

# A Generic Approach for Systematic Analysis of Sports Videos

NING ZHANG

Ryerson University

LINGFANG LI

Institute of Computing Technology

Chinese Academy of Sciences

JUN DU

NEC Research Labs China

LING GUAN

Ryerson University

\*LING-YU DUAN

Peking University

QINGMING HUANG

Institute of Computing Technology

Chinese Academy of Sciences

WEN GAO

Peking University

---

Various innovative and original works have been applied and proposed in the field of sports video analysis. However, individual works focused on sophisticated methodologies with particular sport types and there was a lack of scalable and holistic framework in this field. This paper proposes a solution for this issue and presents a systematic and generic approach which is experimented on a relatively large-scale sports consortia. The system aims at the event detection scenario of an input video with an orderly sequential process. Initially, domain-knowledge independent local descriptors are extracted homogeneously from the input video sequence. Then the video representation is created by adopting a bag-of-visual-words (BoW) model. The video's genre is firstly identified by applying the k-nearest neighbor (k-NN) classifiers on the initially obtained video representation, with various dissimilarity measures are assessed and evaluated analytically. Subsequently, an unsupervised probabilistic latent semantic analysis (PLSA) based approach is employed at the same histogram-based video representation, in characterizing each frame of video sequence into one of four view groups, namely closed-up-view, mid-view, long-view and outer-field-view. Finally, A hidden conditional random field (HCRF) structured prediction model is utilized for interesting event detection. From experimental results, k-NN classifier using KL-divergence measurement demonstrates the best accuracy at 82.16% for genre categorization. Supervised SVM and unsupervised PLSA have average classification accuracies at 82.86% and 68.13%, respectively. The HCRF model achieves 92.31% accuracy using the unsupervised PLSA based label input, which is comparable with the supervised SVM based input at an accuracy of 93.08%. In general, such a systematic approach can be widely applied in processing massive videos generically.

---

This article extends the previous work by the authors appearing under the title "Automatic sports genre categorization and view-type classification over large-scale dataset," [Li et al. 2009].

Authors' addresses: N.Zhang, L.Guan, Ryerson Multimedia Research Laboratory, Ryerson University, 350 Victoria St. Toronto, ON. M5B2K3, Canada; email: {n2zhang,lguan}@ee.ryerson.ca; L-Y.Duan, W.Gao, Institute of Digital Media, Peking University, Beijing, 100871, China; email: lingyu@pku.edu.cn, wgao@jdl.ac.cn; L.F.Li, Q.M.Huang, Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China; email: {lfl, qmhuang}@jdl.ac.cn; J.Du, NEC Research Labs China, Beijing, 100084, China; email: dujun@research.nec.com.cn

\*Correspondence should be addressed to Ling-Yu Duan, lingyu@pku.edu.cn.

Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.

© 2010 ACM 0000-0000/2010/0000-0001 \$5.00

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*abstract methods; indexing methods*

General Terms: Algorithms, Measurement, Performance, Experimentation

Additional Key Words and Phrases: Generic framework, Genre categorization, View classification, Event detection

---

## 1. INTRODUCTION

Living in the information era, we are surrounded by an enormous scale of digital contents. According to Bohn and Short [2010], the size of estimated newly created digital data in the year of 2011 is about 1800 exabyte (1 exabyte=1 billion gigabytes), roughly 100 times of the production in the year of 2002 (2 ~ 3 exabyte). This is equivalent to a 10-fold growth rate annually in average. In terms of the image and video content, YouTube has a 73M videos uploaded every year at a rate of 15 hrs/min; while the number of digital images on the internet is about 5 billion, in which only 5% – 10% of them are labeled [Yan and Hsu 2009].

Among aforesaid explosive growth of multimedia data, sports videos contribute significantly to the total collections of the digital content. Analysis on sports video has drawn more and more attention in the research community, due to its huge popularity and vast commercial values. The sources of the sports video collections are also various: from daily basis public recreations to professional sports games broadcasting; from amateur digital camcorder to professional TV broadcasting, and plenteous but low-quality online streamed videos. Most of literature works focus on particular sports and tasks, utilizing domain-knowledge and production rules. [Xu et al. 2001; Ekin and Tekalp 2002; Xu and Li 2003; Nepal et al. 2001; Zhu et al. 2009]. Supervised learning is another important characteristics adopted by these works to fill the semantic gap. Although aforementioned methods have their merits and brilliance, most of them are stand-alone with little inter-connection. They also suffer from a lack of generality and scalability to the large-scale data with two reasons. Firstly, with various video contents of different themes and cinematographic techniques, domain-knowledge associated methods have difficulties in extensibility. Secondly, labeled data is required for the supervised learning, while the majority of multimedia data currently available is unlabeled. In order to tackle these two issues in sports videos, our proposed approach focuses on developing a domain-knowledge independent feature selection and video representation with an unsupervised learning technique.

In this paper, a generic and systematic framework is proposed with experimentations on a relatively large-scale sports video dataset. We use the term *relatively large-scale* to accurately describe this dataset we engaged with, which is not truly large-scale yet, but with the same complexity of video types and contents. Three tasks are introduced in a systematic and generic manner, such that the output from the previous tasks are utilized as the input to the next task. Event detection is the 3rd and final quest with two preceding tasks, video genre categorization and semantic view type classification. By accomplishing these three tasks, an event detection can be achieved with minimum domain-knowledge and insufficient labeled

data.

The contribution of this paper is two-fold:

(1) A comprehensive survey is conducted targeting on existing works of sports video analysis from aspects of low-level genre, middle-level views/shots and high-level semantics. Individual literature works have their own merit and credit in the field. They either focus on a generic method using probabilistic modeling for instance; or focus on a systematic approach. However emphasizing on the generic property tends to be a non-systematic approach, while pursuing a systematic approach is prone to ignore the generality.

(2) Such a deficiency we are aware of during the survey study leads to the proposed work, a generic and systematic approach on analyzing the sports video. This is the second contribution of this paper, which can be further divided into the following three sub-contributions. (2.1) Initially, a domain-knowledge free local descriptors are extracted using a homogenous process. A bag-of-visual-words (BoW) model is used to build a histogram based distribution to represent video clips. The BoW model with local features is a natural selection for generically processing videos due to its domain-knowledge free property. (2.2) Subsequently, since unlabeled data takes the major portion of all digital content, an unsupervised classifier taking the homogeneously processed representation of part (2.1) is preferred, such that an automatic and systematic process can be deployed towards a large-scale dataset. Since sports videos have well defined semantic view types from their production characteristics, local features and the BoW model is a perfect candidate in view classification as which have been proved successful in computer vision and object recognition fields. Therefore, a probabilistic latent semantic analysis (PLSA) based method for semantic view classification is preferred due to its unsupervised nature and fitting to the BoW model input. (2.3) Lastly, a structured prediction model is a suitable in taking labeled middle-level agents as input to achieve high-level semantics. This is because that sports videos have distinguishable temporal patterns often consisting of sequences of middle-level agents. In our work, since semantic view types have been classified in part (2.2), an appropriate approach is to take view results as input and achieving semantic events detection. Therefore, hidden conditional random field (HCRF) is introduced as a rational choice. The significance of the HCRF is its generalized modeling, which resides in both the relaxation of the Markov property and incorporation with hidden states of the conditional random field (CRF) modeling.

The rest of this paper is organized as follows. In section 2, an extensive review in video analysis is provided. An overview of the proposed system with a flowchart is given in section 3. Proposed techniques achieving various tasks are addressed in the next three sections. The generic feature extraction, BoW model using the proposed bottom-up structure in codebook generation, and the genre categorization technique are presented in section 4. In section 5, middle-level view classification is analyzed by adopting the PLSA based unsupervised model. Section 6 presents a discriminant HCRF structured prediction model on high-level event detection. Experimental results are given in section 7. Finally, the paper is concluded in section 8.

## 2. RELATED WORKS

This section reviews the related works in the domain of sports video analysis. This survey appreciates each individual work for its contribution and value to the research field. Although various researches reviewed in this work are inspirational and innovative, there is a lack of work focusing on a holistic aspect, from an angle of generality and systematic property. Most of the literature works focus on a single aspect. Some works focused on specific sport types with sophisticated techniques. Some researches targeted on generic approaches but lack of systematic analysis. And other works proposed systems with automatic process, but lack of the generic and scalable properties. In the following, we are going to examine both the merits and disadvantages of the literature works with the following order, from low-level feature extraction with video genre categorization, middle-level view classification, to high-level semantic event detection.

### 2.1 Genre categorization

Video genre and its categorization was one of the earliest video analysis drawn researchers' interest. The main task of this genre categorization starts from different big group of videos such as sports, music, news, movies etc., and gradually moves to more delicate categorization such as to identify the sports types. Various works have been highlights in the following. However, a major and common disadvantage for these works are their heavy dependencies on the domain-knowledge.

Fischer *et al.*[1995] first proposed a classification method based on 5 different video genres. Brezeale and Cook [2008] provided an extensive survey in this field. Incorporating the survey and most recent works, a concise summary is provided in Table I. Color feature with C4.5 decision tree were used in [Truong et al. 2000]. Camera motion feature with statistical classifiers were chosen to classify 6 sports genre in [Takagi et al. 2003]. A principal component analysis (PCA) modified audio-visual feature was taken to train a Gaussian mixture model (GMM) classifier in [Xu and Li 2003]. Semantic shots(views) were used to help in genre categorization in [Jaser et al. 2004]. Motion and color as well as audio features were applied in [Wang et al. 2006]. Color features with a hierarchical support vector machine (SVM) were used in [Yuan et al. 2006]. High-level MPEG-7 features were extracted and applied in multi-modality classifiers in [Glasberg et al. 2008]. The best classification result at the moment is with an accuracy of 95% using a dataset of 8 different genres [Montagnuolo and Messina 2009]. These methods used various domain-knowledge with supervised classifiers to achieve the automatic genre categorizations.

As defined in [Ekin et al. 2003], domain-knowledge based features can be divided into two categories, cinematic-based features and object-based features. The cinematic feature involves middle to high level semantics from common video composition or production rules such as shots/views or events, while object-based features are described by their spacial property, such as color, shape, texture as well as spatial-temporal based object motions. As Table I shows, all reviewed works are domain-knowledge dependent, either object-based or cinematic-based. A lack of diversity, i.e. the number of different genres in the database, restricts these methods from generality.

Table I. Summary of previous video genre categorization methods.

Authors and Year Published	Number of Genres	Size of Database (hrs)	Domain-knowledge		Genre Categorization Method	Accuracy rate
			Object Based	Cinematic Based		
[Truong et al. 2000]	4	8	Yes	Yes	C4.5 decision tree	83%
[Takagi et al. 2003]	6	33.75	Yes	Yes	statistics based	n/a
[Xu and Li 2003]	5	5	Yes	No	PCA & GMM	86.5%
[Jaser et al. 2004]	4	n/a	Yes	Yes	decision tree and HMM	91.6%
[Wang et al. 2006]	3	16	No	Yes	pseudo-2D-HMM	n/a
[Yuan et al. 2006]	6	33.33	No	Yes	hierarchical SVM	94%
[Glasberg et al. 2008]	5	5	Yes	Yes	Multi-model	88.5%
[Montagnuolo and Messina 2009]	8	100	Yes	Yes	Parallel Neural Networks	95%

Table II. A comparison of view classification techniques in literature, emphasizing on features utilization and classification methods. In the "Global Features" column with "Others (yes/innov)" category: "yes" means other than color and texture global features are used while not innovative, while "innov" means newly designed features are used. For the "View Classification Method" column,  $S$  indicates the supervised method, while  $UnS$  indicates the unsupervised method.

