Automatic Sports Genre Categorization and View-type Classification over Large-scale Dataset

Lingfang Li[1,4], Ning Zhang[2], Ling-Yu Duan[3], Qingming Huang[1,4], Jun Du[1,5], Ling Guan[2]
[1]Key Lab of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences,Beijing 100080, China
[2] Ryerson Multimedia Research Laboratory, Ryerson University, Toronto, Ontario, Canada
[3]Institute of Digital Media, Peking University, Beijing 100871 P.R. China
[4]Graduate University of Chinese Academy of Sciences, Beijing, 100190 P.R. China
[5]NEC Research Labs China, Beijing, 100084 P.R. China
{lfli,qmhuang}@jdl.ac.cn {n2zhang,lguan}@ee.ryerson.ca
lingyu@pku.edu.cn dujun@research.nec.com.cn

ABSTRACT

This paper presents a framework with two automatic tasks targeting large-scale and low quality sports video archives collected from online video streams. The framework is based on the bag of visual-words model using speeded-up robust features (SURF). The first task is sports genre categorization based on hierarchical structure. Following on the second task which is based on automatically obtained genre, views are classified using support vector machines (SVMs). As a consequence, the views classification result can be used in video parsing and highlight extraction. As compared with state-of-the-art methods, our approach is fully automatic as well as domain knowledge free and thus provides a better extensibility. Furthermore, our dataset consists of 14 sport genres with 6850 minutes in total. Both sport genre categorization and view type classification have more than 80% accuracy rates, which validate this framework's robustness and potential in web-based applications.

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

Keywords

Genre categorization, Scene classification, Codebook

1. INTRODUCTION

Sports video categorization has been an active research area for many years. Some researchers try to classify sports video from other video sources. For instance, Huang *et al.* proposed a method to categorize basketball, football as well as news, weather forecast and commercials based on color, motion and audio features [1]. In the meantime, other researchers focused on categorizing sports video, as shown in Table 1 [3]. The number of sports, source of videos, size of databases, as well as classification methods and performances are listed. Domain knowledge related features have been used by all groups, either through color [11, 12, 7, 13], camera motion [10] or cues/objects [8]. Based on the claimed results, those methods have been proven very effective in small number of genres and datasets. However, as the dataset and diversity of sport genres become greater, maintaining a high performance using domain knowledge would be difficult and sometimes impossible. In addition, as Table 1 indicates, most previous methods use shot or key-frames to represent the whole sequence. This is a drawback in processing large-scale database, since in order to extract key-frames or detect shots, automatic methods would introduce errors and influence robustness while manual process is heavy and unfeasible.

On the other hand, previous sport view type classification methods also suffer from the problem of utilizing too much domain knowledge. Ekin *et al.* applied color and object-based features to obtain shot/view types [6]. Duan *et al.* proposed a unified framework for semantic shot/view type classification to facilitate semantic analysis in sports videos [5]. Although good performances have been obtained in these methods, the extensibility and flexibility towards many different sports videos are limited. Moreover, in these works, prior knowledge about sport genre can only be obtained with human involvement. The realization of full automaticity has thus far failed.

In order to be efficient in processing a diverse as well as large-scale dataset, we propose a framework of automatic sports genre categorization and view-type classification based on extracting speeded-up robust features (SURF) of uniformly sampled images from decoded video frame sequences. As a consequence, both genre categorization and view classification tasks are domain knowledge free as well as fully automatical.

The contribution of this paper is two-fold. First, a dataset including 14 different sports with a total of 6850 minutes approximately is considered in our framework. Previous domain knowledge related methods would be difficult and sometimes impossible to process this large-scale dataset. Sec-

MM'09, October 19–24, 2009, Beijing, China.

Table 1: Summary of previous sports genre categorization (n/a: not available)

Author and Reference	Number of Sports	Source of Videos	Size of Database (mins)	Domain Knowledge Features	Shot or Key-frame Based	Categorization Method	Accuracy rate
Truong & Dorai [11]	4	TV	480	Yes	Yes	C4.5 decision tree	83%
Wang *et al.* [12]	3	n/a	960	Yes	No	pseudo-2D-HMM	n/a
Takagi *et al.* [10]	6	TV	2025	Yes	Yes	statistics based	n/a
Gilbert *et al.* [7]	4	TV	220	Yes	No	two HMMs	93%
Jaser *et al.* [8]	4	n/a	n/a	Yes	Yes	decision tree and HMM	91.6%
Yuan *et al.* [13]	6	TV	2000	Yes	Yes	hierarchical SVM	94%

ondly, all video data are from online streaming with high compression and low quality, which makes it more difficult for them to be analyzed compared to regular ones. By working with these type of videos, we want to demonstrate a comparable result with those using higher quality data such as from TV broadcastings. Our methods can then be used in practical web-based applications (such as retrieval, browsing, etc), with efficient data storage and streaming server management.

