Detection**

I. Introduction

With dramatic increasing volume of video data from various application domains, accurate semantic concept detection is becoming a very important but challenging research topic. Using the technology, video data can be automatically clustered based on ”similarity of concept” for effective organization and faster searching. Due to its critical role in video information retrieval and management, many different approaches have been proposed in recent years [4], [1], [2], [3]. In general, their basic idea is to model the problem as a statistical classification process, which includes two steps. They include,

- Feature extraction - The main functionality is to model video (visual and audio) information and the one of the most popular format is to use feature vectors as content representations (video signatures).
- Concept detection - A categorization scheme to quantify the distance between video clips and concept label on the basis of feature extracted.

Generating high quality video signature has a considerable impact on determining final performance of detection process. However, very limited research study has been carried out in this domain. Existing approaches [11], [8], [9], [10] generally suffer from two main shortcomings including,

- For most of the existing approaches [11], [8], [9], [10], the process of video signature generation is typically isolated from the process of classifier training. This could easily lead to a suboptimal feature space, which might not provide good quality training for classifiers used to detect video concept.
- Another shortcoming with almost all of the existing approaches is that very few approach consider how to combine diverse knowledge source for detection accuracy improvement. Video data could contains a large amount of heterogenous information such as text, audio and image. It may not be appropriate to assume that each of those information provides equal contribution on concept identification. However, design a proper scheme to effectively combine them is still a open research question.
- Low level feature (audio and visual) used to represent content of video information is typically high dimensional. This can easily lead to inefficient identification process based on existing statistical machine learning scheme. Reducing the dimensionality of data becomes a very natural solution for the problem.

Based on above observation, we propose a new video concept detection framework. It is designed using ”wrapper model” feature selection principle, which requires predetermined classification algorithm. The basic architecture contains two core components including 1) CM*F based video signature generation scheme and 2) CM*F based video concept detector. Both components are developed based on CM*F scheme [7], [5] and the main goal for CM*F based video signature generation scheme is to produce high quality feature sets. CM*F is a nonlinear dimensionality reduction method based on neural network and PCA. It can fuse various kinds of video feature to form final content representation of video. Using the signature, accurate event detection process can be carried out using CM*F based classifier. In addition, another main contribution is a set of comprehensive empirical study and relative result analysis using two large video test collections containing TRECVID05 and TRECVID07. It demonstrates various kinds of advantages for the proposed approach including effectiveness and robustness.

The rest of the paper is structured as follows: Section 2 provides detail introduction of the proposed video concept detection framework. Section 3 describes the experimental configuration for empirical study including test collection and evaluation metric. In Section 4, we present experimental results

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

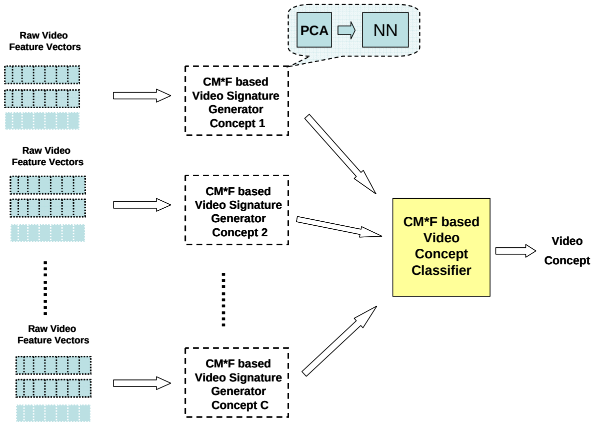

Fig. 1. Illustration of video concept detection system architecture. It contains three basic components including low level video feature extraction, CM*F based video signature generation scheme and CM*F based video concept detector.

TABLE I
DETAIL INFORMATION ABOUT LOW LEVEL VIDEO FEATURES CONSIDERED.