Authors and Year Published	Nature of data	Global Features			Local Feature Based	View Classification Method
		Color Based	Texture Based	Others (yes/innov)		
[Xu et al. 2001]	Soccer	Yes	No	No	No	thresholding ( $S$ )
[Ekin and Tekalp 2002]	Soccer	Yes	No	Yes	No	morphological operations ( $S$ )
[Duan et al. 2003]	4 Sports	Yes	Yes	innov	No	Decision Tree ( $S$ )
[Tong et al. 2004]	Soccer	Yes	Yes	Yes	No	Decision Tree ( $S$ )
[Wang et al. 2007]	4 Sports	Yes	Yes	Yes	No	spectral clustering ( $UnS$ )
[Bennokhtar et al. 2008]	Soccer	Yes	Yes	Yes	No	Neural-network ( $S$ )
[Zhong et al. 2008]	3 Sports	Yes	No	No	No	Spectral-division algorithm ( $UnS$ )
[Kolekar and Palaniappan 2009]	Soccer	Yes	No	Yes	No	Decision Tree ( $S$ )

## 2.2 View classification

Views(shots) are considered as middle agents to link low-level features and high-level semantic events [Ekin and Tekalp 2002; Duan et al. 2003]. Supervised approach is a favourite choice in the research community. Although the labeling effort is not the primary concern because of the size and diversity dealt by current researches; such task becomes more and more unaffordable along the growth of the dataset. Therefore, approaches using unsupervised learning techniques with generality and efficiency ought to be sought for analyzing large-scale multimedia consortia. We summarize related works so that readers could compare popular supervised means with proposed unsupervised PLSA in this paper. Additionally, there are only two works using unsupervised techniques sought by our extensive study, we present them for the completeness of the review [Wang et al. 2007; Zhong et al. 2008].

Although there might be different nomenclatures, the fundamental purpose of the middle-level views(shots) is to involve certain production rules to help high-level tasks. This frame-based label concept was first introduced by Xu *et al.*, whom defined three groups of views: global, zoom-in and close-up [2001]. Ekin and Tekalp [2002] used a slightly different long-shot, middle-shot, close-up/out-of-field notation. Duan *et al.*[2003] used a finer view/shot groups classification, supported by innovative semantic features. These mentioned pioneering methods along with other works such as [Tong et al. 2004; Wang et al. 2005; Kolekar and Palaniappan 2009] focus on using decision tree classifiers to link the low-level features to view/shot types. Xu *et al.*[2001] and Ekin *et al.*[2002] applied color-based grass detector and field/object size to determine view types. Incorporating previously mentioned features. Tong *et al.*[2004] added head-area detection as well as grey level co-occurrence matrix(GLCM) to improve the decision tree on classification. Wang *et al.*[2005] used field region extraction, object segmentation and edge detection for view type decision making. Duan *et al.*[2003] firstly extended the research from single genre (soccer) to multiple genres (four sports) using individual genre based decision trees. Different from previous visual feature extraction methods, Kolekar and Palaniappan [2009] took an top-down approach. They first used audio feature to find the exciting video clip. Subsequently, the motion features of the whole image volume along with the background color information are used for view types classification. Benmokhtar *et al.*[2008] took an approach on feature level fusion using dynamic PCA with information coding neural-network (NN). At the classification level, another NN is used to fuse multi-modality inputs. However, these supervised methods are limited by the labeled data and thus constrained from expanding to larger scales.

Some other researchers pursued unsupervised methods for view classification. Wang *et al.*[2007] proposed an information-theoretic co-clustering method, in which mutual information was maximized by treating shot classes and features as two random variables. As a consequence, color histogram and perceived motion energy features are used with a test set of four sports video genres. Zhong *et al.*'s method was inspired from spectral theory conventionally used to solve segmentation in graph theory [2008]. They proposed a spectral-division algorithm to find the proper video shots clustering, which were tested in three sports videos using HSV space color feature. Although good performances have been obtained in above mentioned

methods, the extensibility and flexibility towards diverse genres and large-scale are very limited. This is again due to the domain-knowledge dependency of the extracted features.

Table II compares aforementioned methodologies from angles of feature utilization and classification techniques. Color and texture are two major global features used by most works. Duan *et al.*'s work is the only one proposing middle level features developed from low-level global features. The rest of works either adopted additional popular global feature schemes such as audio feature, Gabor feature as well as some production rule-based features; or didn't utilize any. While various global features are used, none of the local features have been applied. Moreover, most of the supervised methods (except Duan *et al.*'s work) focus on single type soccer sport, while unsupervised techniques employed various sports types.

### 2.3 Event detection

As one of the most popular semantic tasks in video analysis, event detection has been a popular topic from the beginning of the multimedia research. Despite of different definitions of event detection by different researchers, commonly acknowledged properties of an "event" can be summarized as follows. An event occupies a period of time and is described using the salient aspects of the video sequence input, which consists of smaller semantic units or building blocks [Lavee et al. 2009]. Lavee *et al.* also summarized and classified event detection algorithms into three categories, including: a) pattern-recognition models, b) semantic event models, and c) state event models. Pattern-recognition models focus on direct classification from low-level features but are lack of semantic linkage. Semantic models target on high-level semantic rules and constraints with domain-knowledge. This requires a lot of human involvement in creating rules and regulations using prior information. State models utilize abstracted middle-level agents as well as intrinsic structure of the event itself.

By comparing the above three categories of event modeling with examples in literature, we think that the pattern-recognition model is heavily dependent on the classifiers, which at the moment are not intelligent enough to understand all semantics from low-level features. On the other hand, the semantic model considerably relies on human expertise and thus underestimates the accuracy and efficiency provided by classification tools. From our experience, the state model incorporates the strength of pattern-recognition at low-level with classifiers at high-level, so that it utilizes both feature extraction power and classification intelligence. Moreover, the state model also accommodates an automatic process and unsupervised learning, which reduces human input into the system. Therefore, state event models are suitable for analyzing large-scale dataset, from both generic and systematic point of views. A coarse-to-fine strategy fits well into such state event models, by roughly localizing the event firstly with context information and then precisely detecting the event using advanced structure model. A detail description is presented in section 6.

Although we prefer the state event model for its natural fit to the proposed systematic approach in this work, other two models are still appreciated for their efficiencies in analyzing the sports video and utilizations in other applications. In the following, state of the art works are summarized and compared.

Table III. A comparison of event detection models emphasizing the feature utilization from both low-level features and middle-level semantic agents. In the "Low-level Multimodal Features" column, various features are utilized, including audio (A), visual (V), text (T), motion feature (M), and video shot detection (S), as well as a "n/a" label in the case when no low-level feature mentioned in the related works.

Event-Detection Algorithm Category	Authors and Year Published	Nature of data	Number of Events	Low-level Multimodal Features	Visual Features		Middle-level Semantic Agents
					Global-Based	Local-Based	
Pattern-Recognition Model	[Xu et al. 2003]	Tennis	5	AVM	Yes	No	Yes
	[Sadtler and O'Connor 2005]	4 field sports	2	AVS	Yes	No	No
	[Ye et al. 2005]	Soccer	1	n/a	n/a	n/a	Yes
Semantic Event Model	[Li et al. 2009]	Basketball	5	VM	Yes	No	No
	[Babaguchi et al. 2002]	Football	3	VST	Yes	No	No
	[Zhang and Chang 2002]	Baseball	2	VT	Yes	No	No
State Event Model	[Ekin et al. 2003]	Soccer	3	VS	Yes	No	Yes
	[Nepal et al. 2001]	Basketball	1	AVMT	Yes	No	No
	[Duan et al. 2003]	Tennis/Soccer	16	AVMT	Yes	No	Yes
State Event Model	[Tong et al. 2004]	Soccer	2	VM	Yes	No	Yes
	[Wang et al. 2006]	Soccer	5	AVM	Yes	No	Yes
	[Zhang et al. 2007]	Basketball	5	VT	Yes	No	Yes
	[Tien et al. 2008]	Tennis	4	AVS	Yes	No	No
	[Mei and Hua 2008]	Soccer	3	VM	Yes	No	Yes
	[Xu et al. 2008]	Soccer/Basketball	17	VTS	Yes	No	Yes
	[Zhu et al. 2009]	Soccer	6	VMTS	Yes	No	Yes



Support vector machine (SVM) is a popular pattern-recognition model approach [Lavee et al. 2009]. Some groups use rich audiovisual features such as face-detection, scoreboard information as well as geometry of the field to find certain semantic events. Saldier and O'Connor [2005] used SVM to classify "scoring" events for 4 different field sports. Xu *et al.*[2003] analyzed tennis video by using hierarchical-SVM applying on fused audio-visual modalities. Similarly, Ye *et al.*[2005] utilized middle-level view labels as well as shot length and camera motions descriptors. A SVM-based incremental learning scheme using updated data is proposed in detecting soccer events, along with a pre-defined temporal structure. A similar approach combining SVM and pre-defined temporal structure was proposed by Li *et al.*[2009], targeting on basketball events using optical flow patterns.

Some semantic event models using rules and logic and semantic relationships are presented. Babaguchi *et al.*[2002] used closed caption text stream with audiovisual feature and the intermodal correlation among them to search a "touch down" event from 4 hours American football videos. Zhang *et al.*[2002] also focused on superimposed caption frames and used decision trees to decide the event such as "scoring" or "last pitch" for baseball games. Ekin *et al.*[2003] incorporated production rules and soccer sport rules to detect certain events such as "goal", "referee", and "penalty-box".

In terms of state event models, one of the earliest work targeting on structures of videos was from Nepal *et al.*[2001], who empirically studied the temporal model in basketball videos based on manual observation, using heuristic method and low-level audio-visual features. Duan *et al.*[2003] also created a temporal structure using multi-modality with heuristic experience on tennis events. Another approach of learning temporal structure is from the data-mining perspective, where Tien *et al.*[2008] focused the tennis match events detection by creating a max-subpattern tree and learning the frequent patterns from it.

Another important branch of state event models are structured prediction models, such as hidden Markov models (HMMs) and their variations, Bayesian networks, as well as discriminative conditional random fields (CRFs). Zhang *et al.*[2007] proposed an HMM-based statistical method in classifying middle-level agents generated from web-casting texts. Tong *et al.*[2004] used Bayesian networks to classify "shoot" and "card" events in soccer videos, by applying decision-tree based intermediate layer concept units. Mei and Hua [2008] proposed an innovative mosaic based middle-agent for key-event mining using HMMs. Wang *et al.*[2006] proposed a CRF model on detecting semantic soccer events and the performance turned out to be better than both SVM and HMMs. A similar approach was also proposed by Xu *et al.*[2008] using CRFs for basketball and soccer event detection, where a webcast text feature was obtained to achieve middle-level concepts. An interesting event tactic analysis is proposed by Zhu *et al.*[2009], which is beyond the conventional event and adopts the cooperative nature and tactic patterns of the team sports. Extensive experiments were conducted on the soccer sports.

Table III provides a comparison of aforementioned literature works, from feature utilization point of view. Most of the methods utilizes the multimodality schemes of features input. By comparing the number of events processed, it appears that the state event model has a better scalability in examining various event scenarios.

