Cricket Video Events Recognition using HOG, LBP and Multi-class SVM

Qamber Abbas¹, Youmeng Li¹,*
¹College of Intelligence and Computing, Tianjin University, Tianjin, China
*Corresponding author e-mail: liyoumeng@tju.edu.cn

Abstract. The world has witnessed a growth in multimedia data, especially video data over the past few years due to increased internet bandwidth and higher processing power of computers. Having a large number of video data also require techniques to store, summarize, index and information retrieval. More attention has been given in recent years to develop techniques which summarize, index and retrieve sports videos due to its commercial aspects. This paper proposes a framework which classifies a cricket video into one of the four events namely Bowled Out, Caught Behind, Catch Out and LBW Out. The framework uses training videos from each category of event and summarizes the videos into key frames. HOG and LBP features are computed for key frames and fused to form a single feature vector which will be labeled accordingly which represents a single event in video. Feature vector is given to a Multi-Class SVM which classifies the video into one of the four events. The experimental results show that the Precision of our technique is 77.23%, Recall is 77.86%, F-Measure is 77.55% and the Accuracy is 65.62%. The evaluation metrics of our technique are promising, because in the literature there is no other technique present so far for event detection & classification in cricket videos.

1. Introduction

The world of multimedia has witnessed exponential growth in video data. Video data has grown in recent years mainly due to improvements in multimedia technology, improved processing power, faster and robust networks. Improvements in the technology has led to generation of a vast amount of video data, belonging to different areas like movies, sports, surveillance and news etc. [1].

Due to huge money-making capacity and large number of TV viewership in a game of cricket, the automatic highlights generation is considered to be of utmost importance [3]. Most of the viewers are only interested in watching a compact version of a completed game rather than a full match. The main contributions of proposed framework are as follows:

- An automatic event detection and classification from cricket video is presented in the proposed framework. The proposed framework is one of a kind, as there is no other technique present in the existing literature which detects and classifies important events from a cricket video. There exist other techniques which are based on cricket videos but they do not explicitly do detect and classify the important events from those videos.
- There does not exist any standard dataset for training and testing of events present in a cricket video. The dataset for this purpose has been manually developed, by separating those clips...
from a large cricket video which belonged to an event. Thus 160 clips each belonging to four events namely Bowled Out, LBW Out, Caught Behind and Catch Out were created manually.

The rest of the paper is organized as follows. Section 2 of this paper discusses the related work, Section 3 presents the proposed framework, in Section 4 the experimental results are discussed and in Section 5 the paper is concluded.

2. Related Work

In this section we discuss previous research in the field of sport video processing and analysis. In the past few years there has been an extensive research in the field of semantic analysis of sports videos. Automatic sport video annotations, sports video indexing, video retrieval and automatic highlights creation by making use of semantically important sports video content along with multimodal data was focused in the research [4]. Semantic analysis of sports video is considered as a challenging task, the presence of huge amount of sport video, different numbers of sports video broadcasters and existence of semantic gap between the high level and low level features [2]. The existing research in the field of sports video analysis can be broadly divided into two categories genre-specific and genre-independent. Most of the research in the field of semantic analysis of sports video is genre-specific, mainly because every sport is played differently and has different number of rules and actions. Genre-specific research focuses on specific sports video like soccer [2], Baseball [6], Volleyball, Tennis, Cricket [3,8], Golf and Basketball. For detection of an event in a particular sports video, it is not wise to claim that a genre-independent solution would provide feasible solution because no two sports are the same and every sport has its own set of rules and structure. American Football videos have been summarized by making use of textual overlays present in the videos by [9].

In this paper we propose a framework which detects and classifies significant events from a cricket video, which are Bowled Out, Catch Out, LBW Out and Caught Behind Out. Our framework is unique in a sense that there is no other technique which exists in the literature which detects and classifies events from a cricket video. There was no standard dataset of cricket videos, so dataset containing events from a large set of cricket videos was created manually by separating video clips containing an event. Each video in the dataset belongs to a particular event and these videos are used for training and testing purposes.