Feature type	Details	Dimensionality
Audio features	Timbre (33D), Rhythm (18D) Pitch (18D)	69D
Visual features	Color (37D), Texture (30D) Shape (30D), Color Layout (30D)	127D

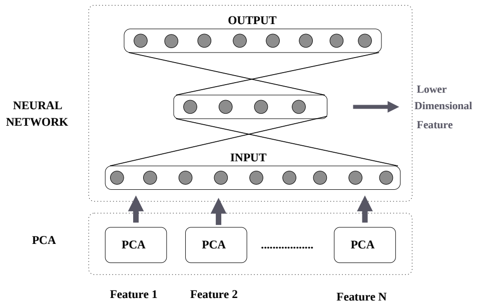

Fig. 2. Illustration of CM*F architecture containing linear PCA and neural network. PCA is used as preprocessing scheme to speed up training and information processing speed. This figure is also shown in [7].

and the related analysis. Finally, Section 5 gives conclusions of this research study and summarizes a few potential directions for future study.

II. A UNIFIED FRAMEWORK FOR VIDEO CONCEPT DETECTION

In the section, we provide the description about each functionality module of our framework. As illustrated in Figure 1, its core architecture includes three components - low level video feature extraction, CM*F based video signature generation scheme and CM*F based video concept detector.

A. Video Feature Extraction

The main functionality of low level feature extraction component is to calculate various kinds of low level physical features to represent content of raw video stream. The whole process can be partitioned into two main steps. In the first step, our system segments each video stream into a set of small shots. After that, different kinds of features are extracted from the individual video segment. They include audio features and visual features. For audio features, we consider timbre, rhythm and pitch information [20]. In addition, information about color, texture, shape and color layout is computed from each video segment [19]. The details can be found in Table I and more information about extraction procedure, please refer to [7].

B. Architecture of Video Concept Detection

The second part of our proposed framework contains multiple CM*F based feature mappers organized with a parallel structure. Each of them is trained for one video concept. Once the low level feature extraction process is finished, task-specific descriptors for video concept can be produced with CM*F based feature mapper. As the basic layout shown in Figure 2, CM*F based feature mappers has a hybrid structure, integrating two traditional dimensionality reduction methods - PCA [12] and multilayer preceptron neural network [15], into single structure. PCA serves as ”pre-processing” step for neural network to reduce the size of raw input video features. It can significantly improve convergence speed of training process for neural network. The multilayer preceptron neural network provides complex maps approximately to produce effective low dimensional feature space.

Comparing to traditional approaches, CM*F based feature mapper enjoys a few unique characteristics. The first one is to enable comprehensive fusion of features from different diverse knowledge sources and produce small but comprehensive video descriptor. At the same time, it also can effectively combine human classification information via training process and learning examples. The corresponding transformation process can be treated as a nonlinear mapping Φ, which projects the raw input feature space Υ onto the concept sensitive feature space Ω, as shown by

$$\Phi : \Upsilon \rightarrow \Omega \tag{1}$$

where each feature vector in the input space Υ is a composite vector containing various kinds of information about audio and visual information for video clip. Since mappers are trained separately, the output feature space Ω can be represent by a set of statistical independent numerical vectors $[cv_1, .., cv_c.., cv_C]$. They can be directly applied as inputs for

CM*F based video concept detector, whose main functionality is to make final decision on the concept label for given video input. To train both CM*F feature mapper and CM*F based video concept detector, we apply the quick-prop learning algorithm as learning algorithm due to its simplicity and effectiveness. In this study, the related learning examples are randomly selected from test collection and its size is about 10% of whole testset. Based on the system just introduced above, we can summarize video concept detection process as below,

- Video boundary detection and partition video stream into a set of video segments.
- For each video segment, extract low level video features and feature types include audio features and visual features.
- Advanced video signature generated by CM*F based feature mapper.
- Video concept detection using CM*F based classifier.

III. Experimental Configuration

This section provides information about two test collections and evaluation metric used for the empirical study. Our empirical study was carried out based on two video collections used in the recent TREC Video Retrieval Evaluation (TRECVID). They include data from TRECVID 2005 (TRECVID05) and TRECVID 2007 (TRECVID07). For TRECVID05, it contains totally 86 hours of news video and those data is from 6 different TV channels including CNN, NBC, MSNBC, CCTV, NTDTV and LBC. Except that news from CCTV and NTDTV are Chinese, the others are in English. This 86-hour video footage includes 61,901 video shots. Figure 3 shows a few examples for video concepts from TRECVID 2005. For TRECVID 2007, main content is very mixed and the topics covered include news magazine, science news, news reports, documentaries, and educational programming. In this 50 hour footage, there are 21,532 shots.