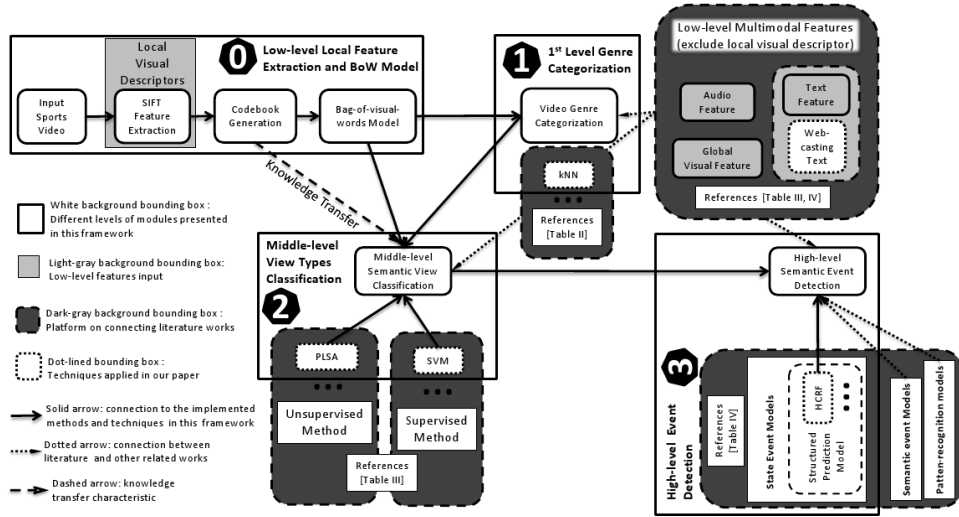


Fig. 1. A flowchart of the proposed generic framework, with one module of generic video representation and three task modules in sequence. Besides the three-level modules in the *white background bounding boxes*, this framework also highlights the relationship between our system and existing literature works, which are shown in the *dark-gray background bounding box*. Associated Table references are also indicated in each module. Multimodal features excluding local visual features are also introduced at various stage by literature works. The *Dotted arrows* are used to represent these associations. The *solid arrows* denote the proposed and implemented techniques in our work. The *dashed arrow* represents a knowledge transfer characteristic of the generated codebooks. In summary, codebooks generated from certain sports with abundant resources, can be transferred and utilized in classifying other sports materials with scarce resources. The detail analysis is introduced in the section 7.2.

It is also interesting to point out that local visual feature hasn't been utilized in any of the method. In addition, a lot of method, especially the state event models, requires middle-level semantic agents to bridge the gap between the low-level features and the high-level events. And such middle-level agents have to be labeled data. However, for the generic approach presented in this work, we tackle the event detection problem using the input obtained by unsupervised learning and unlabeled data.

### 3. OVERVIEW

This section provides a system overview from a holistic aspect illustrated in Figure 1, such that the input sports video is analyzed systematically using a generic and sequential framework. This is interpreted such that the result from a preceding process is input to the next process with a consistent and coherent fashion. There are four modules in total, while module 1–3 are tasks introduced in this work, while module 0 is the infrastructure effort in generic low-level feature extraction and video representation. The highlights of this framework include: (1) a generic foundation using domain-knowledge free local feature is developed to represent input sports videos. This method would fits the general framework in sports video analysis and provides an alternative solution to alleviate generality, scalability, and extension

issues. (2) a thorough and systematic structure starting from genre identification is presented, which was ignored in some related works by assuming the genre type as the prior knowledge. (3) a general platform is introduced to associate our approach to abundant and valuable existing literature works as well as various and innovative features input.

At the module 0, the low-level local feature utilization incorporating with the codebook generation and the BoW model provides an expandable groundwork for the semantic tasks of genre categorization, view classification and high-level event detection. As our survey shows, the local feature is rarely explored in the domain of the sports videos, though it has been broadly adopted and proved effective in the field of computer vision. Most of the literature works involve domain-knowledge and production rule at the feature extraction level. In our structure, a homogenous process is first introduced for extracting domain-knowledge independent local descriptors. A BoW model is used to represent an input video by mapping its local descriptors to a codebook, which is generated from an innovative bottom-up parallel structure. The histogram based video representation are treated as sole input (no other feature models) to both the genre categorization and the view classification modules. Such a concise representation built from the BoW model benefits users in homogeneously extract visual features and represent videos in a compact and collective form.

At the *1st* module, the videos are categorized by its genre. Video genre nomenclature is used to describe the video type, which is defined as the highest level of granularity in video content representation. Since the video genre categorization task directly relies on low-level features, the proposed feature extraction of target video sequence is used in categorization. In large-scale videos, a successful identification of the genre serves as the first step before attempting higher level tasks. For instance, in sports event detection, an unknown "shooting" event is the target quest, which could be from the ball game or the shooting sport. By indiscriminately treating the entire dataset, this event will be searched through all types of sports. However, since sports like figure-skating and swimming have no "shooting" at all, the effort in searching this event at those non-relevant sports becomes infeasible. Instead of treating all data indifferently, a more efficient approach is to identify the genre of the query video first, and then deploy middle/high-level tasks consequently. As the survey shows in sports video analysis, most of the related works on view classification and event detection assume the genre by default. This framework, however, provides a system that automatically identify the genre from various types of sports data before further analysis.

In the middle-level and the *2nd* module, semantic view types are classified using an unsupervised PLSA learning method to provide labels for input video frames. View describes an individual video frame by abstracting its overall content. It is treated as a bridge between low-level visual features and high-level semantic understanding. In addition, unsupervised learning saves a massive amount of human effort in processing large-scale data. Moreover, the supervised methods can also be implemented upon our proposed platform. Therefore, a SVM model is executed as the baseline for the comparison purpose.

Finally at the *3rd* module, a structured prediction HCRF model using labeled

inputs is a natural fit to the system in detecting semantic events. This can be justified that a video event occupies various length along the temporal dimension. Thus, the state event model based HCRF is suitable to deploy. Less comprehensive baseline methods such as the hidden Markov model and the conditional random field can also be applied in this platform.

In the following section, module 0 and module 1 are combined and presented including feature extraction, bag-of-visual-words model as well as genre categorization.

#### 4. FEATURE EXTRACTION, BAG-OF-VISUAL-WORDS MODEL, AND GENRE CATEGORIZATION

This section covers the first part of our proposed framework, generic feature extraction with the BoW model, and systematic genre categorization. Figure 2 illustrates details of each process.

##### 4.1 Feature extraction

Local invariant features are chosen for homogenous feature extraction due to its domain-knowledge free property. The scale, rotation, and illumination invariant properties make these descriptors good candidates in preserving the similarities for semantic objects and events matching and detection. Global features, on the other hand, rely on domain-knowledge and have difficulties in robust concept and event detection, especially in the presence of noise and occlusion [Jiang et al. 2010]. Scale-invariant feature transform (SIFT), developed by Lowe, is selected as feature descriptors in this work [2004]. SIFT method extracts key-points of an image and describes these points using local neighborhood regional information. Since no prior and domain knowledge required, SIFT is an ideal option in the large-scale automatic and homogenous process. By processing image sequences sampled from video clips, each frame is represented by a magnitude of hundreds of SIFT descriptors. After homogenous local descriptor extraction, the BoW model is applied, whose effectiveness relies on a robust codebook design. In order to achieve this resiliency, we propose a two-level bottom-up K-means clustering for codebook generation. The advances of the bottom-up structure are efficiency, scalability and robustness.

##### 4.2 BoW model with two-level bottom-up codebook generation

BoW is a widely recognized model for successfully utilizing key-point based local features and has shown great results in concept detections of images [Jiang et al. 2010; Lay and Guan 2006; Yang et al. 2007]. A representative codebook is synthesized using codewords, which are exemplars of combining all SIFT descriptors. A video clip is then characterized by mapping its SIFT feature points to the generated codebook and a histogram distribution is obtained. This compact representation preserves the information with a small size in storage. In addition, random noisy features can be suppressed in terms of a frequency-based histogram representation.

With the large-scale dataset, efficiency and robustness of the codebook formation have been an important concern for the BoW model. Heuristically, the larger the codebook size, the better the classification results, with certain saturation limitations [Philbin et al. 2007; Yang et al. 2007]. Different codebook sizes have been explored, ranging from several hundred [Lazebnik et al. 2006; Zhang et al. 2007],

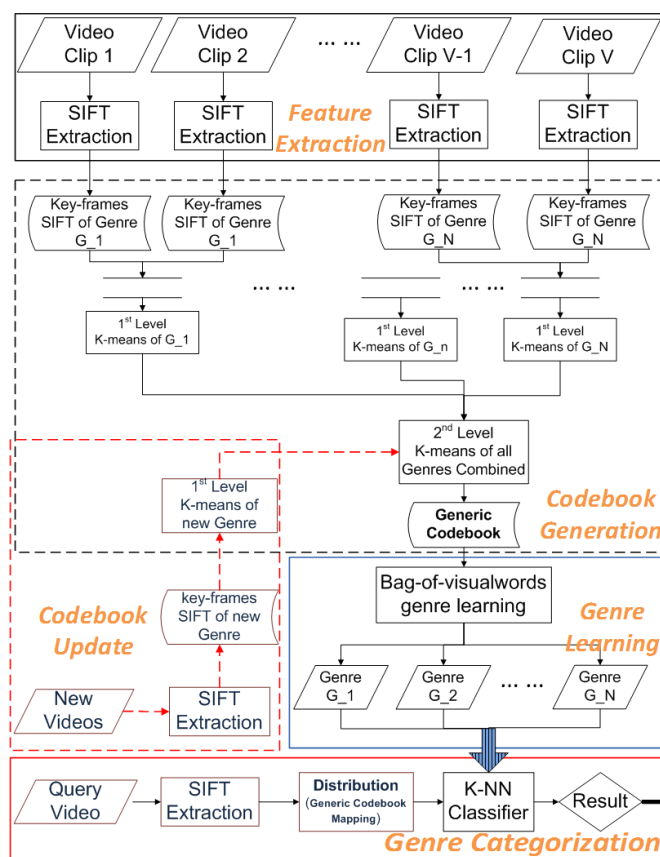


Fig. 2. Feature extraction and genre categorization framework using data parallelism and bottom-up structure for codebook generation.

to thousands [Sivic and Zisserman 2003], to hundreds of thousands [Philbin et al. 2007]. Since using different datasets, there is no conclusion drawn for a decision rule. In this article, choices of codebook sizes are based on the empirical studies.

K-means clustering is utilized to generate a codebook by finding and appointing cluster centers as codeword values. In a large-scale domain, satisfactory performance has been reported using a top-down structure for categorization [Li et al. 2009]. In that work, a two-layer top-down structure is used for sports genre categorization. At the first-layer, a general codebook (size 800) is generated using single K-means, in which a query video is only categorized to one of the pre-defined bigger groups consisting of several genres. Such a group is determined by those sports sharing similar semantics. At the second-layer after a bigger group belonging is identified, an individual codebook (size 200) for this bigger group is used to decide the video genre. For instance, Judo and Boxing are combined into a bigger group named *martial arts*, where the *martial arts* is used as the first-layer candidate. Subsequently, Judo and Boxing are differentiated in the second-layer categorization. Although good classification accuracy has been reported, efficiency

and robustness are problems of such a method in creating a general codebook using single K-means clustering. This is because that most computation of K-means lies in calculating the distances between individual points to their cluster centers in each iteration. A single K-means clustering using large-scale data is heavy in computation and sometimes inaccurate due to K-means own limitations. Since more than 3 million high-dimensional SIFT points are used for building the codebook in our application, one single K-means clustering becomes inefficient.

Therefore, a two-level bottom-up structure is proposed in this work for efficient codebook generation. At the bottom structure, individual genre codebooks are generated in 1<sup>st</sup>-level K-means clustering. At the upper structure, the 1<sup>st</sup>-level codebooks are used as the input for the 2<sup>nd</sup>-level K-means to create the generic codebook. By using this bottom-up structure, we reduce the heavy computation in measuring individual point-to-cluster-center distance in K-means algorithm. Moreover, since the 1<sup>st</sup>-level K-means are independent from each other, distributed computing methods can be applied to further reduce the computation time. The numerical analysis can be referred in section 7.1.

