VIDEO EVENT DETECTION WITH COMBINED DISTANCE-BASED AND RULE-BASED DATA MINING TECHNIQUES

Zongxing Xie^1, *Mei-Ling Shyu*^{1*}, *Shu-Ching Chen*^{2†}

^{1}Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33124, USA
^{2}School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
z.xie1@umiami.edu, shyu@miami.edu, chens@cs.fiu.edu

ABSTRACT

In this paper, the *rare event detection* issue in video event detection is addressed through the proposed data mining framework which can be generalized to be domain independent. The fully automatic process via the combination of distance-based and rule-based data mining techniques can greatly reduce the number of negative (non-event) instances and the feature dimension to facilitate the final event detection, without pruning away any positive (event) testing instance along the process. The effectiveness and efficiency of the proposed framework are demonstrated over the goal event detection application based on a large collection of soccer videos with different styles.

1. INTRODUCTION

Data mining techniques have been increasingly developed to provide solutions for semantic event detection in diverse types of videos [1]. Video events are normally defined as the interesting events which capture user attentions [2]. For example, a soccer goal event is defined as the ball passing over the goal line without touching the goal posts and the crossbar. Most current research for video event detection heavily depends on certain artifacts such as domain knowledge and priori model [3], and thus making them hard to be extendible to other domains or even other data sets. Though some work have been conducted to deal with the general video event extraction, they can only achieve rough detection capability [4] due to the well-known *semantic gap* and *rare event detection* issues [5]. In particular, the *rare event detection* issue, also known as *imbalance data set* problem, occurs in most video event detection applications. This issue is referred to as a very small percentage of positive instances versus negative instances, where the negative instances dominate the detection model training process, resulting in the degradation of the detection performance.

In order to develop a generalized event detection framework applicable to different application domains, a necessary step is to relax the need of domain knowledge such as those domain-specific rules with pre-defined fixed thresholds [6] and additional domain-based high-level features [5], which are often used to increase the percentage of the positive instances in the data set. Toward such a demand, our proposed framework attempts to achieve a fully automatic video event detection procedure via the combination of distance-based and rule-based data mining techniques, which are two basic and widely used measurements in data mining. We validate our proposed data mining framework using soccer goal event detection as the testbed, and the experimental results on 27 soccer videos collected from different broadcasters demonstrate the powerfulness and potential of integrating distance-based and rule-based data mining techniques.

The remainder of this paper is organized as follows. In Section 2, the architecture of the proposed data mining framework is presented. Empirical study and performance evaluation are discussed in Section 3. Finally, conclusions are given in Section 4.

2. THE PROPOSED FRAMEWORK

Figure 1 shows the architecture of our proposed data mining framework that consists of *Video Parsing and Feature Extraction*, *Distance-based Data Mining*, and *Rule-based Data Mining* phases.

2.1. Video Parsing and Feature Extraction

In our soccer goal event detection testbed, raw soccer video is parsed via a video shot detection subcomponent and multi-level features are extracted based on multimodal theory shown in details in our previous work [5]. Here, the feature set F

[*] This research was supported in part by NSF ITR (Medium) IIS-0325260.

[†] 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.

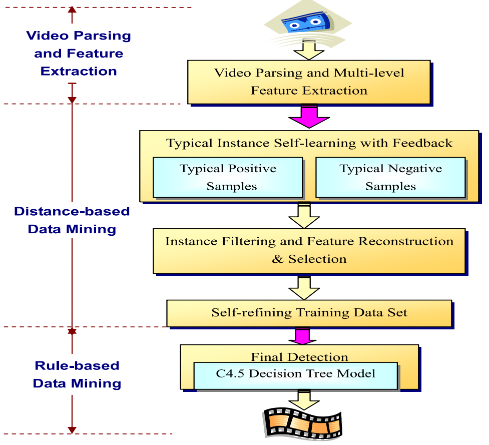
Video Parsing and Feature Extraction
Video Parsing and Multi-level Feature Extraction
Typical Instance Self-learning with Feedback
Typical Positive Samples
Typical Negative Samples
Distance-based Data Mining
Instance Filtering and Feature Reconstruction & Selection
Self-refining Training Data Set
Rule-based Data Mining
Final Detection
C4.5 Decision Tree Model

Fig. 1. Architecture of the Framework

contains totally 17 features including 10 audio features, 5 visual features, and 2 temporal features (i.e., volume_sum and nextfirst_mean) [5].

2.2. Distance-based Data Mining

The proposed schemes in this phase are the keys to achieve fully automatic detection in addressing the *rare event detection* issue. To our best knowledge, in other existing event detection frameworks, certain artifacts, especially domain knowledge, are required to alleviate the issue originating from an imbalanced data set to obtain promising recall and precision performance.

