An Online Learning Framework for Sports Video View Classification

Jun Wu[1*], Xian-Sheng Hua[2], Jian-Min Li[1], Bo Zhang[1], Hong-Jiang Zhang[2]

[1] Department of Computer Science and Technology, Tsinghua University,
Beijing 100084, China
wujun01@mails.tsinghua.edu.cn,
{lijianmin,dcszb}@mail.tsinghua.edu.cn

[2] Microsoft Research Asia, 5F Sigma Center, 49 Zhichun Road,
Beijing 100080, China
{xshua,hjzhang}@microsoft.com

Abstract. Sports videos have special characteristics such as well-defined video structure, specialized sports syntax, and some canonical view types. In this paper, we proposed an online learning framework for sports video structure analysis, using baseball as an example. This framework, in which only a very small number of pre-labeled training samples are required at initial stage, employs an optimal local positive model by sufficiently exploring the local statistic characteristics of the current under-test videos. To avoid adaptive threshold selection, a set of negative models are incorporated with the local positive model during the classification procedure. Furthermore, the proposed framework is able to be applied to real time applications. Preliminary experimental results on a set of baseball videos demonstrate that the proposed system is effective and efficient.

1 Introduction

Structure analysis is an elementary step for mining the semantic information in videos. Semantic structure parsing for general videos is difficult, while for sports videos, the task is much easier due to the well-defined temporal and spatial structures in this types of videos, as well as many domain specific rules can be applied.

In [1], three types of views (global, zoom-in and closeup) of soccer games are classified according to grass ratio of the video frames, and play/break statuses are detected as basic segments using heuristic rules. Hidden Markov model and dynamic programming are applied to play/break segmentation in [2]. Gong *et al* [3] classify soccer videos into various play categories based on a prior model consisting of four major components: a soccer court, a ball, the players and the motion vectors.

In some other research efforts, sports videos are parsed by detecting canonical views, such as serve view in tennis [4][5] and pitch view in baseball [6]. In [6], a simple "play" model is defined as the basic unit, and a general sports video summarization scheme is proposed based on the definition of "play".

Zhong *et al* [4] proposed a unified framework for scene detection and structure analysis by combining domain-specific knowledge with supervised machine learning methods. A verification step based on object and edge detection is used in order to obtain more reliable results. However, this framework has two major disadvantages.

[*] Part of the work was performed when the author was visiting Microsoft Research Asia as a student.

One is that good generalization capacity of the pre-trained models typically requires a sufficient amount of training data while this condition is difficult to be satisfied in most cases. The other disadvantage is that the best thresholds in classification procedure are varied for different videos. Adaptive threshold selection may help solve this issue to some extent, but the optimal thresholds are still difficult to be determined.

In this paper, a more intelligent and general online learning framework is proposed based on sufficiently exploring the local distribution properties of the video data, as well as incorporating negative models in the learning and classification procedures. This framework has the following unique features: 1) A dynamic local positive model is online calculated for the current under-analysis video by sufficiently exploring the local features in this video using a small amount of unlabeled samples. Furthermore, this online learning process is also able to be applied dynamically to update the local model periodically, thus making the classification results more reliable. 2) Negative models are sufficiently utilized to facilitate the determination of the view types, which makes the view classification results more robust and not sensitive to the thresholds. 3) The proposed system only requires a very small number of labeled training samples, e.g. about 10 to 40 samples.

The remainder of this paper is organized as follows. The next section is concerned with feature extraction and feature diversity analysis. In section 3, the online learning framework is proposed in detail. Experimental results are provided in section 4, followed by conclusion remarks and future works in the last section.

2 Feature Extraction and Feature Diversity Analysis

Correct classification of various kinds of views in sports video is essential for further content analysis such as event detection. In this paper, pitch view in baseball is taken as an example to describe and test our system. Due to camera motion is not prominent within typical pitch view shots, we only use the spatial features in the key-frames of each shot. It is observed that there are special distribution characteristics for grass and sand regions in pitch views. Accordingly, in the proposed system, the feature vectors are derived from the block-wise grass and sand distribution in the key-frames.