3. Proposed Method

In this section, all the processing steps involved in our proposed framework are explained in detail. Our video dataset $D$ is divided into training and testing dataset $D_{\text{training}}$ and $D_{\text{testing}}$, where the dataset $D$ consists of video clips from four different types of events i.e., Bowled Out, Catch Out, LBW Out and Caught Behind Out. The training dataset is represented as $D_{\text{training}} = \{ V^1_T, V^2_T, V^3_T, ..., V^{N/4}_T \}$, and testing dataset is represented as $D_{\text{testing}} = \{ V^{3(N/4)+1}_T, ..., V^N_T \}$ where N is the total number of videos in our video dataset $D$. During the training process videos in the training dataset are summarized into five key frames which are extracted for further processing by employing the summarization technique proposed in [10]. For instance, a video $V_T^i$ from our dataset is taken and a set of frames are extracted from it $F_T^i = \{ F_T^{i,1}, F_T^{i,2}, ..., F_T^{i,n} \}$, where $n = 5$ and $F_T^{i,1}$ represents one particular video from the database and $F_T^{i,1}$ represents the first frame of the first video. The experimental results show us that five key frames are sufficient for further processing and five key frames represents the entire video clip. Every extracted key frame’s size is adjusted to 125 by 250, it is converted into grayscale and each image is enhanced by applying Median Filter and Histogram Equalization to remove any noise and blur. In the next step HOG descriptor is used for every key frame and combined to form a HOG feature vector for a single video clip. Then LBP descriptor is used to extract features of all five key frames of the video and combined to form a feature vector. In the next step HOG and LBP feature vectors of all the five key frames of a single video clip are fused to form a single feature vector.
Figure 1. Proposed Framework for Cricket Event Classification.

Table 1. Description of the symbols used in the proposed framework.

| Symbol | Description |
|--------|-------------|
| D → Video Dataset | $F_i^{V,HOG}$ → $n^{th}$ frame’s HOG feature of $i^{th}$ video |
| $D_{Training}$ → Training Video Dataset | $F_i^{V,HOG}$ → HOG features extracted from all five frames of video $F_i^V$ |
| $D_{Testing}$ → Test Dataset | $F_i^{V,LBP}$ → LBP features extracted from all five frames of video $F_i^V$ |
| $V_i^T$ → $i^{th}$ video | $F_i^{V,LBP}$ → $n^{th}$ frame’s LBP feature of $i^{th}$ video |
| $F_i^V$ → Key frames of $i^{th}$ video | $F_i^V$ → Fused feature vector for frames of all videos $F_i^V (HOG \oplus LBP)$ |
| $F_i^{VL}$ → First frame of the $i^{th}$ video | $L_i$ → Event labels |
| $F_i^{VN}$ → $n^{th}$ frame of the $i^{th}$ video | SVM $\zeta$ → Multi-class SVM classifier |

In case of $F_i^V$ which is the set of extracted frames from any video, if HOG is applied on this set of images and combined together then it will be represented as a feature vector $F_i^V$ HOG, after applying LBP on the frames we will get its combined feature vector $F_i^V$ LBP. Once we get the features $F_i^V$ HOG and $F_i^V$ LBP, we fuse the two features together to get a single feature vector $F_i^V (HOG \oplus LBP)$. Feature vector is computed for all the extracted video frames and they are assigned a label according to the events they belong to as $L_i = \{L_1, \ldots, L_4\}$ where $i$ represents all four events. For training purposes these labeled features are given to a multi-class SVM $\zeta$. In testing phase a video is given as an input from the training dataset $D_{Training}$, key frames are extracted from the video and preprocessed, their features are extracted accordingly and represented as $F_i^V (HOG \oplus LBP)$ and are tested against the their labels $L_i$. The proposed framework is shown in the Figure 1. Details of all the symbols of the framework are shown in Table 1.

3.1. Frame Selection and Pre-processing

In the first step each video $V_i^T$ from $D_{Training}$ is given to the system as an input. If the number of extracted key frames from each video is different, then the size of fused feature vector of a particular video clip would also be different which will pose a problem during testing phase. Once the key frames of a video are extracted, certain preprocessing operations are applied on all the key frames. In order to remove any noise in the images, a media filter of size $3 \times 3$ is applied on all images of a single video. The filter replaces the neighborhood pixel’s value to the median value and helps in removing salt and pepper noise. Histogram equalization is applied on the images so that the image contrast is enhanced and provides better features for accurate classification.
3.2. Feature Extraction

In the proposed methodology, two feature Histogram of Oriented Gradient and Local Binary Pattern have been used for all extracted images belonging to each class of videos. HOG features of the selected frames are combined and similarly LBP features of frames are fused to form a consolidated one dimensional vector, which represents a complete event and labeled accordingly.