Another advantage of bottom-up K-means clustering resides in the system update and scalability. In the case of new genre videos added to the dataset, a codebook update module is applied to find the new genre's individual codebook. The result together with existing codebooks is used to generate the new generic codebook by only re-running the 2<sup>nd</sup>-level K-means. In the case that new videos are imported for an existing genre, the corresponding 1<sup>st</sup> level K-means is applied to achieve the updated individual codebook and then 2<sup>nd</sup>-level K-means is re-run to update the generic codebook.

At the next step, training data is characterized by frequency-based histogram representation. The individual genre is modularized as a distribution denoted by  $P$  using training data of its own kind.

### 4.3 Genre Categorization

In the final genre categorization stage, a query video is expressed as a histogram  $Q$ , also using the generic codebook and BoW model. Then, a k-Nearest Neighbor (k-NN) classifier is applied with a defined dissimilarity measurement between the query  $Q$  and a trained individual genre  $P$ . Consequently, the query video is identified as the genre whose distribution the query is closest with in measure. Technical detail is presented in section 7.1.

By identifying the genre of this query video, subsequent processes are confined to a focused group and the scale of computation is decreased. Therefore, advanced and sophisticated techniques can be used in middle/high-level video analysis.

## 5. MIDDLE-LEVEL VIEW TYPES CLASSIFICATION

This section introduces the middle-level view classification, where the previously built BoW model is also used in feature representation of view types. As this work targets on large-scale videos, an unsupervised based solution is more applicable and realistic. Therefore unsupervised probabilistic latent semantic analysis (PLSA) based model is focused. PLSA has demonstrated promising result in analyzing co-occurrence data of words and documents in text retrieval [Hofmann 2000]. From a matrix factorization point of view, PLSA belongs to a subgroup called

non-negative matrix factorization, where the factorized matrices are non-negative [Hofmann 1999]. Because the codebook paradigm with codewords is adopted in mapping visual features to a probability based histogram which has to be non-negative, PLSA becomes a more suitable selection comparing to other factorization techniques such as singular value decomposition or principle component analysis.

PLSA relies on the likelihood function of multinomial sampling and aims at an explicit maximization of the predictive power of the model. Incorporating the PLSA plate notation in Figure 3 with the view classification application, the observed state  $w$  is defined as codewords with a total of pre-defined codebook size  $M$ . An individual video frame is denoted by  $d$  with a total number of training frames  $N$ . The latent state  $z$  is the view type and the parameter  $K$  is the total number of view classes, and in this work  $K$  equals 4. The likelihood function is given in Equation 1. The probabilistic distribution is defined as  $p(w_i|d_j)$ , where  $w_i$  is an individual codeword and  $d_j$  is a training frame. Such distribution can be represented by a sum-of-product of two distributions  $p(w_i|z_k)$  and  $p(z_k|d_j)$ . The former is interpreted as an impact on codewords by a view type, while the latter is the probability of a particular view type given a training frame. The number counted of codeword  $w_i$  appearing in a frame  $d_j$  is denoted as  $n(w_i, d_j)$ . The argument of maximum posterior (MAP) estimate  $z^*$  is optimized by using expectation maximization (EM) as shown in Equation 2.

$$\begin{aligned} L &= \prod_{i=1}^M \prod_{j=1}^N p(w_i|d_j)^{n(w_i, d_j)} \\ &= \prod_{i=1}^M \prod_{j=1}^N \left( \sum_{k=1}^K p(w_i|z_k) p(z_k|d_j) \right)^{n(w_i, d_j)} \end{aligned} \quad (1)$$

$$z^* = \arg \max_z p(z|d) \quad (2)$$

Since SVMs have demonstrated a great performance in the field of classification, it is adopted in our view classification task for comparison purpose. In general, supervised models tend to yield better results but require pre-defined knowledge. A typical radial basis function (RBF) is used as the non-linear kernel in SVM [Chang and Lin 2001] and shown in Equation 3. In this equation,  $x_i$  and  $x_j$  represent the codewords, and  $\gamma$  is the kernel parameter of the RBF.

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0. \quad (3)$$

Four view types are defined, namely close-up-view, mid-view, long-view and outer-field-view. This definition is also popular among other works in this field [Xu et al. 2001; Ekin and Tekalp 2002; Duan et al. 2003]. For the PLSA based model, the number of view types is required in terms of human effort, and no labeling requires for individual frames. On the contrary, SVM based model demands both semantic pre-defined view types as well as all frames labeled with ground-truth, which could be unaffordable when the video is large-scale in size.

As the result of view classification task, the query video sequence is labeled with

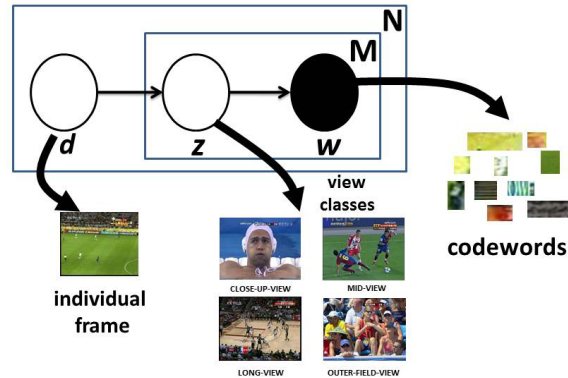


Fig. 3. PLSA model in plate notation is used in the view types classification. Parameter  $z$  is the latent state, with a total of  $K = 4$  view classes, defined as {closed-up-view, mid-view, long-view, and outer-field-view}.  $d$  is the individual frame.  $w$  is the codeword. The two pre-defined constants  $M$  and  $N$  are the codebook size and the total number of training frames, respectively.

view types. In the next section, models which take labeled video sequence as input for detecting interesting events are introduced.

## 6. HIGH-LEVEL EVENT DETECTION

Content-based video event detection is among the most popular quest for the high-level semantic analysis. Different from video abstraction and summarization which target on any interesting events happening in a video rush, event detection is only constrained to a pre-defined request type, such as the third goal or the second penalty kick in a particular soccer match. In sports video, a consumer's interest of events resides in the actual video contents, more than just the information delivered. For instance, a user wants to watch particular goals in basketball games, or replays in soccer matches. S/he is not only interested in the information like what/who/how scored, but more importantly the visual contents rendered from the sport clip. On the other hand, sports videos also have a very strongly correlated temporal structure. In a way, such the structure can be interpreted as a sequence of video frames which have patterns and internal connections. This pattern existence is ubiquitous due to the nature of the sports, a competition where players learn from the standard in order to excel. Therefore, an intuitive approach is to find such patterns using certain representation and learn the temporal structure. Luckily, the PLSA approach provides such a labeled frame sequence and what we need is a clever technique on which portion of the video to analyze and what robust structured prediction model to use. Following, we will introduce a coarse-to-fine scheme and hidden conditional random field (HCRF) for the event detection.

### 6.1 Event detection using coarse-to-fine scheme

Before learning the tempo and patterns, a starting and entry point of an event needs to be seized. A two-stage coarse-to-fine event detection strategy is suitable for this scenario. The first stage is a rough event recognition and localization utilizing rich and accurate text-based information either from web-casting text or optical



character recognition (OCR) techniques of the score-board update. In the second stage, precise video contents associated to the semantic event has been detected in terms of the event boundary detection and accuracy analysis.

The coarse-to-fine techniques have been proved effective and accurate from our previous works. Web-casting text for coarse stage event detection and video alignment was studied and analyzed such as replay scenes and various goal and shot scenes detection in soccer video. [Dai et al. 2005; Xu et al. 2006] At the coarse stage, we captured the text event by extracting keywords from either the well-structured or freestyle web-casting text. Then, the extracted text event provided a time stamp for the visual event entry point. At the second fine-stage, our previous work continued to rely on the web-casting textual information such as text/video alignment, and the accurate information match such as the detail process of the event, including players' involvement in the event [Xu et al. 2006]. Since the experiment conducted in this work focus on fine-stage process with basketball data, we won't repeat the previous work using web-casting text for video analysis. Above mentioned related works can be referred in details for those who are interested.

## 6.2 Hidden conditional random field (HCRF) model

In this paper, since the proposed framework targets on the generic learning model that can be extended to large-scale, we rely on the visual contents, i.e. the local features extracted and middle-level views classified from such features. To demonstrate the effectiveness of the proposed model, we focus on a particular basketball score event detection. We adopted the previously developed score-board update detection method for a coarse-stage process in order to obtain the time stamp [Miao et al. 2007]. The fine-stage process focuses on a robust and accurate visual content detection from the score event. The video sequence is analyzed by distinguishing the actual score event from false alarm events such as time-out or intermission which are also concurrent with score board information. We propose a HCRF based structured prediction model utilizing previously classified views, and completing the generic approach. For example, the HCRF model can be used to detect the score event in basketball for exciting events and highlights. Such a HCRF technique belongs to the state event model defined in the related works. Therefore, the HCRF takes the labeled sequences as input in a natural and seamless fashion. On the other hand, the HCRF is a comprehensive model which can be degraded to hidden Markov models (HMM) or conditional random fields (CRF) with certain constraints. The merits of HCRF comparing the other two models are its resilience and robustness with combination of both the hidden states and the Markov property relaxation. Technical details are examined in the following.

There are several advantages of using the HCRF in large-scale datasets than HMM or CRF models. Firstly, HCRF relaxes the Markov property which assumes that the future state only depends on the current state. In our generic framework, video frames are uniformly decimated and sampled, regardless of the temporal pace of video itself. In some cases, several consecutive frames have the same labeling while in other cases, different labels are assigned. Markov property based model such as HMM is appropriate for the former scenarios but not suitable for the latter ones, since the future state in HMM only cares about the current state label but not previous states. On the other hand, HCRF is flexible and takes surrounding

states from both before and after the current state. Thus, HCRF is more robust for dealing with large-scale homogeneous process and uniform sampling with no prior knowledge. For instance, if a key frame immediate preceding the current state is missed due to the uniform sampling, such an information loss could be compensated by including and summing up distant informational frames (both previous and future) from uniform sampling without misclassifying the event.

Secondly, HCRF has merit in its hidden states structure, which helps to relax the requirement of explicit observed states. This is also an advantage in dealing large-scale uniformly sampled video frames. It is because that in computation, the CRF model outputs individual result label (such as event or not event) per state and requires separate CRFs to present each possible event [Xu et al. 2008]. In HCRF, only one final result is presented in terms of multi-class events occurring probabilities. From the robustness point of view, a CRF model can be easily ruined by semantically unrelated frames due to the automatic uniform sampling. A multi-class HCRF on the other hand, can correct the error introduced by such unrelated frames using probability-based outputs [Quattoni et al. 2007].

Moreover, the HCRF is also appealing in allowing the use of not explicitly labeled training data with partial structure [Quattoni et al. 2007]. From literature, HCRF has been successfully used in gesture recognition [Wang et al. 2006; Quattoni et al. 2007] and phone classification [Gunawardana et al. 2005].