Our proposed distance-based data mining scheme is motivated by the powerfulness and robustness of our novel distance-based anomaly detection algorithm called Representative Subspace Projection Modeling (RSPM) [7] under different application domains and diverse types of data sets. Accordingly, we have developed (1) a feedback based self-learning positive instance selector to select typical positive instances and typical negative instances automatically for further instance and feature filtering; (2) two classifiers sequentially trained with the selected typical positive instances and typical negative instances to greatly decrease the number of data instances and the number of features for further rule-based detection model training; and (3) a linear analysis method to refine the training data instances based on the cluster of the score value corresponding to the first principal component.

2.2.1. *Typical Instance Self-learning with Feedback*

Let $\mathbf{X}$ = $\{\mathbf{x}_{ij}\}$ be a matrix of size $p \times N$, containing N p-dimensional column vectors, representing N original data instances after video parsing and features extraction. Among them, assume that there are $N1$ labeled positive instances $\mathbf{X}^1$ = $\{\mathbf{x}_{ij}\}$ and $N2$ labeled negative instances $\mathbf{X}^2$ = $\{\mathbf{x}_{ij}\}$. The proposed **typical positive instances** are defined as selected $T1$ ($T1$ < $N1$) positive instances in $\mathbf{X}^e$ = $\{\mathbf{x}_{ij}\}$. The normalized matrix $\mathbf{Z}^e$=$\{\mathbf{z}_{ij}\}$ of $\mathbf{X}^e$ can be obtained via Equation (1), where $\bar{\mu}_i$ and s_{ii} are the sample mean and variance of the i^{th} row in $\mathbf{X}^e$.

$$\mathbf{z}_{ij} = \frac{\mathbf{x}_{ij} - \bar{\mu}_i}{\sqrt{s_{ii}}}. \tag{1}$$

Let (λ_1^e, $\boldsymbol{E}_1^e$), (λ_2^e, $\boldsymbol{E}_2^e$), $\ldots$, (λ_p^e, $\boldsymbol{E}_p^e$) be the p typical positive eigenvalue-eigenvector pairs of the robust correlation matrix [7] for $\mathbf{Z}^e$, where $\lambda_1^e \geq \lambda_2^e \geq \ldots \geq \lambda_p^e \geq 0$. Also, let the typical positive instance score matrix $\mathbf{Y}^e$= $\left\{\mathbf{y}_{ij}\right\}$ be the projection of $\mathbf{Z}^e$ onto the p-dimensional typical positive eigenspace. Those principal components whose corresponding score row vectors $\boldsymbol{R}_i^e$= $(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{iT1})$ of $\mathbf{Y}^e$ satisfying Equation (2) are selected as the representative components to model the similarity of the typical positive instances.

$$STD(\boldsymbol{R}_m^e) < \text{a}, \tag{2}$$

where $STD(\boldsymbol{R}_m^e)$ is the standard deviation of the score row vectors satisfying Equation (2) and a is the arithmetic mean of the standard deviation values from all $\boldsymbol{R}_i^e$. Next, a class-deviation equation (Equation (3)) is designed to differentiate the normal and anomaly instances in the view of typical positive instances, where M is the selected index vector from Equation (2).

$$\mathbf{c}_j^e = \sum_{m \in M} \frac{(\mathbf{y}_{mj})^2}{\lambda_m^e} \tag{3}$$

The maximum value $\mathbf{c}_{max}^e$ of all $\mathbf{c}_j^e$ for $\mathbf{X}^e$ is selected as a threshold to justify if an incoming instance is statistically normal to the typical positive instances. In other words, if the score matrix $\mathbf{Y}$= $\left\{\mathbf{y}_{ij}\right\}$ is the projection of matrix $\mathbf{Z}$ (the normalized matrix of $\mathbf{X}$ via Equation (1)) onto the typical positive eigenspace and any $\mathbf{c}_j$ for $\mathbf{X}$ is less than $\mathbf{c}^e{}_{max}$, the corresponding instance will be considered as normal and otherwise anomaly.

Accordingly, in order to identify the best group $\mathbf{X}^e$, we randomly select $T1$ instances from $\mathbf{X}^1$ for K times. Each time, the percentage of the recognized instances in $\mathbf{X}^1$ and the percentage of the rejected instances in $\mathbf{X}^2$ are recorded, and $\mathbf{X}^e$ is defined as the one that can recognize 100% instances in $\mathbf{X}^1$ and at the same time reject the maximal percentage of instances in $\mathbf{X}^2$, when K and $T1$ are big enough.

Similarly, the proposed **typical negative instances** $\mathbf{X}^n$ is defined as the selected $T2$ ($T2$ < $N2$) negative instances which can reject 100% instances in $\mathbf{X}^1$ and at the same time recognize the maximal percentage of instances in $\mathbf{X}^2$. Then, the typical negative eigenvalue-eigenvector pairs (λ_1^n, $\boldsymbol{E}_1^n$), (λ_2^n, $\boldsymbol{E}_2^n$), $\ldots$, (λ_p^n, $\boldsymbol{E}_p^n$) and their typical negative eigenspace can be obtained. Finally, the index vector M' for typical negative instances can be automatically determined.

