Semantic Event Extraction Using Neural Network Ensembles

Min Chen
Department of Computer Science
University of Montana
Missoula, MT, USA
mchen005@cs.fiu.edu

Chengcui Zhang
Department of Computer & Information Science
University of Alabama at Birmingham
Birmingham, AL, USA
zhang@cis.uab.edu

Shu-Ching Chen
School of Computing and Information Sciences
Florida International University
Miami, FL, USA
chens@cs.fiu.edu

Abstract

This paper proposes a novel semantic content analysis framework for reliable video event extraction which is essential for high-level video indexing and retrieval. In this work, we target to address the unique challenges posed in rare event detection, where positive examples (i.e., eventful data points) are vastly outnumbered and thus overshadowed by negative ones (i.e., noneventful data points). The proposed framework tackles this issue by integrating the strength of multimodal content analysis and neural network ensembles. Specifically, due to the rareness of the target events, the boostrapped sampling method is adopted to reduce the effect of class imbalance and a group of component neural networks are constructed consequently. Thereafter, a weighting scheme is applied to intelligently traverse and combine the component network predictions. The effectiveness of the proposed framework is demonstrated over a large collection of soccer video data with different styles produced by different broadcasters.

1. Introduction

The amount of accessible video information has been increasing rapidly. Consequently, efficient video indexing and retrieval have emerged as challenging yet appealing research topics in multimedia applications. As humans tend to use high-level semantic concepts when querying and browsing multimedia databases, recent studies on video indexing and retrieval have been greatly shifted from low-level feature-based approaches to high-level semantic analysis.

Events are the most important part in video semantic contents, which are defined as representing the temporal interactions and variations that compose the story line in a video program [14]. In terms of event detection, a variety of learning algorithms or stochastic methods, such as Dynamic Bayesian Network [2], Neural Network [8], Support Vector Machines (SVMs) [9], and Hidden Markov Model (HMM) [7], have been adopted to derive the interested events from the video features. However, the occurrence of interested events in many applications is fairly scarce, e.g., traffic accident events in transportation videos and security threat events in surveillance videos, etc. In this case, event detection becomes even more difficult with imbalanced training data where positive training examples (i.e., eventful data points) are vastly outnumbered by negative ones (i.e., noneventful data points). Such problem is called rare event detection in this paper. Most of the existing learning approaches have difficulties in addressing this issue. For instance, the standard SVM treats positive and negative examples equally such that the margin of the SVM hyperplane to negative training examples is equal to the one to positive ones. In the case of rare event detection, as the positive examples are so rare that they are not representatives of the genuine distribution of positive examples, the performance of SVM drops significantly [13]. The similar issue exists in Neural Network and Bayesian Network where positive examples are overshadowed by the negative ones.

In addition, it is of great importance to develop an extensible framework capable of detecting various events in different video types. However, for a given indexing or event detection task, it is unfeasible to assume that a unique solution for all genres of videos exists [9]. Hence, in many so-called generalized approaches, various mid-level or high-level features were explored, such as ball location [11] and field line orientation [9], where their extraction either is computationally expensive or remains as an open issue. Consequently, it is critical to reach a reasonable tradeoff between efficiency and extensibility.

Moreover, as discussed in [1], the performance of a learning algorithm is measured by its generalization, which means the ability to correctly classify data points not in the training set. The framework presented in [2][7] did not justify its generalization performance in the sense that the generalization error is not discussed theoretically nor tested empirically.

To address the above-mentioned issues, in this paper, a novel framework using multimodal content analysis and neural network ensembles is proposed by taking into consideration of the unique challenges posed by rare event detection. For the sake of efficiency, where possible, the feature detectors are implemented from low-level features taken directly or derived easily from the audio/visual bit-stream. Neural network ensembles have been applied for soccer goal event detection in our earlier work [12]. In this study, the framework is further extended to identify the corner events including corner-kicks, free-kicks near the penalty box, and line throws from the corner (see examples in Figure 1). These events are considered as interesting events as they give an opportunity for one team to dominate over the other and possibly lead to a goal event. These events normally keep the sports fans at the edge of their seats hoping for something exciting. Since the characteristics of goal events vary greatly from those of corner events, the focus of this paper is to testify the extensibility of our framework using a large set of soccer videos and strict cross-validation to estimate the generalization error.