Figure 4(a) illustrates a HCRF structure, in which a label  $y \in Y$  of event type is predicted from an input  $\mathbf{X}$ . This input consists of a sequence of vectors  $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m, \dots, \mathbf{x}_M$ , with each  $\mathbf{x}_m$  representing a local state observation along the HCRF structure. In order to predict  $y$  from a given input  $\mathbf{X}$ , a conditional probabilistic model defined in [Quattoni et al. 2007] and defined in Equation 4 is adopted. In the equation, the model parameter  $\theta$  is used to describe the local potential function  $\psi$  which is expanded in Equation 6. A sequence of latent variables  $\mathbf{h} = h_1, h_2, \dots, h_m, \dots, h_M$  are also introduced in Equation 4, which are not observable from the structure of Figure 4(a). Each  $h_m$  member of  $\mathbf{h}$  corresponds to a state of  $s_m$ . The denominator  $Z(\mathbf{X}; \theta)$  is the normalization factor which is expanded in Equation 5.

$$P(y|\mathbf{X}, \theta) = \sum_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{X}, \theta) = \frac{\sum_{\mathbf{h}} e^{\psi(y, \mathbf{h}, \mathbf{X}; \theta)}}{Z(\mathbf{X}; \theta)} \quad (4)$$

$$Z(\mathbf{X}; \theta) = \sum_{y', \mathbf{h}} e^{\psi(y', \mathbf{h}, \mathbf{X}; \theta)} \quad (5)$$

$$\psi(y, \mathbf{h}, \mathbf{X}; \theta) = \sum_t \sum_k \theta_k^1 f_k^1(y, h_t, \mathbf{X}) + \sum_t \sum_k \theta_k^2 f_k^2(y, h_{t-1}, h_t, \mathbf{X}) \quad (6)$$

In the event detection application, each  $\mathbf{x}_m$  from  $\mathbf{X}$  is a vector descriptor called local observation. In the notation, the  $\mathbf{x}_m$  value at a time  $t$  is defined as  $\mathbf{x}_m(t) = [p_{ws_1}(t), p_{ws_2}(t), p_{ws_3}(t), p_{ws_4}(t), p_{wc}(t)]$ , with each entry of  $\mathbf{x}_m(t)$  calculated from an average result of a sliding window centering at time  $t$ , as Figure 5 shows. The first four entries of  $\mathbf{x}_m(t)$  are the probabilities of four possible view types, where  $p_{ws_j=1,2,3,4}(t)$  associates with close-up-view, mid-view, long-view and outer-field-

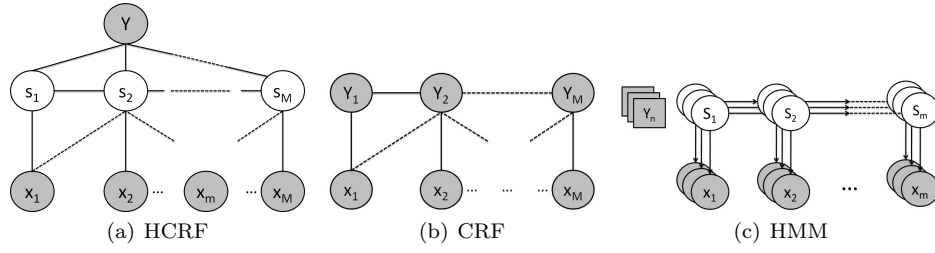


Fig. 4. Structured Prediction Models. (a): Hidden conditional random field (HCRF). (b): Conditional random field (CRF). (c): Hidden Markov Model (HMM).

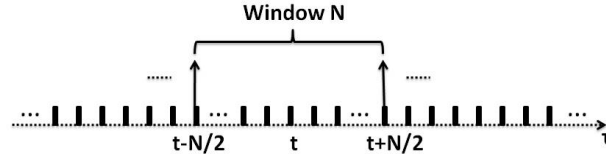


Fig. 5. HCRF input shown in Equation 7, by sliding window average result on view types of decoded image sequence.

view by  $j = 1, 2, 3, 4$  respectively. The fifth  $p_{wc}(t)$  value is an associated directional motion descriptor, introduced by Tan *et al.*[2000]. The formula to calculate the average values at the time stamp  $t$  are given in Equation 7, where individual frame based probabilities are  $p_{s_{j=1,2,3,4}}$  and  $p_c$ .

$$p_{ws_j}(t) = \frac{1}{N} \sum_{\tau=t-N/2}^{t+N/2} p_{s_j}(\tau) \quad \text{with } j = 1, 2, 3, 4$$

$$p_{wc}(t) = \frac{1}{N} \sum_{\tau=t-N/2}^{t+N/2} p_c(\tau) \quad (7)$$

A label and training sequence pair is defined as  $(y_i, \mathbf{X}_i)$  with the index number  $i = 1, 2, \dots, n$ . For each pair,  $y_i \in Y$  and  $\mathbf{X}_i = \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \mathbf{x}_{i,m}, \dots, \mathbf{x}_{i,M}$  are the event label and observed states as Figure 4(a) depicts. For instance,  $\mathbf{x}_{i,m}$  is interpreted as the  $m^{\text{th}}$  sampled time state of the  $i^{\text{th}}$  training sequence, where  $\mathbf{x}_{i,m}(t) = [p_{i,ws_1}(t), p_{i,ws_2}(t), p_{i,ws_3}(t), p_{i,ws_4}(t), p_{i,wc}(t)]$ .

During HCRF training, the parameters  $\theta_k^1$  and  $\theta_k^2$  need to be learned. As Equation 6 shows,  $\theta_k^1$  and  $\theta_k^2$  are coefficients for the state feature function  $f_k^1$  which only involves a single hidden state, and the transition feature function  $f_k^2$  involving two adjacent hidden states, respectively. In order to find the optimal parameters, a log-likelihood objective function is used as shown in Equation 8, with a second term called shrinkage prior to avoid the parameters getting too large. A limited-memory version of Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) quasi-Newton gradient ascent method [Morency et al. 2008] is applied to find the optimal  $\theta^* = \text{argmax} \mathcal{L}(\theta)$ . L-BFGS algorithm is chosen due to this method's efficiency and performance from both theory [Sha and Pereira 2003] and application [Xu et al. 2008].

In the optimization process, the conditional probability in Equation 8 is substituted by the explicit form in Equation 4 to get Equation 9. Then partial derivatives of a training sample  $\mathcal{L}_i(\theta)$  with respect to  $\theta_k^1$  and  $\theta_k^2$  are derived in Equations 10 and 11 respectively.

$$\mathcal{L}(\theta) = \sum_i \log p(y_i | \mathbf{X}_i, \theta) - \frac{1}{2\delta^2} \|\theta\|^2 \quad (8)$$

$$\mathcal{L}(\theta) = \sum_i \log \left( \frac{1}{Z(\mathbf{X}_i; \theta)} \sum_{\mathbf{h}} e^{\psi(y_i, \mathbf{h}, \mathbf{X}_i; \theta)} \right) - \frac{1}{2\delta^2} \|\theta\|^2 \quad (9)$$

$$\begin{aligned} \frac{\partial \mathcal{L}_i(\theta)}{\partial \theta_k^1} &= \sum_t P(h_t | y_i, \mathbf{X}_i) f_k^1(y_i, h_t, \mathbf{X}_i) \\ &\quad - \sum_{t, y'} P(h_t, y' | \mathbf{X}_i) f_k^1(y', h_t, \mathbf{X}_i) \end{aligned} \quad (10)$$

$$\begin{aligned} \frac{\partial \mathcal{L}_i(\theta)}{\partial \theta_k^2} &= \sum_t P(h_{t-1}, h_t | y_i, \mathbf{X}_i) f_k^2(y_i, h_{t-1}, h_t, \mathbf{X}_i) \\ &\quad - \sum_{t, y'} P(h_{t-1}, h_t, y' | \mathbf{X}_i) f_k^2(y', h_{t-1}, h_t, \mathbf{X}_i) \end{aligned} \quad (11)$$

### 6.3 Connection with conditional random field (CRF) and hidden Markov model (HMM)

For the comparison purpose, we also utilized a conventional CRF model as depicted in Figure 4(b). By following definitions in [Lafferty et al. 2001], the conditional probability function is shown in Equation 12, with the normalization factor in Equation 13. The potential function is defined in Equation 14, where  $v_j(Y_{t-1}, Y_t, \mathbf{x})$  is a transition feature function between state positions  $t$  and  $t-1$  with the entire observation sequence; while  $s_k(Y_t, \mathbf{x})$  is a state feature function at state position  $t$ .  $\lambda_j$  and  $\mu_k$  are parameters to be estimated for transition and state feature functions, respectively.

$$P(\mathbf{Y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \cdot \exp \left( \sum_{t=1} F(\mathbf{Y}, x, t) \right) \quad (12)$$

$$Z(\mathbf{x}) = \sum_{Y'} \exp \left( \sum_{t=1} F(Y', \mathbf{x}, t) \right) \quad (13)$$

$$F(Y, \mathbf{x}, t) = \sum_j \lambda_j v_j(Y_{t-1}, Y_t, \mathbf{x}) + \sum_k \mu_k s_k(Y_t, \mathbf{x}) \quad (14)$$

HMM algorithm is also provided in Equation 15 and depicted in Figure 4(c).

$$\begin{aligned} P(Y | X) &= P(X, Y) / P(X) \\ &= \prod_t P(X_t | Y_t) \cdot P(Y_t | Y_{t-1}) \end{aligned} \quad (15)$$

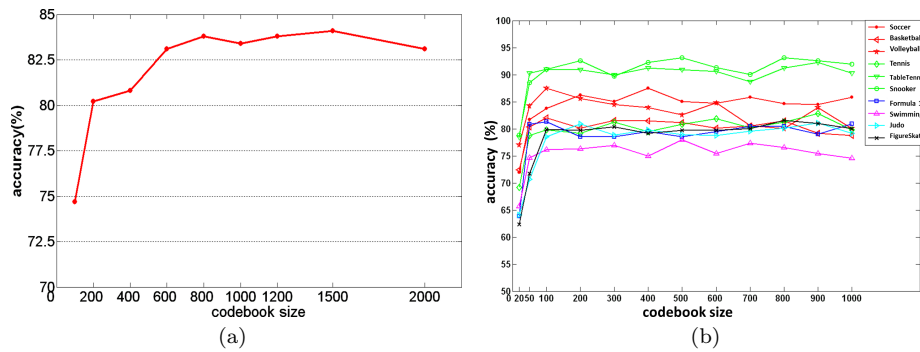


Fig. 6. Empirical studies on the codebook size selection. (a): Average sports accuracy performance for genre categorization, with sports listed in the column plot 0??. (b): Individual sport accuracy performance for view classification.

Different methods are used for detecting an event in decision stage of aforementioned three structured prediction models. For the HMM, the query sequence will be tested and the highest likelihood of the HMM provides the final decision in event detection. On the other hand in the CRF model, since each state variable  $Y(t)$  requires a label as Figure 4(b) shows, a majority-rule voting scheme in which the most event labels along the  $Y$  sequence decide the event result. For the HCRF model depicted in Figure 4(a), a multi-class training process recognizing all classes at the same time is adopted. Therefore, a detected event with the highest probability is considered as the final result for the query sequence.

## 7. EXPERIMENTS AND RESULTS

In the following, experimental results are presented to justify the properties of the proposed generic framework, specifically using a relatively large-scale video collection including 23 genres with a total of 145 hours gathered by the authors, named as 23-sports dataset. To our best knowledge, this dataset is the most diverse one in video genres, which was collected from both internet and TV recordings. All the video clips have the same length of 167 seconds with a total of 500 uniformly sampled frames at a sampling rate of 3 frames per second. This dataset is composed of 3122 clips. In training, 1198 clips are used, in which a subset of 46 clips (2 clips per sport) are used in codebook generation with a total of 3,112,341 SIFT points. In testing, the other 1924 clips are selected.

Various codebook sizes are studied at firstly. Then the proposed system is evaluated by three experiments, with a particular event detection as its ultimate measurement: (1) genre categorization using the proposed bottom-up codebook generation is analyzed; (2) view classification results are assessed and compared using both supervised and unsupervised classifiers; (3) finally, the coarse-to-fine event detection is examined by investigating the basketball score event. The validity on the score event detection can be extended to other event scenarios with labeled video sequence. The detailed argument can be found in section 7.3.