The proposed framework of genre categorization and scene classification is presented in the following section. Experimental results and discussions are shown in section 3. Finally, the paper is concluded in section 4.

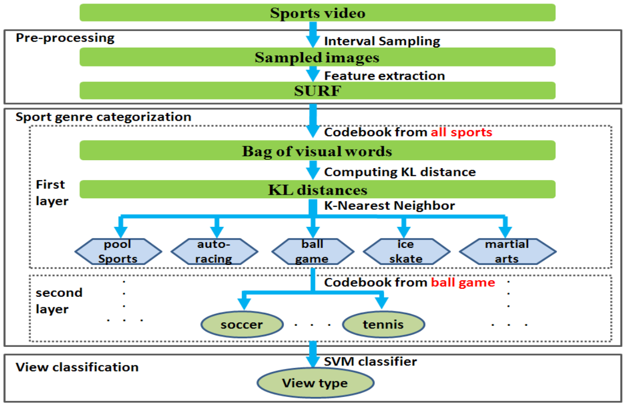

Figure 1: The proposed framework, with detailed two-layer hierarchical structure for sports genre categorization.

2. FRAMEWORK OF OUR METHODS

The proposed framework is depicted in Figure 1, where a query sport video has gone through three processes in sequence. In the first preprocessing stage, the video is automatically decimated by uniform sampling, with SURF descriptors then extracted. In the second sports categorization stage, a hierarchical categorization scheme is proposed to compare metric differences between query video and training videos for each sport genre, based on their designated codebooks. After the genre of query video is successfully discovered, at last stage, a supervised support vector machines (SVMs) model is applied to classify view types based on cinematographic definition. In theory, unsupervised model such as probabilistic latent semantic analysis (PLSA) can also be used to obtain non-predefined views automatically even without sport genre information. In section 2.3 and section 5, we will discuss about why we prefer SVM than PLSA in detail.

2.1 Pre-processing

In our framework, SURF is used in extracting local significance for each frame due to its accelerated speed and comparable performance to conventional Scale Invariant Feature Transform (SIFT) descriptors [2, 9]. After obtaining SURF from training video, a codebook is generated using k-mean clustering and each codeword value is defined by the exemplar vector of each cluster. By mapping individual frame SURF descriptors to a codebook, each frame can be represented by a histogram, which shows appearance frequency of each codeword in the codebook. Clearly, this codebook based representation is domain knowledge free and thus can be extended to large-scale dataset. Detailed experiments and discussions on codebook generation and selection mechanisms are given in section 5.

2.2 Sports Genre Categorization

A heuristic two-layer hierarchical structure is proposed based on sports genre semantic category. For instance, soccer, basketball and volleyball are all considered as a large sports group, i.e. the ball game. A detailed category of group sports and individual sport is listed in Table 2. As Figure 1 illustrates, in the first layer, each query video is classified into a bigger group of either pool sports, auto-racing, ball game, ice skate or martial arts. The codebook for this layer is based on a composition of all 14 different sports. In the second layer, a sports group based categorization takes place to further classify what individual sport this query video belongs to. In this part, codebook is generated by sports only from this group. From information theory, Kullback-Leibler (KL) divergence is a natural and effective distance measure between two probability distributions. Since query video codeword distribution Q is obtained from online by combining histograms of frames sequence and training category codeword distribution T is offline, in both layers, we utilize KL divergence to measure the difference between them, as equation 1 shows. Parameter i is the index of codeword in codebook. After that, the k-nearest neighbor (k-NN) algorithm is used to categorize genre based on KL divergence. Compared with other machine learning classifiers shown in table 1, K-NN is effective and easy to realize. Furthermore, K-NN is a prior knowledge free method which can efficiently work at a diverse scale of sports genre.

$$D_{KL}(Q||T) = \sum_i [q_i \cdot \log(q_i/t_i)] \tag{1}$$

2.3 View Classification

In the view classification task, four view types including close-up, mid-view, long-view and outer-field-view are classi-

Table 2: Group and individual sport with number of video clips

Group Sports	*Ball Game*						*Auto-Racing*	
Individual Sport	Soccer	Basketball	Volleyball	Tennis	Table-Tennis	Snooker	F1	Motorcycling
Num of Videos	94	98	89	106	85	87	84	75
Group Sports	*Pool Sports*		*Martial Arts*		*Ice Skate*		X	
Individual Sport	Swimming	Diving	Boxing	Judo	Figure Skate	Speed Skating	x	x
Num of Videos	94	70	88	85	83	76	x	x

fied. Either supervised or unsupervised classifier can be employed to perform view classification. Supervised classifiers tend to yield better results at the cost of prior knowledge about sport genre and human labeling work. Comparably, unsupervised classifiers are independent from genre information and involve much less labeling work, but at the risk of worse performances.