3.3. Event Classification

Our proposed framework uses Support Vector Machine (SVM) for classification of the cricket events from videos. SVM is a machine learning algorithm based on supervised learning, which means that labeled data is required for classification by SVM. SVM is a linear model which learns a line often called as a hyperplane, this hyperplane separates the data into two classes. Conventional SVM does binary classification i.e., it classifies the data into two classes only. In our proposed framework we need to classify the cricket events into four different classes, so we will make use of Multi-class SVM. SVM was designed for binary classification but it can be tweaked into a classifier, which does multi-classification by employing several techniques like one versus rest or one versus one, which converts a single multi-classification problem into several binary classification problems. In our proposed framework we have utilized the one versus all technique. Since the amount of features is on the higher side, so we have made use of SVM with a linear kernel $\delta$ which is described as follows (See Eq. 1):

$$\delta \left( \mathbf{f}_i^T (\text{HOG} \oplus \text{LBP}), \mathbf{\Psi}_i \right) = 1/(1 + e^{f_i^T (\text{HOG} \oplus \text{LBP})^T \mathbf{L} \mathbf{j}})$$

Given our labeled training sample which is represented as $(\mathbf{F}_i^T (\text{HOG} \oplus \text{LBP}), \mathbf{L}_j), \mathbf{L}_j \in \{1,2,3,4\}$ and the classification can be shown as (See Eq. 2):

$$\zeta \left( \mathbf{F}_i^T (\text{HOG} \oplus \text{LBP}) \right) = \text{sign} \left( \sum_{i=0}^{n} \alpha_i \mathbf{L}_j \delta \left( \mathbf{f}_i^T (\text{HOG} \oplus \text{LBP}), \mathbf{\Psi}_i \right) + b \right)$$

Where $\alpha_i$ the Lagrange multipliers of dual optimization problems, $\delta$ is the kernel function and $b$ is bias of the hyperplane.

4. Experimental Results

The main goal of this experiment is to evaluate the efficiency of the proposed system in detecting different cricket dismissals. Each of the four cricket dismissals as discussed in the previous section are tested on the test dataset or the cross validation dataset. Each dataset is comprised of video clips, where each video clip contains one event. Experimental results have demonstrated the performance and robustness of the proposed framework in providing flexibility to detect different events. The proposed system is implemented in Matlab (2017a), using Digital Image and Computer Vision Toolbox. All experiments were performed on Intel Core i7 CPU frequency of 1.8 GHz and 6GB of RAM.

4.1 Dataset

Since there was no standard dataset which could be used for event detection, so dataset was developed manually from cricket videos on the internet. In our experiments, the videos in all domains totaled 640, belonging to four events and there are a total of 160 video clips for each event. A total number of 480 videos were adopted for training four events. The summary of event learning using our dataset is provided in Table 2.

| Event Category in videos | Total videos | Event |
|--------------------------|--------------|-------|
| Catch Out                | 160          | 01    |
| LBW Out                  | 160          | 01    |
| Bowled Out               | 160          | 01    |
| Caught Behind Out        | 160          | 01    |
We evaluated our proposed system using three common metrics of measurements that are accuracy, detection rate and positive predictive value (PPV). Where TP (True Positive) indicates that the system correctly detects an event; TN (True Negative) indicates that the system performs correct rejection, FP (False Positive) indicates the false alarm from the system and FN (False Negative) indicates the miss detection from the system. We manually annotated the ground truths where events and non-events are identified for each of the dataset. The average accuracy for the four events detectors is 65.625%; average of detection rate is 77.8% and average PPV is 77.23%.

4.2. Classified Visual Results

The video is given as input to frame extraction algorithm as explained in the proposed methodology section. Our algorithm summarizes the video clip into five key frames as shown in Fig. 2. Then these key frames are given as input to the classifier. These frames are classified into five frames for four events namely ‘Bowled Out’, ‘Caught Behind’, ‘Caught Out’ and ‘LBW Out’. Bowled category of the video clips contains all Bowled Out event videos and same as LBW, Catch and Caught Behind. First of all, the video is given as an input to the key frame extraction algorithm as explained in the proposed methodology section. Then these five frames are given as input to the proposed classifier which accurately classifies the video as Bowled, LBW, Catch and Caught Behind category.