To investigate the codebook size effectiveness, a subset of the 23-sports dataset including 14 sports is used. The clip numbers of these sports range from 70 to 106

at an average of 87, while individual clip has a uniform 167 seconds in length. Two experiments are conducted on the codebook size selection for genre categorization and view classification, respectively. For genre categorization, the average accuracy performance of all sports as a function of different codebook sizes is shown in Figure 6(a). The plot plateaus after the codebook size 800 and starts to drop at 1500. For view classification, the accuracies of individual sports as a function of different codebook sizes are shown in Figure 6(b). Although various accuracy levels are observed for each sport, the individual performance follows the similar plateau trend. Based on these empirical studies, it is concluded that the performances are proportional to codebook sizes, with stable results at codewords ranges of 800-1500 and 800-1000 for genre categorization and view classification, respectively. This study is also consistent with existing researches [Philbin et al. 2007; Yang et al. 2007; Jiang et al. 2010]. In the following experimentation for genre categorization with a total of 23 sports types, it is predicted that the codebook size should be bigger than in the tested 14 sports case. Therefore, a codebook size of 1600 is chosen, and a codebook size of 800 is also applied as a comparison analysis. For the view classification involving 14 sports, a codebook size of 800 is selected.

### 7.1 Genre categorization using a K-nearest neighbor (k-NN) classifier

In genre categorization, a K-nearest neighbor (k-NN) classifier is applied. Three different dissimilarity measurements are compared, including Euclidian distance (ED), earth mover's distance (EMD), and Kullback-Leibler divergence (KL-div). ED is used for measuring the spatial distance in Euclidian space in between two histograms. EMD is a distance function to achieve the minimal cost in transforming one histogram into the other [Rubner et al. 2000]. The KL-div is a non-symmetric measurement between two probability distributions  $Q$  and  $P$ , defined as  $D_{KL}(Q||P) = \sum_i q_i \cdot \ln(q_i/p_i)$  [Duda et al. 2001]. In this work,  $q_i$  and  $p_i$  are individual codewords for the query video  $Q$  and the trained genre model  $P$ , respectively.

Before the accuracy performance analysis on genre categorization, codebook generation schemes are examined by comparing both the proposed two-level bottom-up (BU) structure and the baseline single K-means (SK) clustering method. As pointed by Jain *et al.*[1999], K-means clustering is considered as a partitional algorithm using the squared error to reach the optimum solution. Sum of squared errors (SSE) is a widely used criterion function for clustering analysis, which quantitatively measures the total difference between all individual points to their clustering centers [Duda et al. 2001]. A SSE deviation percentage  $\delta_{dev}$  is defined in Equation 16. Let  $\xi_{BU}$  and  $\xi_{SK}$  represent the SSEs of the bottom-up based clustering and the single K-means clustering at the end of each algorithm, respectively. The numerator is the absolute value of the difference between  $\xi_{BU}$  and  $\xi_{SK}$ , and the denominator is  $\xi_{SK}$ . As Table IV shows, the SSE deviation percentages at codebook sizes of 800 and 1600 are 1.4% and 3.7%, respectively. Thus we can conclude that using the bottom-up structure instead of the single K-means clustering for codebook generation, the deviation of SSE is trivial.

$$\delta_{dev} = \frac{|\xi_{BU} - \xi_{SK}|}{\xi_{SK}} \cdot 100\% \quad (16)$$

Table IV. SSE deviation percentage  $\delta_{dev}$  and computation time (hours) in codebook generation, using bottom-up (BU) and single K-means (SK) structures.

Codebook Size	cb <sub>BU</sub> =800	cb <sub>SK</sub> =800	cb <sub>BU</sub> =1600	cb <sub>SK</sub> =1600
$\delta_{dev}$	1.4 %		3.7 %	
Computation	4hrs	350hrs	9hrs	648hrs

Codebook computation effort of the bottom-up structure is also compared with single K-means clustering in Table IV. Both bottom-up and single K-means clustering are employed on a single Quad CPU at 2.40GHz with 4.0G RAM machine, in which the bottom-up is only simulated as parallel computing in a serial sequence. To generate a codebook with size 800, the single K-means clustering uses 350 hours, while the bottom-up based clustering only takes 4 hours. When the codebook size is doubled to 1600, the computation for single K-means and bottom-up based clustering are 648 hours and 9 hours respectively. With a truly distributed processing environment using multiple computers, bottom-up based processing time will be further reduced. This demonstrates that our generic framework using robust bottom-up based clustering for codebook generation can replace the single K-means in dealing with large-scale and diverse datasets.

For the accuracy performance using k-NN and various dissimilarities, Table V shows the average genre categorization results for 23 different sports. The proposed bottom-up codebook generation manifests a better and more robust performance than single K-means codebook generation in both EMD and KL-div measurements. By comparing the row-wise various dissimilarities, the bottom-up structure is more consistent with codebook sizes of 800 and 1600. On the contrary, the single K-means based codebook generation is unstable for both histogram and mLDA based distributions. For instance, the performance at a codebook size of 800 using EMD has about 7% increment from ED dissimilarity (75.33% vs. 68.31%), while the counterpart at a codebook size of 1600 using EMD has dropped 1.1% from ED dissimilarity (64.28% vs. 65.39%). One reason is that the single K-means clustering on over 3 million input SIFT points hardly reaches the optimal value. As a summary, KL-div performs the best among three dissimilarity measures. Using the bottom-up structure, results of the codebook size 1600 outperforms the cases with the size of 800 in all measurements with consistency. Oppositely, single K-means clustering results are not consistent.

Another merit of the bottom-up structure is its preservation of individual genre characteristics from the 1<sup>st</sup>-level K-means. On the contrary, single K-means codebook generation covers all the data, thus a weakly distinguishable genre is easily overruled by a strong one. This explains why with the increase of codebook size from 800 to 1600, the bottom-up process has about 4% improvement for KL-div, while the single K-means process has only 2% increment for KL-div.

The individual sport genre classification result is illustrated in Figure 7. In average, a codebook size of 1600 gives an average of 3.6% higher than the codebook size of 800, which agrees with the empirical studies from other research groups [Jiang et al. 2010; Yang et al. 2007].

To evaluate the generic and extensive properties of our proposed approach, ex-

Table V. Average categorization results (%) of 23-sports data with codebook size 800 and 1600. (BU: codebook generated using bottom-up structure. SK: codebook generated using single K-means structure.)

Measurement	ED	EMD	KL-div
$\text{cb}_{\text{BU}}=800$	61.54	75.80	78.59
$\text{cb}_{\text{SK}}=800$	68.31	75.33	73.49
$\text{cb}_{\text{BU}}=1600$	65.68	78.94	82.16
$\text{cb}_{\text{SK}}=1600$	65.39	64.28	75.75

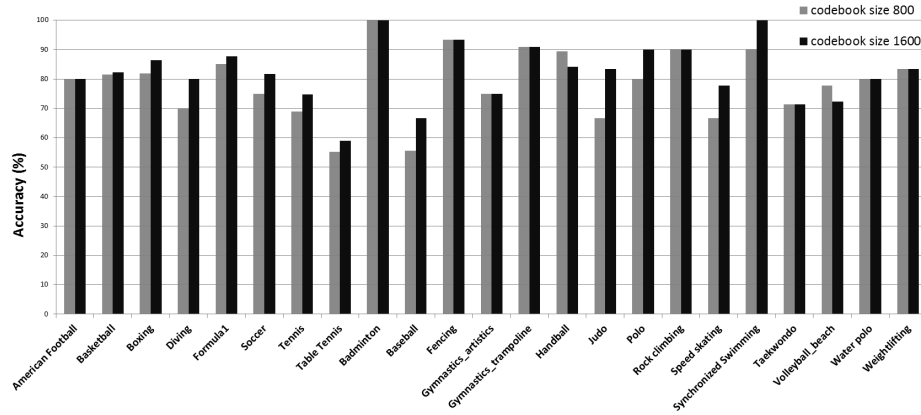


Fig. 7. Genre categorization for the 23-sports dataset, with codebook size of 800 and 1600.

periment results on 23-sports dataset are compared with results in Li *et al.*'s work [2009], where a top-down process was used with single K-means as its top layer general codebook. The best performance in a two-layer and a single-layer structures are 83.83% and 81.2% respectively [Li et al. 2009]. In their work, a speeded up robust features (SURF) based method is adopted. Similar to SIFT, SURF is also a scale and rotation-invariant interesting point feature extraction algorithm, which focuses on the computational efficiency [Bay et al. 2006]. Although SURF and SIFT adopt different key points detection techniques, these two descriptors are comparable in characterizing local features of sampled frames from a video sequence. Therefore, such a comparison is valid in genre categorization performances, regardless of the feature extraction difference. Considering the increment of data in scale about 27% (145hrs *vs.* 114.2hrs), while in diversity about 64% (23 genres *vs.* 14 genres), using the bottom-up structure with a codebook size of 1600 and KL-div measurement, our experimentation provides comparable results of 82.16%, with a degradation of 1.67%.

Although the performance is maintained averagely, we also observed that the individual performance has been fluctuant. This is mainly due to the nature of the adopted k-NN classifier, where distance-based measurement can be overruled by a strong representation in large and sparse dataset. We acknowledge such a fact that k-NN may not be the most robust approach towards the very large-scale dataset. However, the k-NN is an efficient method in batch processing. It can be used as



Table VI. Genre categorization accuracy between various video clips with the uniform sampling based and the key-frame/shot based methods.

<i>3 Minutes Clip</i>		<i>10 seconds Clip</i>	
<b>Uniform Sampling</b>	<b>Key-frame/Shot</b>	<b>Uniform Sampling</b>	<b>Key-frame/Shot</b>
83.83%	79.41%	71.90%	63.10%

a coarse and preliminary execution to quickly prune off the large portion of the irrelevant data.

From a different perspective, generic property of the proposed approach is assessed using various video clip lengths and frame sampling methods. As detailed in Table VI, better performance is acquired using longer length of video clips, while a generic and automatic uniform sampling method outperforms the key-frame based sampling. This is because the proposed approach is based on local key-point descriptors. Therefore, a longer video clip with denser sampling frames provides more key-points and consequently builds a better distribution than a shorter clip with less sampled key-frames/shots. Such an experimentation demonstrates the merit of proposed generic approach towards a truly large-scale dataset.

## 7.2 View classification analysis using supervised SVM and unsupervised PLSA

Experiments in this part focus on middle-level view classification, utilizing extracted low-level histogram based representations. A subset of 14 sports of all 23 sports were used as test data which is detailed in the column plot 0???. Figure 8 compares both supervised SVM and unsupervised PLSA results as the 1st and 2nd columns respectively. In average, supervised SVM has a classification accuracy of 82.86%, and unsupervised PLSA has an average of 68.13%, in which the SVM technique outperforms the PLSA approach with 14.73%.

It needs to be pointed that the above evaluation is based on the pre-determined semantic view types which is in favor of the SVM approach in nature. This is because such a semantic definition has considerably involved in the SVM training, while barely used in the PLSA training. In the SVM method, labeled training data associated with each pre-defined view types are indispensable for building the classifier. On the other hand, the PLSA model training merely requires a specified number of view types, which is similar as the number of clusters needed for training a K-means clustering. Thus it is anticipated that the supervised SVM method has better performance than the unsupervised PLSA algorithm.

However, the PLSA model is advanced in its unsupervised characteristics such that the labeled data is avoidable in training. This feature makes the PLSA more suitable than the SVM and significant in supporting the generic framework dealing with large-scale datasets, where automatic process and minimum human and expertise interventions are essential. For evaluating our proposed framework, a trade-off in the classification accuracy can be afforded, if the ultimate event detection results are comparable using either the PLSA or the SVM view results.

In order to analyze the generic and scalable property, a subset with small-scale 5-sports dataset is applied including {soccer, basketball, volleyball, table tennis, tennis}. The SVM and PLSA view classification performance of this small-scale dataset is presented at the first 5 sports with 3rd/4th columns of the Figure 8,

respectively. Baseline with the small-scale data, the 14-sports has 0.27% performance drops in SVM, while improved 1.76% in PLSA. With the similar results, comparing with the 5-sports small-scale data, 14-sports view dataset has a lot more data in both variety and volume.