(a) Corner-kick (b) Corner-throw (c) Free-kick

Figure 1. Example corner events

The rest of the paper is organized as follows. Section 2 details the proposed framework. In Section 3, performance evaluation is presented. Finally, Section 4 concludes this paper.

2. Overall framework

Figure 2 shows our proposed framework. It consists of three phases. The first two phases (syntactic segmentation and descriptor extraction) involve the modeling of video into a computer understandable format, where the modeling strategy and the parameters are selected to facilitate the final decision making process (Phase 3). In the following subsections, each phase will be discussed with the main focus on phases 2 and 3 in terms of corner event detection.

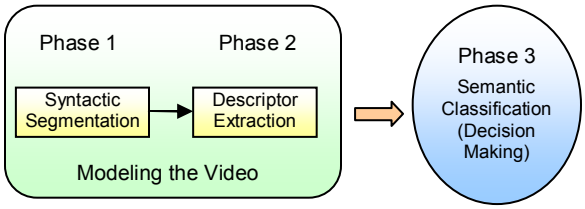


Figure 2. Proposed semantic highlight extraction framework

2.1. Syntactic segmentation

Syntactic segmentation involves temporal partitioning of the video sequence into meaningful units. These segments then serve as the basic unit for descriptor extraction and semantic annotation.

Various syntactic units have been proposed in the literature such as frame level, shot level, scene level, and clip level. Among them, shots are widely accepted as a self-contained and well-defined unit. Therefore, the basic syntactic unit in this work is defined in the shot level. For this purpose, our shot-boundary detection algorithm [4] is employed, which consists of pixel-histogram comparison, segmentation map comparison, and object tracking. Here, the segmentation map and object information are extracted by using the unsupervised object segmentation method SPCPE proposed in our previous work [5]. In essence, the basic idea of the shot-boundary detection algorithm is that the simpler but more sensitive checking steps (i.e., pixel-histogram comparison) are first carried out to obtain a candidate pool, which thereafter is refined by the methods that are more effective but with a relatively higher computational cost.

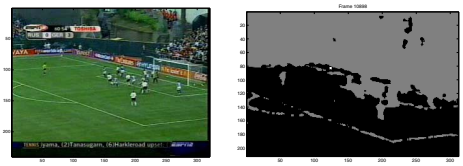
2.2. Descriptor extraction

In our earlier work [3], 5 visual features (pixel-change, histo-change, grass-ratio, background_mean, background_var) and 15 audio features (4 volume features, 7 energy features, and 4 spectrum flux features) were extracted for goal event detection. In brief, visual features are captured with the assistance of color analysis and object segmentation techniques, whereas audio features are exploited in both time-domain and frequency-domain. Most of these features are low-level visual/audio features and can be readily extracted from the video stream with minor computational requirements.

For corner event detection, the multimodal feature set proposed in [3] is adopted together with the following three types of derived features.

- Upper foreground object ratio

Normally, with the occurrence of a corner event, the focus is on a certain corner of the play field with large off-field areas (e.g., audience area, billboard, etc.) shown on the scene, as can be seen from Figure 1. To capture this idea, the key frame, currently the 8th frame of each shot is segmented and analyzed. As illustrated in Figure 3(b), the pixels in each frame are grouped into two different classes, where the foreground area is marked with the gray color and background is displayed in black. Then the percentage of foreground objects in the upper one-third region of the scene is calculated, which is called "upper foreground object ratio" in this study.



a) Corner scene (b) Segmentation result

Figure 3. Example corner scene and its segmentation result

- Shot view label

In the literature, various approaches have been proposed for shot view classification. Most of the existing works utilize grass ratio as an indicator of the shot view type, assuming that a global view (e.g., Figure 4(a)) has a much greater grass ratio value than that of a close view (e.g., Figure 4(b)) [3][11]. However, close view shots such as the one shown in Figure 4(c) could have large grass ratio values. Thus, the use of grass ratio alone can lead to misclassifications. In this study, we propose a simple yet efficient 'object detector' coupled with grass ratio for shot view classification.