Firstly, each key-frame is divided into $N \times N$ blocks. Then grass and sand ratios are extracted from each block respectively. It is observed that typically, the backgrounds of pitch views, such as audiences or buildings, are generally at the top of video frames, though they are varied for different stadiums or channels. Therefore, several top rows of the key-frames are ignored to make the extracted features more accurate. We define an "ignored ratio" (denoted as "*IR*") as "the number of ignored rows" divided by "the total number of rows in the frame". For example, if *IR* is set to 0.5, only $((1-0.5)\times N\times N)$ blocks at the bottom of the currently processing key-frame are considered, as demonstrated in **Fig. 1**.

As aforementioned, though difficult, the optimal threshold for each under-test video is necessary to be determined adaptively, in order to obtain optimal classification results. To more clearly illustrate this issue, the optimal thresholds for a view classification method based on a simple supervised learning scheme are investigated below. Five baseball videos are used in this investigation, in which 1500 shots are taken as test data and totally there are 211 pitch view shots.

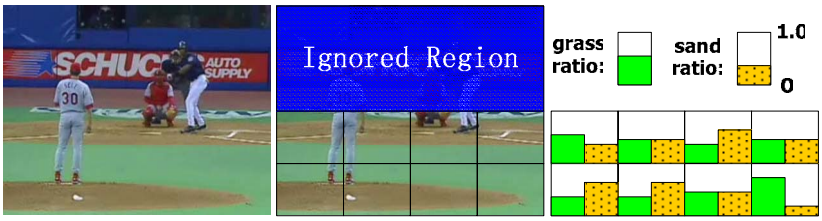

Fig. 1. Grass and sand distribution feature extraction. *Left*: an example of pitch view. *Middle*: This frame is divided into 4×4 blocks, and the two top rows are ignored (*IR*=0.5). *Right*: grass ratio (green block) and sand ratio (yellow block) are extracted from each block, respectively.

The major idea of the "simple learning" scheme is similar to the one in [4]. Firstly, the training samples of a specific view type are classified into a set of clusters, which results in a set of candidate view models represented by these cluster centers. Then, the key-frames of the shots in the initial segment (say, the first 10 minutes) of the under-test video are used to vote the above candidate models in order to select the best-matched model. Finally, in the classification procedure, the sample that is sufficiently close to the best-matched model is recognized as the corresponding view type.

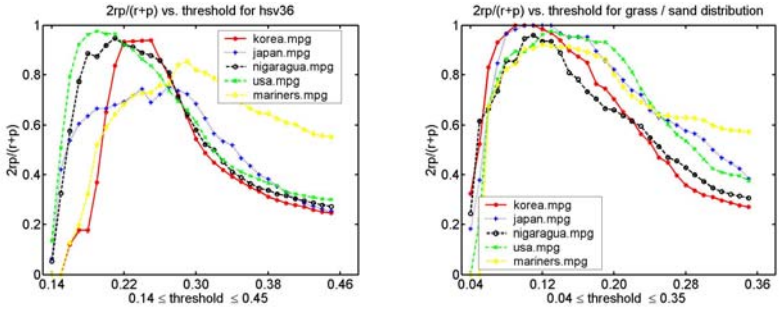

Fig. 2. Two sets of performance curves ($2rp/(r+p)$ vs. threshold) for the two types of features using the simple learning scheme. *Left*: color histogram is applied under the corresponding classification threshold ranging between 0.14 and 0.45. *Right*: grass/sand distribution (N = 4 and IR = 0.5) is employed under the classification threshold ranging between 0.04 and 0.35.

The features used here are quantized (36 bins) color histogram in HSV color space (Hereinafter referred as **hsv_36**) and grass/sand distribution (Hereinafter referred as **g_s_16**). And the performance is evaluated by $2rp/(r+p)$ since it is more discriminating than the average of p and r, where p denotes *precision* and r denotes *recall* [7]. **Fig. 2** shows the two sets of curves of $2rp/(r+p)$ under different thresholds for the five videos when using **hsv_36** and **g_s_16**, respectively.

From **Fig. 2**, it can be seen that the grass/sand distribution feature is better than color histogram in the above learning method. And for each under-test video, the performance of view classification is basically satisfactory under the best threshold. However, the optimal thresholds for the five videos are quite different and the uniform threshold is difficult to be determined when using either feature.