Based on the above analytical results, the extrapolated performance from current relatively large-scale to a truly large-scale dataset should be maintained especially for the PLSA method. The reasoning of this claim is two-fold. Firstly, large-scale data is normally sparse, PLSA as a generative model, has a characteristics in probabilistically mapping data from a high-dimensional space to a low-dimensional space. Hence more information brought by the new data can help in finding significant representatives in the lower dimensional space. Secondly, since the number of view classes are fixed 4 types, more variety and volume won't affect the performance much.

Additionally, a knowledge transfer property is investigated by using the same 5-sports dataset. It can be seen that individual sport from insufficient resources {basketball, volleyball, table tennis, tennis} can be assisted by borrowing the codebook from an abundant sport resource {soccer}. As Figure 8 depicts on these limited-source four sports at 5th/6th columns, the codebook transfer mechanism has improved about 2.07% and 5.05% for the SVM and PLSA in average, respectively. The margin of the improvement using the PLSA is bigger than the counterpart in the SVM. This can be explained by the nature of two different techniques. PLSA is a probabilistic-based dimensional reduction technique. Therefore, more data will provide a more thorough characterization of the low-dimensional model. On the contrary, SVM is a technique mapping from a low dimensional space to a higher dimensional space. More information brought by the codebook may be overwhelmed by the SVM process and not necessarily provides a better classification in the higher-dimensional space. Therefore, such a knowledge transfer property could help the unsupervised PLSA in further improving its performance for the sports of scarce resources.

### 7.3 Basketball score event detection using coarse-to-fine scheme and HCRF based structured prediction model

In previous experiments, the proposed framework provides an application to identify video genres by directly utilizing domain-knowledge free SIFT descriptors and a BoW model. After the genre is determined, individual frames of the query video sequence are labeled by the middle-level semantic views via either supervised or unsupervised classifiers. In this experiment, the task on basketball score event detection is investigated by employing this labeled video sequence. Two-staged coarse-to-fine scheme is adopted with firstly detecting scoreboard information change introduced by Miao *et al.*[2007]. By adopting this technique, an entry point of an interesting event is located. However, this coarse detection only provides a static frame based rough estimation as an entry point. Since scoreboard information not only appears in score events, but also in time-out events or intermission events, individual frame based detection without temporal structured information cannot provide robust and satisfactory result. Therefore, a fine tuning process in finalizing detection is adopted to ensure that the query video truly conveys the score event as its semantic theme. The proposed HCRF model is deployed as such process after the first stage

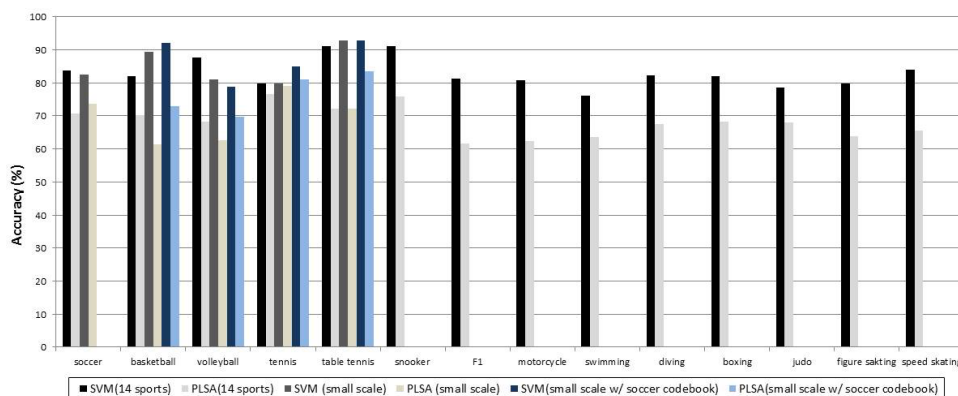


Fig. 8. View type classification using supervised SVM and unsupervised PLSA. First two columns are with the codebook size 800 for 14 sports. 3rd and 4th columns are SVM and PLSA performances of a smaller group with 5 sports {soccer, basketball, volleyball, table tennis, and tennis}. 5th and 6th columns are the SVM and PLSA performances of the above 5 sports data excluding the soccer sport. The difference of 5th/6th from 3rd/4th is that the generated codebook is borrowed from the abundant soccer sport.

Table VII. Precision and recall results of basketball score events detection at the first (coarse) stage.

Correctly Detected Score (true positive)	Detected Score (correct result)	Correct Total Score (obtained result)	Precision (%)	Recall (%)
231	251	268	92.03	86.19

coarse detection. Experimental results of using this HCRF model are compared with CRF and HMM baselines.

Two video groups consisting of four matches are utilized, which are defined as (a) Dataset A: using two NBA games for training, and using another two Olympic Games for testing; (b) Database B: using one NBA game for training, and using another NBA game for testing. Frame-based views from the PLSA model and the SVM model are applied to Dataset A and B. Therefore, four combinations of view labels and datasets are defined as  $PLSA+A$ ,  $PLSA+B$ ,  $SVM+A$ , and  $SVM+B$ . Each video clip used in both training and testing is automatically decimated and consists of 500 uniformly sampled frames. We use a window size  $N = 20$  which is introduced in Figure 5 and Equation 7 from section 6, with a window  $N$  sliding every 10 frames. The final number of the states sequence for HCRF is thus calculated as  $49 = 500 / (20 - 10) - 1$ .

The number of approximated events detected after the first stage is given in Table VII. The precision and recall of the coarse stage basketball score detection are 92.03% and 86.19% respectively. In the second stage, the proposed HCRF-based model and state of the art HMM and CRF models are evaluated and compared. The advantage of HCRF over HMM is its relaxation on the Markov property that the current state  $S_t$  can be inferred from both current observation as well as surrounding observations. This is illustrated in Figure 9. In the experiment, the

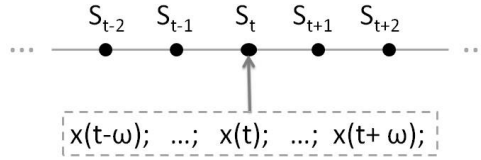


Fig. 9. Current state influenced by surrounding observed states.

Table VIII. Performance comparison on score event detection in basketball. Dataset A: NBA matches as training, Olympic matches as testing. Dataset B: NBA matches for both training and testing.

	Accuracy			
	Dataset A (NBA/Olympics)		Dataset B (NBA/NBA)	
	SVM+A (%)	PLSA+A (%)	SVM+B (%)	PLSA+B (%)
HMM $\omega = 0$	78.28	75.29	87.50	85.94
CRF $\omega = 0$	78.16	74.57	87.43	86.52
CRF $\omega = 1$	79.52	76.82	88.52	87.89
HCRF $\omega = 0$	80.93	75.53	90.00	90.77
HCRF $\omega = 1$	83.26	80.24	93.08	92.31
HCRF $\omega = 2$	82.09	77.88	91.46	91.77

circumferential range number is selected at  $\omega = 0, 1, 2$ . As shown in Table VIII, the HCRF has better performance than the CRF for the same  $\omega$  values, while both models outperform the HMM baseline. When using different  $\omega$  values for both CRF and HCRF,  $\omega = 1$  provides better results than  $\omega = 0$ , in which neighboring information assists in a better decision making. However, when  $\omega = 2$  is used for HCRF, the performance has been dropped for all cases comparing with  $\omega = 1$ . This can be viewed as an over-fitting issue, in which adding more surrounding information limits the structured prediction ability. A similar over-fitting problem is also observed in gesture recognition research using HCRF [Quattoni et al. 2007]. In summary, the proposed HCRF based model with parameter  $\omega = 1$  outperforms both CRF and HMM models. The best results are obtained at 93.08% and 92.31% by taking SVM and PLSA based input labels, respectively.

On the other hand by comparing the input of basketball videos, the performance discrepancy of event detection has been shortened as we compare column-wise SVM with PLSA in both datasets, although the input views after classification shown in Figure 8 has PLSA (70.14%) been outperformed by SVM (82.00%) for 11.86%. For dataset A, the average difference shows that SVM outperforms PLSA by 3.65%, while in dataset B, such a difference is only 0.47%. This demonstrates the robustness and resilience of structured prediction models in accommodating not well labeled video sequences from PLSA, yet achieving comparable performance as with input from SVM learning. Therefore, the event detection presented in this work achieves similar results by both unsupervised learning and supervised learning approaches. However, due to PLSA's much less human involvement, the unsupervised classifier is preferred in the large-scale video analysis.

Experimental result discrepancies using Dataset A and Dataset B are also compared. Although both datasets belong to basketball sport, Dataset B using NBA

matches for both training and testing outperformed dataset A with NBA matches for training but Olympics matches for testing, by 10.9% in average. It suggests that albeit from the fact which datasets A and B are with the same genre and event detection task, a significant difference exists. This can be explained by assuming NBA and international basketball (FIBA) are two different styles of a same genre. In terms of computer vision and structured prediction, NBA and FIBA have related but different temporal pattern even in the same semantic event. Thus, by training/testing in the same style, it is expected to have a better detection rate than training/testing using different styles. This is an example of the semantic gap that semantic event recognition with discrepant conditions is still not perfect.

Although there is only one event detection example discussed in this paper, It is believed that the approach can be extended and generalized to a bigger pool of the event scenarios. The reason is four-fold. Firstly, the experiment data of basketball score event is multi-source and non-simplex. Videos are collected from both internet and TV recordings, as well as different production rules of NBA and Olympics basketballs. Secondly, the video representation module using local feature and BoW model is domain-knowledge free and no production rules involved. Such generic approach has been approved to be effective in genre categorization of 23 sports, and view classification of 14 sports, and the basketball score event. Thirdly, the event detection approach utilizing HCRFs as well as baseline HMMs and CRFs are all structured prediction model and belong to the category of state event model. By comparing the number of the events analyzed using different event models from Table III, the state event model is a popular approach in recent years with a lot more events handled than other two model types. In addition, among the state event models, most methods utilize middle-level semantic agents. In our work, the adopted four-category view types definition is one of the most popular classification scheme in literature. Lastly and most importantly, the input of our event detection model is a sequence of labeled views which are the results of a domain knowledge free method (either PLSA or SVM), using a generic video representation. With a better accuracy achieved by the proposed HCRF based model than baselines HMM and CRF based models, the performance should be maintained with other labeled sequences which could from various event scenarios. Moreover, utilizing sequences labeled by the middle-level agents as input, are also popular among the peers' works with state event models [Tong et al. 2004; Wang et al. 2006; Zhang et al. 2007; Xu et al. 2008].

## 8. CONCLUSION

This paper introduces a generic framework for analyzing a relatively large-scale diverse sports video dataset, with three video analysis tasks in a coherent and sequential order. By processing all data indifferently at the feature extraction stage using domain-knowledge free local SIFT descriptors, various video data are represented by compact and concise BoW models. Then a systematic approach is employed for event detection targeting on a query video sequence, which may embody an interesting event. In this approach, after its genre identified firstly using k-NN classifier, the query video is evaluated with semantic views assignment as the second stage with the PLSA model. Both tasks utilize the initially processed

video representation as input. Finally in the third task, the interesting event is detected by feeding the view labels into a HCRF structured prediction model.

Overall, this framework demonstrates the efficiency and generality in processing voluminous data from a relatively large-scale sports collection and achieves various tasks in video analysis. The affectiveness of the framework is justified by extensive experimentation and results are compared with benchmarks and state of the art algorithms. As a conclusion, with little human expertise and effort involvement in both domain-knowledge independent video representation and annotation free unsupervised view labeling, the proposed generic and systematic approach is promising in processing sports video dataset, with a potential to be extended to real large-scale and diversified dataset.