The F-measure of the proposed system is given in Table. 3. In this table, the first row represents the values for Bowled event category. The next three rows contain the value of LBW, Catch and Caught Behind event category. Where Caught Behind event category has the highest TP value. Then, a total is calculated by adding the values of all the four events. In the end Precision, Recall, F1-score and Accuracy are calculated. Table. 3 is the subjective analysis results of the suggested technique to check over different descriptors and find the precision and recall of each descriptor, the performance result of our proposed method have high precision and recall value. The below graph (See Fig. 3) shows the comparison of different video events with ground truth that has been proved in Table. 3. The black line shows the values of the Bowled event category and the red, blue and green lines show the values of LBW, Catch and Caught Behind event category.

**Figure 2.** Visual Results of Our Proposed Technique.
Table 3. Overall F-measure of the Proposed System.

| Category          | TP  | FP  | TN  | FN  |
|-------------------|-----|-----|-----|-----|
| Bowled            | 24  | 5   | 0   | 11  |
| LBW               | 24  | 8   | 5   | 3   |
| Catch             | 22  | 6   | 4   | 8   |
| Caught behind     | 25  | 9   | 1   | 5   |
| Total             | 95  | 28  | 10  | 27  |

Precision: 0.77235  
Recall: 0.77868  
F1-score: 0.77551  
Accuracy: 0.65625

Figure 3. Precision and Recall of Different Events.

5. Conclusion

In this paper a novel framework for detecting events in cricket video is presented. An input video clip is summarized into key frames. HOG and LBP features are computed for each key frame and fused to form a single feature vector and the video clip is labeled accordingly. For training and testing purposes Multi-Class SVM is used. There is no method present in existing literature which detects important events in a cricket video. The evaluation metrics used to evaluate our technique with the ground truth was Precision, Recall, F-measure and Accuracy which shows good results. Our proposed technique uses HOG and LBP feature descriptor. The size of our feature vector is too large due to the descriptors used. In the future work, we would be exploring such descriptors which are not just robust but more efficient than the current descriptors used.

References

[1] Ballan, L., et al., Event detection and recognition for semantic annotation of video Multimedia Tools and Application, 2011. 51(1):p. 279-302.
[2] Hosseini, M.-S. and A.-M. Eftekhari-Moghadam, Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video. Applied Soft Computing, 2013. 13(2): p. 846-866.
[3] Kolekar, M.H. and Sengupta, S.: Hidden Markov Model Based Structuring of Cricket Video Sequences Using Motion and Color Features, ICVGIP, 632-637,(2004).
[4] Xu, H. and Chua, T.: The fusion of audio-visual features and external knowledge for event detection in team sports video, Proceedings of 6h ACM SIGMM international workshop on Multimedia information retrieval, 127-134, ACM,(2004).
[5] Li, Y., Narayanan, S. and Kuo, C.: Content-based movie analysis and indexing based on audiovisual cues, IEEE transactions on circuits and systems for video technology, 14, 1073–1085, IEEE, (2004).

[6] Cheng, C.C. and Hsu, C.T.: Fusion of audio and motion information on HMM-based highlight extraction for baseball games, IEEE Transactions on Multimedia, 8, 585–599, IEEE, (2006).

[7] Duan, L.Y., Xu, M., Tian, Q., Xu, C.S., and Jin, J.S.: A unified framework for semantic shot classification in sports video, IEEE Transactions on multimedia, 7, 1066–1083, IEEE, (2005).

[8] Kolekar, M.H. and Sengupta, S.: Semantic concept mining in cricket videos for automated highlight generation, Multimedia Tools and Applications, 47, 545–579, Springer, (2010).

[9] Babaguchi, N., Kawai, Y., Ogura, T. and Kitahashi, T.: Personalized abstraction of broadcasted American football video by highlight selection, IEEE Transactions on Multimedia, 6, 575–586, IEEE, (2004).

[10] Ejaz, N., Tariq, T.B., Baik, S. W.: Adaptive key frame extraction for video summarization using an aggregation mechanism, Journal of Visual Communication and Image Representation, 23, 1031–1040, Elsevier, (2012).