The basic idea of 'object detector' is to detect the foreground objects and their associated sizes in the play field, where two object detector windows, object search window (the small blue rectangle in Figure 4) and object locate window (the large green rectangle in Figure 4), are applied. Let *fw* and *fh* be the frame width and height, respectively. In our framework, the size of each window and the center of the object locate window are derived empirically as shown in Table 1. The detailed algorithm for shot view label assignment is expressed in Table 2. The thresholds *THu*, *THs* and *THg* are set to 0.7, 0.7 and 0.3, respectively based on experimental observations.

(a) Global view (b) Close view (c) Close view

Figure 4. Example shot views

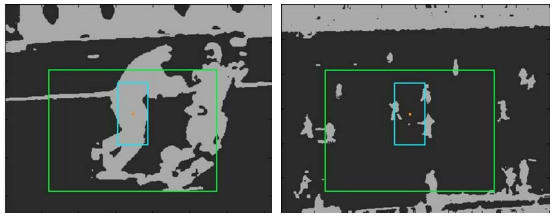


Figure 5. Example object detector windows

Table 1. Object locate window & object search window setup

	Object Locate Window	Object Search Window
Width	0.8**fw*	0.1**fw*
Height	0.6**fh*	0.25**fh*
Center	(0.5**fw*, 0.6**fh*)	(0.15**fw*, 0.425**fh*)

As can be seen from Table 2, the object locate window is a static window, whereas the object search window is a sliding window which slides through the object locate window searching for objects (steps 2-10). If the object size within a certain object search window reaches a predefined threshold, the scene is labeled as a close view (steps 6-7). Otherwise, it is identified as global view if it has a large grass area (steps 11-12). A medium view label is assigned to the scene if none of the criteria is satisfied (step 14).

- Temporal features

Inherently, a video is composed of temporal evolution of series of activities. A single analysis unit (e.g., a shot) which is separated from its context has less capability of conveying semantics [15]. Temporal information in a video sequence plays an important role in conveying video content. In general, an important event is a result of past activities and it might cause effects in the future as well. In terms of corner event detection, without loss of generality, it is assumed that it takes up to four past video shots to evolve into the targeted event. Therefore, shot view labels and grass ratios of the previous four shots are included as the temporal features in the final feature set.

Table 2. The proposed algorithm for view label assignment

1. If upper foreground object ratio > *THu*, the foreground object in the upper one-third region is considered as the off-field area and is not counted as the targeted foreground pixels in the play field.
2. Place the object locate window at the initial position (a, b). Currently, a and b are set to be 0.15**fw* and 0.425**fh*, respectively, as defined in Table 1.
3. **For** r = 0 to 0.35**fh*
4. **For** c = 0 to 0.7**fw*
5. Place the center of the object search window at (a+c, b+r).
6. **If** the percentage of foreground pixels inside the object search window reaches *THs*
7. Assign a close view label to the shot. Stop.
8. **endif**
9. **endfor**
10. **endfor**
11. **If** the grass ratio of the shot is greater than *THg*
12. Assign a global view label to the shot. Stop.
13. **Else**
14. Assign medium view label to the shot. Stop.
15. **endif**