For semantic concepts used for this empirical study, we consider total 39 semantic concepts provided by the Light Scale Concept Ontology for Multimedia (LSCOM- Lite) project [22]. The range of those concepts is quite wide. They include objects (e.g., Car), scenes (e.g., Outdoor), and semantic topics (e.g., Military). To assess the effectiveness of our approach, we use recall R and precision P as evaluation metrics [6].

IV. Experimental Results

The main focus of the first set of experiment is to study effectiveness of the proposed video detection method. To demonstrate its advantage, we compared its performance based on four different low level video feature configurations. They include (audio feature, visual feature), (visual feature), (audio feature) and (linear combination of visual and audio features). All tests use the same data sets and same queries introduced in the previous section. Table II and III summarize the performance results of our approaches based on TRECVID05 and TRECVID07 test collection. Based on results shown in the last row, linear combination leads to the lowest detection accuracy among those four configurations. Moreover, when the system only considers audio feature, it also suffers from poor detection accuracy. Although using visual feature can provide improvement at some level for both test collections, the gain is very limited. In fact, combining different kinds of video feature via CM*F based feature mapper brings signification lift to detection effectiveness. For example, in comparison to the systems based on visual feature or audio feature, additional 15% in precision measurement and 17% in recall can be found for TRECVID05 dataset when both visual and audio features are considered. On the other hand, we can find that detection accuracies obtained using two different test collections are close. For example, there is only 5.7% difference between precision ratio obtained using TREVID05 and TREVID07. The results show the method we propose enjoys superior robustness over different test collections.

Systems	Recall(R)	Precision(P)
AF+VF	0.372	0.426
VF	0.255	0.304
AF	0.197	0.215
AF+VF(L)	0.117	0.187

TABLE II
Comparison of detection effectiveness on TRECVID05. AF+VF, VF and AF denote our method based on audio and visual feature, visual feature, and audio feature. AF+VF(L) denotes linear combination of audio and visual features.

Systems	Recall(R)	Precision(P)
AF+VF	0.315	0.417
VF	0.235	0.284
AF	0.188	0.212
AF+VF(L)	0.125	0.194

TABLE III
Comparison of detection effectiveness on TRECVID07. AF+VF, VF and AF denote our method based on audio and visual feature, visual feature, and audio feature. AF+VF(L) denotes linear combination of audio and visual features.

The training examples play a deterministic role in the quality of learning process for our system. However, the process to acquire training examples generally requires large amount of human labor and high level domain knowledge. One of desirable property for learning based video concept detection systems is to sustain good detection accuracy with smaller size of learning set. To study system behavior in this aspect, we firstly created nine different-sized training sets by randomly selecting 1%, 2%, 5%, 7%, 10%, 12%, 15%, 17% and 20% of TRECVID05 test collection. Then effectiveness based on different settings is investigated. Table IV shows the experimental results and we observe that the accuracy for system degrades significantly as size of training examples is too small (e.g., 1% and 2% of whole collection). This is because very limited information is available for learning. On the other hand, the performance of our approach is relatively robust against the volume of learning data. There is no dramatic decrease in accuracy with relatively small size of training

Fig. 3. Examples of video concepts from TRECVID 2005

Systems	Recall(R)	Precision(P)
20%	0.379	0.431
17%	0.377	0.429
15%	0.372	0.427
12%	0.372	0.426
10%	0.372	0.426
7%	0.372	0.426
5%	0.372	0.426
2%	0.248	0.357
1%	0.215	0.315

TABLE IV
COMPARISON OF DETECTION EFFECTIVENESS ON TRECVID05 WITH DIFFERENT SIZE OF TRAINING EXAMPLE. VIDEO FEATURE CONSIDERED INCLUDE VISUAL AND AUDIO FEATURES.

set, e.g. around 0.426 precision for larger training data (5% of whole collection). The main reason is that the system proposed is constructed using multiple video features and designed based on ”wrapper model” principle, which can result in optimal feature space for classification. From results given above, we can conclude that the proposed video concept detection enjoys both effectiveness and robustness.