In our framework, prior knowledge about sport genre has been obtained in the first task and human labeling work for training is on a small scale. In such cases, supervised methods have more advantages compared to unsupervised methods. Since SVM has demonstrated a great performance in the field of classification in general and is able to achieve optimum solution only using a small training set, this method is adopted in our view classification task. A typical radial basis function (RBF) as the non-linear kernel is used in SVM [4] and shown in equation 2. In this equation, x_i and x_j represent codeword, and γ is the kernel parameter of RBF.

$$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \quad \gamma > 0. \tag{2}$$

To compare the performance, a recently popular unsupervised PLSA model is also used. PLSA relies on the likelihood function of multinomial sampling and aims at an explicit maximization of the predictive power of the model.

3. EXPERIMENT

A total of over 6850 minutes sports videos about 104G in size were obtained from online streams, in standard common intermediate format (CIF) at 352x288 resolutions. This dataset is consisted of 1214 video clips varying from 10 sec to 1 hour, in which 25% of data has 5 minutes or more length, 17% is around 3 minutes, and 25% around 1 minute, 17% for 30 seconds, and 16% for 10 seconds. A detailed sports genres and number of associated video clips can be referred in Table 2. By using the processing power of a duo-core PC with 1.86GHz CPU and 1G RAM, the average speed is about 15 seconds for each 5 minutes video clip in our experiment.

3.1 Codebook Analysis

The codebook is very crucial in our proposed framework. In this section, we will discuss the codebook in more details based on some experimental results. Table 3 illustrates one example of classification performance by incorporating different sports videos into the codebook. Performance is based on the view classification. This is because by checking the eligibility of codebook source, genre categorization is not applicable for individual sport, while view classification is an operational replacement.

It is well known that Table Tennis (TT) & snooker is more relevant than TT & judo. As can be seen in table 3, compared to using TT only, the codebook from TT & snooker provides same result. This shows that by incorporating snooker into the codebook, TT view classification accuracy is maintained. On the contrary, codebook source using TT & judo doesn't match the performance of using TT alone. We also consider a combination of TT and movies, which are not relevant at all. This generates the worst result of all, as predicted. The experimental fact suggests that the more relevant sports used in generating a codebook, the better accuracy results can be obtained. This is the reason that inspired us to use two-layer structure in genre categorization. In the first layer, codebook is based on a combination of all 14 sports videos, while in the second layer, each group codebook is generated by the most relevant sports which belong to this group. From Table 3, it is interesting to see that the accuracy rate by using codebook from TT & movies still remains above 85%. This not only demonstrates that codebooks from less relevant videos still have many similarities, but also proves that it is feasible to compose a single codebook using all genres in general categorization. This accounts for why we can employ a unified codebook for all 14 sports in the first layer categorization.

Table 3: Average accuracy using codebooks generated from different sports (TT: Table Tennis)

Codebok Sources	TT	TT + Snooker	TT + Judo	TT + Movie
Average Accuracy%	90.94	90.29	88.67	85.76

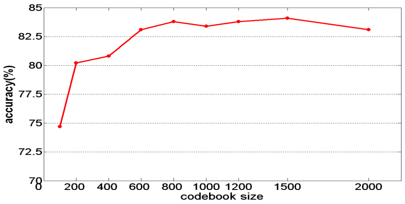

Figure 2: Accuracy performance of different codebook sizes.

3.2 Sport Genre

The very first and important step of sports genre categorization is the codebook size selection. Figure 2 illustrates the categorization accuracy rate with respect to the increase of first layer codebook sizes from 100 to 2000. The performance is boosted and saturated at 600 and then plateaus. While codebook size approaches 2000, the performance starts to drop. The initial boost is due to the lack of ability for a small codebook size to differentiate sports genre. The later performance decay after 2000 is due to the overfitting. As final result, we chose 800 as our codebook size for the first layer categorization. Based on experiments, it is also decided that the optimal codebook size for each individual group sports is 200 in the second layer.

Figure 3 presents precision and recall rate for all 14 sports

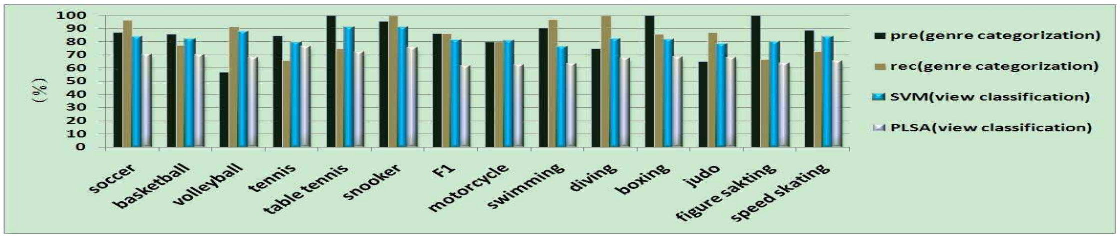

Figure 3: Precision and recall rate of sport genre categorization as well as view classification accuracy rate using SVM and PLSA models on 14 individual sport.