2.3. Semantic Classification

The semantic event detection process can be viewed as a function approximation problem, where the task is to learn a target function f that maps a set of feature descriptors x to one of the pre-defined event labels y. The target function is called a *classification model*. Hence, once the input feature set is determined, the next step is to select the proper methodology for function approximation. As discussed earlier, for rare event detection, special attention should be paid to alleviate the problems associated with the imbalanced nature of the training data set. In this study, the bootstrapped sampling scheme proposed in [12] is adopted for this purpose. In brief, for a given training data set containing event subset E and nonevent subset S, where $|E| << |S|$, S is divided into N groups (called bootstrapped samples) and E is bootstrapped into each of the N groups.

$$N = \lfloor |S|/(r^*|E|) \rfloor \quad (1)$$

Here, r is a constant that determines the non-event to event instance ratio in the bootstrapped samples. The idea is to form a number of samples containing a comparably equal number of instances from both event and nonevent classes. In this study, r is set to 1. N component neural networks are thereafter trained using each of the samples, whose network structure consists of two layers, with 20 and 1 logistic sigmoid neurons in the hidden layer and output layer, respectively. Generally, training the neural network to avoid over-fitting the training set is a topic under discussion in many studies. However, it was shown in [10] that it is actually advantageous to use under regularized component networks which overfit the training data for learning in large ensembles. Therefore, the component networks in our work are trained using a back-propagation algorithm to achieve maximum recall and precision for the training set. Finally, the weighting scheme presented in [12] is applied to intelligently traverse and combine the decisions of the N individual neural networks to reach the final decision. The idea is that each trained component network is tested upon all the N training subsets, where the one with a higher generalization ability gets a higher weight.

In summary, our proposed framework adopts the idea of neural network ensembles since they possess strong capability in reducing the generalization error [6]. Furthermore, in real-world applications, it is more desirable to incorporate a preprocessing phase, where clearly irrelevant data points in the training and testing data sets are removed prior to subsequent modeling, analysis and classification.

3. Experiments

In our experiments, 25 soccer videos were collected from different broadcasters, with total time duration of 8 hours and 25 minutes. These videos are composed of 4,206 shots, out of which 93 are corner events which account for only 2.2% of the whole data set. As discussed earlier, it is quite challenging for such rare event detection. To demonstrate the effectiveness of our proposed framework, a comparative experiment was conducted.

The data set was randomly divided into a training set and a testing set, where the training set constitutes about 2/3 of the data set and the rest go to the testing set. Five such groups were formed randomly to employ 5-fold cross validation. Without addressing the class imbalance issue, the Radial Basis Function (RBF) neural network was adopted for semantic classification and the average recall and precision scores across five groups were 69.23% and 17.48%, respectively.

As a comparison, in our framework, the non-event shots in the training set were randomly subsampled into N sets and the corner events were bootstrapped into each set. In this experiment, we set r to 1. The

experimental results are summarized in Table 3. Here columns 'Iden', 'Mis', 'Misiden', 'R (%)', and 'P (%)' denote 'Identified', 'Missed', 'Misidentified', 'Recall (%)' and 'Precision (%)', respectively. In addition, 'Other' shows the number of 'goal kicks' and 'line throws' events misidentified as the corner events, which were counted as false positives and included in 'Misiden'. From this table, we have the following two observations. First, the proposed framework can achieve high recall values with good precision values using strict cross-validation, which demonstrates its encouraging generalization performance. Second, a large portion of the false positives belong to either goal kicks or line-throws whose broadcast patterns are quite similar to that of corner events.

Table 3. Experimental results for corner event detection

Dataset	Tot	Iden	Mis	Misiden	Other	R (%)	P (%)
1	27	24	3	9	4	88.89	72.73
2	34	31	3	12	7	91.18	72.09
3	39	36	3	14	4	92.31	72.00
4	25	21	4	8	4	84.00	72.41
5	33	31	2	13	6	93.94	70.45
Avg.	**32**	**29**	**3**	**11**	**5**	**90.06**	**71.94**

4. Conclusions

In this paper, an advanced framework for semantic highlight extraction is proposed. It consists of three main components, syntactic segmentation, descriptor extraction and semantic classification. Specifically, video clips are first segmented into a set of basic analyzing units (i.e., shots) and shot-level audio/visual features are then extracted. To address the challenges associated with rare event detection, neural network ensembles with bootstrapped sampling scheme are adopted for data classification. The effectiveness of the proposed framework is evaluated on soccer corner event detection from a considerably large number of soccer videos with various production styles. In addition, the generalization performance is fully testified via the strict 5-fold cross validation scheme.