Our future work will focus on expanding the current dataset to a truly large-scale in size and various-type in diversity, so that the proposed approach can be examined in a more complicated proving ground. We also will conduct more experiments on event scenarios other than the score event, as well as related high-level semantic analysis, such as tactic analysis, automatic broadcast video generation and etc.

## 9. ACKNOWLEDGEMENT

This work was supported in part by grants from Canada Research Chair Program, Canada Foundation for Innovations, Ontario Research Funds; as well as from the Chinese National Natural Science Foundation under contract No. 60902057, in part by the National Basic Research Program of China under contract No. 2009CB320902, and in part by the CADAL Project Program and the NLPR Open Project Program.

## REFERENCES

- BABAGUCHI, N., KAWAI, Y., AND KITAHASHI, T. 2002. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia* 4, 1, 68–75.
- BAY, H., TUYTELAARS, T., AND VAN GOOL, L. 2006. Surf: Speeded up robust features. *Lecture notes in computer science* 3951, 404.
- BENMOKHTAR, R., HUET, B., AND BERRANI, S. 2008. Low-level feature fusion models for soccer scene classification. In *Proc. IEEE ICME*. 1329–1332.
- BOHN, R. AND SHORT, J. 2010. How much information? 2009 Report on American Consumers. *University of California at San Diego, Global Information Industry Center*.
- BREZEALE, D. AND COOK, D. 2008. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 38, 3, 416–430.
- CHANG, C. AND LIN, C. 2001. LIBSVM: a library for support vector machines.
- DAI, J., DUAN, L., TONG, X., XU, C., TIAN, Q., LU, H., AND JIN, J. 2005. Replay scene classification in soccer video using web broadcast text. In *Proc. IEEE ICME*. 1098–1101.
- DUAN, L., XU, M., CHUA, T., TIAN, Q., AND XU, C. 2003. A mid-level representation framework for semantic sports video analysis. In *Proc. ACM MM*. 33–44.
- DUAN, L., XU, M., AND TIAN, Q. 2003. Semantic shot classification in sports video. In *Proc. SPIE*. 300–313.
- DUDA, R., HART, P., AND STORK, D. 2001. *Pattern classification*. Wiley-Interscience.
- EKIN, A. AND TEKALP, A. 2002. Framework for tracking and analysis of soccer video. In *Proc. SPIE VCIP*. Vol. 4671. 763–774.
- EKIN, A., TEKLAP, A. M., AND MEHROTRA, R. 2003. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing* 12, 7, 796–807.
- FISCHER, S., LIENHART, R., AND EFFELSBERG, W. 1995. Automatic recognition of film genres. In *Proc. ACM MM*. Vol. 95. 295–304.

- GLASBERG, R., SCHMIEDEKE, S., MOCIGEMBA, M., AND SIKORA, T. 2008. New Real-Time Approaches for Video-Genre-Classification Using High-Level Descriptors and a Set of Classifiers. In *Proc. IEEE ICSC*. 120–127.
- GUNAWARDANA, A., MAHAJAN, M., ACERO, A., AND PLATT, J. 2005. Hidden conditional random fields for phone classification. In *Proc. Interspeech*. 1117–1120.
- HOFMANN, T. 1999. Probabilistic latent semantic indexing. In *Proc. ACM SIGIR*. 50–57.
- HOFMANN, T. 2000. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. *NIPS* 12, 914–920.
- JAIN, A., MURTY, M., AND FLYNN, P. 1999. Data clustering: a review. *ACM computing surveys (CSUR)* 31, 3, 264–323.
- JASER, E., KITTLER, J., AND CHRISTMAS, W. 2004. Hierarchical decision making scheme for sports video categorisation with temporal post-processing. In *Proc. IEEE CVPR*. Vol. 2. 908–913.
- JIANG, Y., YANG, J., NGO, C., AND HAUPTMANN, A. 2010. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia* 12, 1, 42–53.
- KOLEKAR, M. AND PALANIAPPAN, K. 2009. Semantic concept mining based on hierarchical event detection for soccer video indexing. *Journal of Multimedia* 4, 5, 298–312.
- LAFFERTY, J., MCCALLUM, A., AND PEREIRA, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*. 282–289.
- LAVEE, G., RIVLIN, E., AND RUDZSKY, M. 2009. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 39, 5 (Sept.), 489–504.
- LAY, J. AND GUAN, L. 2006. Semantic retrieval of multimedia by concept languages: treating semantic concepts like words. *IEEE Signal Processing Magazine* 23, 2, 115–123.
- LAZEBNIK, S., SCHMID, C., AND PONCE, J. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proc. IEEE CVPR*. Vol. 2. 2169–2178.
- LI, L., CHEN, Y., HU, W., LI, W., AND ZHANG, X. 2009. Recognition of Semantic Basketball Events Based on Optical Flow Patterns. In *Proc. ISVC*. Springer, 480–488.
- LI, L., ZHANG, N., DUAN, L., HUANG, Q., DU, J., AND GUAN, L. 2009. Automatic sports genre categorization and view-type classification over large-scale dataset. In *Proc. ACM MM*. 653–656.
- LOWE, D. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of computer vision* 60, 2, 91–110.
- MEI, T. AND HUA, X. 2008. Structure and event mining in sports video with efficient mosaic. *Multimedia Tools and Applications* 40, 1, 89–110.
- MIAO, G., ZHU, G., JIANG, S., HUANG, Q., XU, C., AND GAO, W. 2007. A Real-Time Score Detection and Recognition Approach for Broadcast Basketball Video. In *Proc. IEEE ICME*. 1691–1694.
- MONTAGNUOLO, M. AND MESSINA, A. 2009. Parallel neural networks for multimodal video genre classification. *Journal of Multimedia Tools and Applications* 41, 1, 125–159.
- MORENCY, L., QUATTONI, A., CHRISTOUDIAS, C., AND WANG, S. 2008. Hidden-state Conditional Random Field Library.
- NEPAL, S., SRINIVASAN, U., AND REYNOLDS, G. 2001. Automatic detection of ‘Goal’ segments in basketball videos. In *Proc. ACM MM*. 261–269.
- PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J., AND ZISSERMAN, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proc. IEEE CVPR*. Vol. 3613. 1575–1589.
- QUATTONI, A., WANG, S., MORENCY, L., COLLINS, M., DARRELL, T., AND CSAIL, M. 2007. Hidden-state conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 10, 1848–1852.
- RUBNER, Y., TOMASI, C., AND GUIBAS, L. 2000. The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision* 40, 2, 99–121.

- SADLIER, D. AND O'CONNOR, N. 2005. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 10, 1225–1233.
- SHA, F. AND PEREIRA, F. 2003. Shallow parsing with conditional random fields. In *Proc. of HLT-NAACL*. 213–220.
- SIVIC, J. AND ZISSERMAN, A. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*. Vol. 2. 1470–1477.
- TAKAGI, S., HATTORI, S., YOKOYAMA, K., KODATE, A., AND TOMINAGA, H. 2003. Sports video categorizing method using camera motion parameters. In *Proc. IEEE ICME*. Vol. 2. 461–464.
- TAN, Y., SAUR, D., KULKARNI, S., AND RAMADGE, P. 2000. Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Transactions on Circuits and Systems for Video Technology* 10, 1, 133–146.
- TIEN, M., WANG, Y., CHOU, C., HSIEH, K., CHU, W., AND WU, J. 2008. Event detection in tennis matches based on video data mining. In *Proc. IEEE ICME*. 1477–1480.
- TONG, X., LIU, Q., LU, H., AND JIN, H. 2004. Shot classification in sports video. In *Proc. ICSP*. Vol. 2. 1364–1367.
- TONG, X., LU, H., AND LIU, Q. 2004. A three-layer event detection framework and its application in soccer video. In *Proc. IEEE ICME*. Vol. 3. 1551–1554.
- TRUONG, B., DORAI, C., AND VENKATESH, S. 2000. Automatic genre identification for content-based video categorization. In *Proc. ICPR*. Vol. 15. 230–233.
- WANG, J., CHNG, E., AND XU, C. 2005. Soccer replay detection using scene transition structure analysis. In *Proc. IEEE ICASSP*. 433–437.
- WANG, J., XU, C., AND CHNG, E. 2006. Automatic sports video genre classification using pseudo-2d-hmm. In *Proc. ICPR*. 778–781.
- WANG, P., LIU, Z., AND YANG, S. 2007. Investigation on unsupervised clustering algorithms for video shot categorization. *Journal of Soft Computing-A Fusion of Foundations, Methodologies and Applications* 11, 4, 355–360.
- WANG, S., QUATTONI, A., MORENCY, L., DEMIRDJIAN, D., AND DARRELL, T. 2006. Hidden conditional random fields for gesture recognition. In *Proc. IEEE CVPR*. 1521–1527.
- WANG, T., LI, J., DIAO, Q., HU, W., ZHANG, Y., DULONG, C., AND BEIJING, P. 2006. Semantic event detection using conditional random fields. In *Proc. IEEE CVPRW*. 109–114.
- XU, C., WANG, J., WAN, K., LI, Y., AND DUAN, L. 2006. Live sports event detection based on broadcast video and web-casting text. In *Proc. ACM MM*. 230.
- XU, C., ZHANG, Y., ZHU, G., RUI, Y., LU, H., AND HUANG, Q. 2008. Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia* 10, 7, 1342–1355.
- XU, L. AND LI, Y. 2003. Video classification using spatial-temporal features and PCA. In *Proc. IEEE ICME*. Vol. 3. 485–488.
- XU, M., DUAN, L., XU, C., AND TIAN, Q. 2003. A fusion scheme of visual and auditory modalities for event detection in sports video. In *Proc. IEEE ICASSP*. Vol. 3. 189–192.
- XU, P., XIE, L., CHANG, S., DIVAKARAN, A., VETRO, A., AND SUN, H. 2001. Algorithms and system for segmentation and structure analysis in soccer video. In *Proc. IEEE ICME*. 928–931.
- YAN, R. AND HSU, W. 2009. Content-based and concept-based analysis for large-scale image/video retrieval. In *Proc. ACM MM*. 913–914.
- YANG, J., JIANG, Y., HAUPTMANN, A., AND NGO, C. 2007. Evaluating bag-of-visual-words representations in scene classification. In *Proc. ACM MIR*. 197–206.
- YE, Q., HUANG, Q., GAO, W., AND JIANG, S. 2005. Exciting event detection in broadcast soccer video with mid-level description and incremental learning. In *Proc. ACM MM*. 455–458.
- YUAN, X., LAI, W., MEI, T., HUA, X., WU, X., AND LI, S. 2006. Automatic video genre categorization using hierarchical svm. In *Proc. IEEE ICIP*. 2905–2908.
- ZHANG, D. AND CHANG, S. 2002. Event detection in baseball video using superimposed caption recognition. In *Proc. ACM MM*. 315–318.



- ZHANG, J., MARSZALEK, M., LAZEBNIK, S., AND SCHMID, C. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision* 73, 2, 213–238.
- ZHANG, Y., XU, C., RUI, Y., WANG, J., AND LU, H. 2007. Semantic event extraction from basketball games using multi-modal analysis. In *Proc. IEEE ICME*. 2190–2193.
- ZHONG, L., LI, C., LI, H., AND XIONG, Z. 2008. Unsupervised Clustering Algorithm for Video Shots Using Spectral Division. In *Proc. ISVC*. Springer, 782–792.
- ZHU, G., XU, C., HUANG, Q., RUI, Y., JIANG, S., GAO, W., AND YAO, H. 2009. Event tactic analysis based on broadcast sports video. *IEEE Transactions on Multimedia* 11, 1, 49–67.