As to be explained in detail in next section, in our proposed framework, the diversity issue of the features and classification thresholds is overcome by adapting local classification models, instead of adaptively selecting optimal classification threshold.

3 An Online Learning Framework

To explore the local statistic properties of video data, we present an online learning framework for the view classification of sports videos, which dynamically learns the local statistics through the *reference samples* excerpted from part of current under-test video. Furthermore, to avoid adaptive threshold selection, two model sets (the *candidate positive model set* and *negative model set*) are trained by a simple voting process in which only a very small number of training samples are required. In the classification procedure, both a local positive model and a set of negatives are employed. As illustrated in **Fig. 3**, the training process consists of five primary steps.

Step 1. Clustering "reference samples": The key-frames of the shots from the initial part of the under-test video (unlabeled samples) are excerpted as *reference samples*. Standard K-mean clustering algorithm is employed on these samples so as to get a set of *reference clusters*, represented by their centers as $\{ C_1, \ldots, C_K \}$.

Step 2. Computing average inter-distance: For a specific view type, the pre-labeled training samples are clustered into another set of clusters, represented by their centers as $\{S_1, \ldots S_N\}$. The average inter-distance ($\overline{D}$) of these clusters is defined as

$$\overline{D} = \frac{1}{N^2} \sum_{i,j=1}^{N} Dist(S_i, S_j) , \quad (1)$$

where $Dist(S_i, S_j)$ denotes the distance between S_i and S_j. $\overline{D}$ will be used as a distance constraint in next step.

Step 3. Establishing "candidate positive model set": All the pre-labeled training samples are classified into the reference clusters formed in **Step 1** (by voting), among which the cluster with the largest voting number (denoted as M_0^C) is added to the candidate positive model set. And then, any other clusters that satisfy the following two conditions simultaneously are also inserted into the candidate model set.

(1) Its voting number is larger than a certain value, e.g. one percent of the number of total training samples. Otherwise, it is regarded as noise.

(2) The distance between the current cluster and M_0^C, $Dist(M_0^C, C_k)$, is smaller than $\alpha_1 \overline{D}$, where α_1 is a pre-defined coefficient and $\overline{D}$ is defined in Equation (**2**).

(*The distance constraint $\alpha_1 \overline{D}$ is specially designed to control the size of the candidate positive model set in order to make the learned view models more compact.*)

Step 4. Forming "negative model set": After the candidate positive model set is determined, all of the remaining reference clusters are considered as negative models, which are denoted as $\{ M_1^N, M_2^N, \ldots, M_H^N \}$.

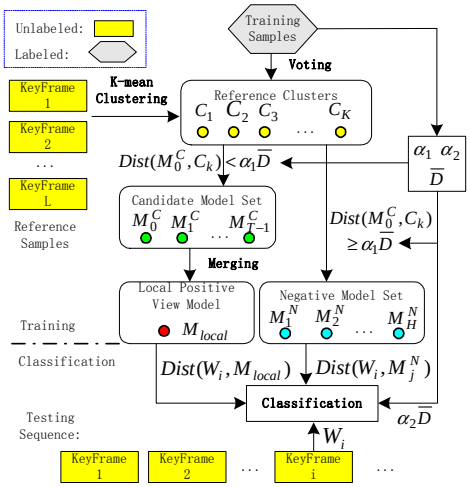

Fig. 3. Block diagram of the online learning framework. The upper part (separated by dash-dot line) shows the training procedure, and the lower part is the classification process.

Step 5. Calculating local model M_{local}: The local positive view model, M_{local}, is obtained by calculating the center of the merged cluster containing all the samples belonging to any of the candidate positive models.

In classification procedure, a view decision has high confidence when it is closer to the positive local model (M_{local}) than it is to the closest negative model. That is, the i-th under-test sample W_i is determined as the specific view type if (1) $Dist(W_i, M_{local}) < \min_{1 \le j \le H} Dist(W_i, M_j^N)$ and (2) $Dist(W_i, M_{local}) < \alpha_2 \overline{D}$, where α_2 is a predefined coefficient and $\overline{D}$ has been calculated in training procedure.

The parameters in the above algorithm, α_1 and α_2, are determined experimentally by balancing the *precision* and *recall* of view classification, as to be presented in the next section.