V. CONCLUSION AND FUTURE STUDY

Video concept detection is becoming an increasing important due to its rich applications. While it has received great attention for a long time, the existing approaches still suffer from low detection accuracy and poor robustness across different concepts. In this research, we study problem how to produce a compact and comprehensive video signature for the purpose of concept detection. A novel approach has been developed to fuse different information from diverse knowledge sources based on CM*F scheme [7]. In addition, we also present a detection framework containing two core components including 1) CM*F based video signature generation scheme and 2) CM*F based video concept detector. To validate the approach, we have carried out an extensive experimental study based on two large video databases. It demonstrates the effectiveness and robustness of our approaches.

A few interesting research questions have been initiated by this research study. For example, in this paper we randomly select training examples without optimization process. Identification performance can be improved further if we can design an optimization scheme to enhance quality of training examples. Another promising direction is to study applications of the framework in other application domains including image classification and speech recognition.

REFERENCES

[1] J. Yang, R. Yan, and A. Hauptmann, “Cross-Domain Video Concept Detection using Adaptive SVMs,” *ACM Multimedia*, 2007.

[2] J. Tang, X. Hua, G. Qi, M. Wang, T. Mei and X. Wu, “Structure-sensitive manifold ranking for video concept detection,” *ACM Multimedia*, 2007.

[3] J. Wang, Y. Zhao, X. Wu, and X. Hua, “Transductive multi-label learning for video concept detection,” *ACM MIR*, 2008.

[4] R. Yan and M. Naphade, “Semi-Supervised Cross Feature Learning for Semantic Concept Detection in Videos,” *CVPR*, 2005.

[5] J. Shen, A. Ngu, J. Shepherd, D. Huynh and Q. Z. Sheng, “CMVF: A Novel Dimension Reduction Scheme for Efficient Indexing in A Large Image Database,” *ACM SIGMOD*, 2003.

[6] C. Manning, P. Raghavan and H. Schtze,, “Introduction to Information Retrieval,” Cambridge University Press, 2008.

[7] J. Shen, “Advanced Query Processing on Large Multimedia Databases,” *UNSW PhD Thesis*, 2007.

[8] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, “Unsupervised mining of statistical temporal structures in video, video mining, eds. a. rosenfeld, d. doermann and d. dementhon,” 2003.

[9] D. A. Sadlier and N. E. ÒConnor, “Event detection in field sports video using audio-visual features and a support vector machine,” *IEEE T-CSVT*, 2005.

[10] Y. Ke, R. Sukthankar, and M. Hebert, “Efficient visual event detection using volumetric features,” in *Proc. of ICCV*, 2005.

[11] M. Shyu, X. Xie, M. Chen, and S. Chen, “Video semantic event/concept detection using a subspace-based multimedia data mining framework,” *IEEE T-MM*, 2008.

[12] F. Fukunaga and W. Koontz, “Applications of the karhunen-loève expansion to feature selection and ordering,” *IEEE TComputers*, 1970.

[13] TREC Video Retrieval Evaluation (TRECVID’05), ”http://www-nlpir.nist.gov/projects/tv2005/tv2005.html”, 2005.

[14] TREC Video Retrieval Evaluation (TRECVID’04), ”http://www-nlpir.nist.gov/projects/tv2004/tv2004.html”, 2004.

[15] S. Roweis and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding”, *Science*, 2000.

[16] T. Zhang, J. Yang, D. Zhao, and X. Ge, “Linear local tangent space alignment and application to face recognition,” *Neurocomputing*, 2007.

[17] E. Kokiopoulou and Y. Saad, “Orthogonal neighborhood preserving projections,” in *Proceedings of Fifth IEEE International Conference on Data Mining*, 2005.

[18] C.O’Toole, A. Smeaton, N. Murphy, and S. Marlow, “Evaluation of shot boundary detection on a large video test suite,” in *Proc. of Challenges in Image Retrieval*, 1999.

[19] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall, 2002.

[20] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE T-SAP*, 2002.

[21] T. Tolonen and M. Karjalainen, “A computationally efficient multipitch analysis model,” *IEEE T-SAP*, 2000.

[22] D. L. L. Definitions and Annotations, “http://www.ee.columbia.edu/dvmm/lscom/.”