Table 4: Categorization accuracy between uniform sampling based and key-frame/shot based preprocessings

3 Minutes Clip		*10 seconds Clip*	
Uniform Sampling	**Key-frame\ Shot**	**Uniform Sampling**	**Key-frame\ Shot**
83.83%	79.41%	71.90%	63.10%

at our selective codebooks. As shown, volleyball, tennis and judo have low precision or recall rate, this is due to their ambiguous features which are easily confused with other sports. Most mis-categorization occurred because SURF had failed in representing the uniqueness of each individual sport. For example, tennis and basketball were misclassified as volleyball due to their similarities in field layout, camera angles, etc. The performances of the proposed two-layer approach and single layer method are also compared. On average over all sports, the two-layer structure has a good accuracy rate at 83.8%, which outperforms the single layer's accuracy rate at 81.2%.

Furthermore, another experiment was conducted based on comparing key-frame/shot-based with uniform sampling based preprocessing, as Table 4 presents. In both cases with video lengths of 3 minutes and 10 seconds, the final accuracy rate of uniform sampling outperforms the key-frame/shot-based method. This justifies that our uniform sampling based sports genre categorization method is not only more efficient in computation but also more effective in accuracy performance than key-frame/shot-based method which is often adopted in previous works.

3.3 View Type

Observed through a large number of experiments, we notice that the view classification performance is robust at codebook size from 100 to 1000. Figure 3 shows the classification results of SVM and PLSA, with codebook size at 100. In average, the accuracy rate of SVM is about 10% higher than PLSA for all 14 sports. The attractive feature of PLSA model is its unsupervised process, in which no prior knowledge of the sport genre is required. This appears as a drawback for the SVM model. However, by incorporating SVM with previous introduced two-layer genre categorization model, the categorized genre can guide view classification so that a completely automatic process is achieved for unknown video.

4. CONCLUSION

We have proposed a unified automatic framework to categorize sports genre as well as classify important view types by using SURF descriptors. This framework has been experimented on a large-scale, diverse dataset consisting 14 different sports genre with 6850 minutes in length. Considering the sampled frame quality due to the high compression rate of streaming videos, promising results have been obtained. The average accuracy rate for categorizing genres of all 14 sports is 83.8%, and the view classification accuracy rate is about 82.8%. In future, based on what we have achieved, we would like to develop some potential web-based applications such as retrieval, browsing, etc.

5. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China: 60773136 and 60833006, in part by National Hi-Tech Development Program (863 Program) of China: 2006AA01Z117, in part by National Basic Research Program of China (973 Program): 2009CB320902, and in part by Beijing Natural Science Foundation: 4092042.

6. REFERENCES

[1] J. Huang, et al., Joint scene classification and segmentation based on hidden Markov model. *IEEE Transactions on Multimedia*, Volume 7, Issue 3, pp. 538 - 550, 2005.

[2] H. Bay, et al., Surf: Speeded up robust features. *Lecture Notes in Computer Science*, 3951:404, 2006.

[3] D. Brezeale and D. Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics*, 38(3):416–430, 2008.

[4] C. Chang and C. Lin. LIBSVM: a library for support vector machines, 2001.

[5] L. Duan, et al., A unified framework for semantic shot classification in sports video. *IEEE Transactions on Multimedia*, 7(6), 2005.

[6] A. Ekin, et al., Automatic Soccer Video Analysis and Summarization. *IEEE Transactions on Image Processing*, 12(7), 2003.

[7] X. Gibert, et al., Sports video classification using HMMs. *IEEE International Conference on Multimedia and Expo*, pp. 345–348, 2003.

[8] E. Jaser, et al., Hierarchical decision making scheme for sports video categorisation with temporal post-processing. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 908–913, 2004.

[9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[10] S. Takagi, et al., Sports video categorizing method using camera motion parameters. *IEEE International Conference on Multimedia and Expo*, pp. 461–464, 2003.

[11] B. Truong and C. Dorai. Automatic genre identification for content-based video categorization. *International Conference on Pattern Recognition*, pp. 230-233, 2000.

[12] J. Wang, et al., Automatic sports video genre classification using pseudo-2d-hmm. *International Conference on Pattern Recognition*, pp. 778–781, 2006.

[13] X. Yuan, et al., Automatic video genre categorization using hierarchical SVM. *IEEE International Conference on Image Processing*, pp. 2905–2908, 2006.