4 Experimental Results

Totally five baseball videos in MPEG-1 format from different stadiums, denoted as $\{V_1, V_2, V_3, V_4, V_5\}$, are used in our experiments. In this section, firstly the selection of parameters is investigated. And then, the performance of the proposed learning framework is evaluated, followed by some further discussions on the generalization capacity of the predefined parameters, the number of training samples, as well as the number of reference samples.

4.1 α_1 and α_2 Determination

To illustrate the generalization capacity of these parameters, the five videos are divided into two sets. One is the *validation data set*, consisting of V_1 and V_2, from

which the optimal values of α_1 and α_2 are determined. The other is the test data set to test the generalization capacity of the parameters, including V_3, V_4 and V_5.

As shown in **Fig. 4**, the highest value for average $2rp/(r+p)$ (denoted by coordinates z in the figure), 0.971, is obtained when setting $\alpha_1 = 1.3$ and $\alpha_2 = 1.0$. The performance on the test set under these optimal parameters is 0.957. It can be seen that the *precision* and *recall* are not very sensitive to the parameters, especially α_1. When α_1 is sufficiently large, the average $2rp/(r+p)$ only decreases about 2% compared with the best case. In section 4.3, more experimental results when choosing different validation data set will be presented.

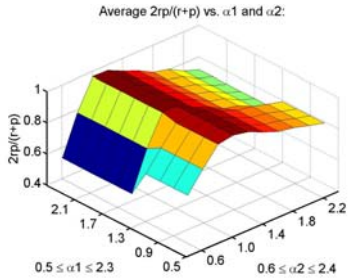

Fig. 4. Plot of the average of $2rp/(r+p)$ vs. α_1 and α_2 on the validation data when using grass/sand distribution as the classification feature, in which α_1 ranges between 0.5 and 2.3 (coordinates y) and α_2 between 0.6 and 2.4 (coordinates x). The increasing step is 0.2.

4.2 Pitch View Classification

In this part, the performance of the proposed online learning framework is evaluated when grass/sand distribution is selected as the feature vector, and the two parameters, α_1 and α_2, are set as the optimal values learned from the validation data set as mentioned in section 4.1.

Aiming at simulating the circumstance of insufficient training data, while each of the five videos takes its turn acting as test data, a small quantity of pre-labeled pitch view samples in the remaining four videos will act as training data. For example, to test V_1, the first 100 shots of V_1 are taken as the reference data and the next 300 shots as the test data. And the training data set consists of 40 pitch view samples randomly selected from the pre-labeled samples of other four videos (ten from each video).

In order to compare with our proposed framework, the performance of the simple learning scheme mentioned in Section 2 is also provided in **Table 1**. The classification threshold of this scheme is optimized by maximizing the classification performance on the validation data set mentioned in section 4.1.

From the table, it can be see that when using grass/sand distribution feature, the *precision* and *recall* of the online learning framework increases from 0.977 to 0.993 and from 0.847 to 0.924, respectively, comparing with the simple learning scheme. Furthermore, compared with color histogram, grass/sand distribution increases *preci-*

sion by 0.091 and *recall* by 0.146 for the simple learning scheme. Therefore, the grass/sand distribution feature is better than color histogram, while the online learning scheme is better than the simple learning scheme.

Table 1. Pitch view classification results for two methods: the simple learning scheme and the online learning framework. The total number of the pitch view shots in the test data set is 144. For the online learning framework, $\alpha_1 = 1.3$ and $\alpha_2 = 1.0$.

Method	Feature	Training Sample No.	Error	Miss	Precision	Recall
Simple Learning	**hsv_36**	400	13	43	0.886	0.701
Simple Learning	**g_s_16**	400	3	18	0.977	0.847
Online Learning	**g_s_16**	40	1	11	0.993	0.924

It should be noted that the online learning framework only uses 10% of the training samples required by the simple learning scheme, while achieves much better performance. As to be explained in section 4.3, the number of required training samples can be further reduced without much affecting the classification performance.

4.3 Discussions

In order to further investigate the generalization capacity of the parameters discussed in Section 4.1, we change the sizes of the validation and test data sets simultaneously as shown in **Table 2**.