5. ACKNOWLEDGEMENT

For Shu-Ching Chen, this work was supported in part by NSF EIA-0220562, NSF HRD-0317692, and Florida Hurricane Alliance Research Program sponsored by the National Oceanic and Atmospheric Administration.

6. REFERENCES

[1] C. Burgess, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, 1998, pp. 121-167.

[2] C.-Y. Chao, H.-C. Shih, and C.-L. Huang, "Semantic-Based Highlight Extraction of Soccer Program Using DBN," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2005, pp. 1057-1060.

[3] S.-C. Chen, M.-L. Shyu, C. Zhang, L. Luo, and M. Chen, "Detection of Soccer Goal Shots Using Multimedia Features and Classification Rules," in *Proceedings of the 4th International Workshop on Multimedia Data Mining*, 2003, pp. 36-44.

[4] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative Shot Boundary Detection for Video Indexing," in *Video Data Management and Information Retrieval*, Edited by S. Deb, Hershey, PA, USA: Idea Group Publishing, 2005, pp. 217-236.

[5] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Video Scene Change Detection Method Using Unsupervised Segmentation and Object Tracking," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME),* 2001, pp. 57-60.

[6] L. K. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, 1990, pp. 993-1001.

[7] Y.-L. Kang, J.-H. Lin, M. S., Kankanhalli, C.-S. Xu, and Q. Tian, "Goal Detection in Soccer Video Using Audio/Visual Keywords," in *Proceedings of 2004 International Conference on Image Processing (ICIP)*, 2004, pp. 1629-1632.

[8] W.-N. Lie, T.-C. Lin, and S.-H. Hsia, "Motion-based Event Detection and Semantic Classification for Baseball Sport Videos," in *Proceedings of IEEE International Conference on Multimedia and Expo*, vol. 3, 2004, pp. 1567-1570.

[9] D. A. Sadlier and N. E. O'Connor, "Event Detection in Field Sports Video Using Audio-Visual Features and a Support Vector Machine," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, 2005, pp. 1225-1233.

[10] P. Sollich and A. Krogh, "Learning with Ensembles: How Over-fitting Can Be Useful," Edited by D.S. Touretzky, M.C., Mozer, and M.E. Hasselmo, Advances in Neural Information Processing Systems 8, MIT Press, 1996, pp. 190-196.

[11] V. Tovinkere and R. J. Qian, "Detecting Semantic Events in Soccer Games: Towards A Complete Solution," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2001, pp. 1040-1043.

[12] K. Wickramaratna, M. Chen, S.-C. Chen, and M.-L. Shyu, "Neural Network Based Framework for Goal

Event Detection in Soccer Videos," in *Proceedings of IEEE International Symposium on Multimedia*, 2005, pp. 21-28.

[13] G. Wu and E. Chang, "Class-Boundary Alignment for Imbalanced Dataset Learning," in *Proceedings of the 12th International Conference on Machine Learning (ICML) Workshop on Learning from Imbalanced Datasets*, 2003, pp. 49-56.

[14] G. Xu, Y.-F. Ma, H.-J. Zhang, and S.-Q. Yang, "An HMM-Based Framework for Video Semantic Analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 11, 2005, pp. 1422-1433.

[15] X. Zhu, X. Wu, A. K. Elmagarmid, Z. Feng, and L. Wu, "Video Data Mining: Semantic Indexing and Event Detection from the Association Perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 5, 2005, pp. 665-677.