Table 2. Pitch view classification performance of the online learning framework for different sizes of the validation data set and the test data set. The five videos are assigned to these two sets as shown in this table, in which the validation data set is utilized to optimize the parameters as described in section 4.1. The feature used here is grass/sand distribution. Note that the reference data size is 100 shots and the total training sample number is 40. **Opt1** means the best average $2rp/(r+p)$ on the *validation data set* with the optimized parameters on the *validation data set* listed in the table. **Opt2** denotes the highest average $2rp/(r+p)$ on the *test data set* after optimizing α_1 and α_2 on the *test data set*.

Data Set		**Training** α_1, α_2			**Classification** Performance			
Validation	Test	α_1	α_2	**Opt1**	p	r	**Eva**	**Opt2**
v1	v2,v3,v4,v5	1.3	1.0	0.958	0.988	0.928	0.957	0.978
v1,v2	v3,v4,v5	1.3	1.0	0.971	0.993	0.924	0.957	0.974
v1,v2,v3	v4,v5	1.3	1.2	0.970	0.990	0.922	0.955	0.973
Average Performance				0.966	0.990	0.925	0.956	0.975

The performance of the online learning framework listed in **Table 2** demonstrates the robust generation capacity for different sizes of the validation and test data set. The average classification performance on the test data sets under the parameters optimized on the validation data sets (**Eva**) is about 0.956, which is only decreased by about 0.02 compared with the average performance on the test data sets under the parameters optimized on the corresponding test data sets (**Opt2**). As a result, α_1 and α_2 have the potential to be applied to a large range of videos after pre-training on a relatively small amount of samples.

Furthermore, as aforementioned, the number of the training samples can be reduced even further. According to our experiments, when it decreases from 40 to 8, the performance only descends from 0.974 to 0.972. Even only four training samples are available, average $2rp/(r+p)$ still remains at a relatively high value (0.777).

In addition, the size of the reference data currently used is 100 shots. If the application permits exploring more data, for example, the size is increased to 200 or 300 shots, the performance only improves 0.01 at most.

5 Conclusions

In this paper, we have proposed an online learning framework for sports video view classification by dynamically adapting the local classification models, in which the local statistic characteristics of the under-test videos are sufficiently explored. In this proposed framework, view classification is based on an adaptive local positive model and a set of negatives obtained by appropriately combining a small portion of unlabeled samples in the under-test video and a small amount of pre-labeled training samples, which make the results more robust and not sensitive to the parameters involved in the online learning and classification procedure. Experimental results demonstrate the effectiveness and efficiency of the proposed framework, even when only a very small quantity of training samples are available. Except for testing the proposed framework on more and wider range of videos, our future research effort will be focused on automatic feature evaluation and selection. Furthermore, an incremental learning framework, which dynamically increases the positive and negative model sets, will be explored to further improve the classification performance.

Acknowledgement

This work was supported in part by NSF Grant 60135010, 60321002.

References

[1] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, H. Sun, Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video, ICME, Tokyo, Aug. 22-25, 2001.

[2] Lexing Xie, Shih-Fu Chang, Ajay Divakaran, Huifang Sun, Structure Analysis of Soccer Video with Hidden Markov Models, ICASSP 2002,Volume: 4, pp13-17.

[3] Y.Gong, T.S. Lim, and H.C. Chua, Automatic Parsing of TV Soccer Programs, IEEE International Conference on Multimedia Computing and Systems, May, 1995, pp. 167 - 174.

[4] Di Zhong; Shih-Fu Chang; Structure Analysis of Sports Video Using Domain Models, ICME, Tokyo, Aug. 22-25, 2001, pp713 -716.

[5] G. Sudihir, J.C.M. Lee, A.K. Jain, Automatic classification of tennis video for high-level content-based retrieval, in Proc. IEEE International Workshop on Contented-based Access of Image and Video Database, Jan, 1998, Bombay, India.

[6] Baoxin Li, M.Ibrahim Sezan, Event Detection and Summarization in Sports Video, IEEE Workshop on Content based Access of Image and Video Libraries (CBAIVL), 2001

[7] Raaijmakers, S.; den Hartog, J.; Baan, J.; Multimodal topic segmentation and classification of news video, ICME, Aug 26-29, 2002, pp33-36, Vol2.