2.2.2. Instance Filtering and Feature Reconstruction & Selection

Before the execution of this step, the original data set $\mathbf{X}$ is randomly split into two disjoint subsets, namely a training data set $\mathbf{X}^A$ (with known class labels) and a testing data set $\mathbf{X}^B$ (with unknown class labels). We can further assume that the labeled positive instances in $\mathbf{X}^A$ are $\mathbf{X}^{Ae}$ and the labeled negative instances are $\mathbf{X}^{An}$. From this step, data mining event detection is gradually achieved via distance-based filtering followed by rule-based classification. In this step, $\mathbf{X}^e$ and $\mathbf{X}^n$ are trained sequentially to conduct a rough classification to increase the percentage of positive instances, which addresses the *rare event detection* issue.

$\mathbf{X}^e$ is first trained to reject negative instances in $\mathbf{X}^{An}$ and $\mathbf{X}^B$, i.e., to recognize the possible candidate positive instances and to exclude all other instances. The recognized normal data instances are passed to the second classifier trained with the selected $\mathbf{X}^n$. In the second classifier, the recognized normal instances are removed again from the refined $\mathbf{X}^{An}$ and $\mathbf{X}^B$ since these instances can be considered as "counterfeit positive instances" as they are double recognized by both typical positive and negative instances. In addition, all the remaining $\mathbf{X}^{An}$, $\mathbf{X}^{Ae}$, and $\mathbf{X}^B$ after two classifiers are projected onto the typical negative eigenspace and the score values corresponding to the index vector M' are extracted to replace the original feature set. In other words, the original feature set F are reconstructed and filtered to be F' with the dimension p' $(p' < p)$. The remaining and reconstructed data matrix can be defined as $\mathbf{Y}^{Ae}$=$\{\mathtt{Y}_{ij}\}$ for positive instances in the training data set, $\mathbf{Y}^{An}$=$\{\mathtt{Y}_{ij}\}$ for negative instances in the training data set, and $\mathbf{Y}^B$=$\{\mathtt{Y}_{ij}\}$ for all testing data where $(i = 1, 2, \ldots, p')$.

It is worth to mention that all positive instances would be kept in the filtered data set after the execution of the two classifiers. In the soccer event detection testbed, more than 90% of the negative instances are filtered without removing any positive instances. The proposed scheme effectively addresses the *rare event detection* issue and thus provides a good platform for the next rule-based data mining algorithm.

2.2.3. Self-refining Training Data Set

The self-refining of the training data set is achieved by linear analysis on the score row vectors $\boldsymbol{R}_1^{Ae}$ and $\boldsymbol{R}_1^{An}$ which correspond to the first selected principal component for $\mathbf{Y}^{Ae}$ and $\mathbf{Y}^{An}$, respectively. It is noted that the first selected principal component is always the first one in M' [7] which presents the most data information, and thus is used to refine the training data set with the attempt to improve training model accuracy in the later rule-based algorithm. In this single dimension, most $\boldsymbol{R}_1^{Ae}$ and $\boldsymbol{R}_1^{An}$ values are separated except several interlaced ones. As we use the typical negative eigenspace for obtaining the projected scores, most $\boldsymbol{R}_1^{Ae}$ (the relative anomaly ones) would have a higher value than that of the $\boldsymbol{R}_1^{An}$ ones. Though there are only several interlaced instances, they may greatly degrade the performance of the rule-based classifier. For example, in our experiment testbed, less than 30 (or 15) positive instances are used as training (or testing) data, and thus one or two misclassified instances may greatly change the rule construction and impact the final detection performance.

For this purpose, the following two self-learning rules are proposed to refine the interlaced instances in the training data set. (1) Any instance whose corresponding value in $\boldsymbol{R}_1^{An}$ is larger than the average value of $\boldsymbol{R}_1^{An}$ will be removed; (2) Any instance whose corresponding value in $\boldsymbol{R}_1^{Ae}$ is smaller than half of the average value of $\boldsymbol{R}_1^{Ae}$ will be removed.

2.3. Rule-based Data Mining

In the proposed framework, the C4.5 decision tree [8] is used for final event detection as it is a well-know rule-based algorithm good at learning the associations among different features of a set of pre-labeled instances. The filtered training data by the distance-based data mining scheme are fed into the C4.5 classifier to construct the tree model. Each data instance consists of the filtered features as well as the classification label, either "yes" for positive instances or "non" for negative instances. The filtered testing data are processed by the constructed tree model for final goal event detection.

3. EMPIRICAL STUDY

Soccer goal event detection is used as the experiment testbed to validate our proposed data mining framework. In our empirical study, 27 soccer videos were collected from different Internet broadcasters from 1998 to 2003, including 7 FIFA2003 soccer videos and 20 other videos, with a total time duration of 9 hours and 28 minutes. Among 4885 video samples, only 41 are positive (goal event) instances, i.e., $N = 4885$, $N1 = 41$, $N2 = 4844$, and the positive instances only account for about 0.84% of the total instances. It is worth to mention that the collected videos possess lower visual/audio quality as compared to videos used in other existing studies.

3.1. Experiment Setup

The $2/3$ of the video data (18 videos) are randomly selected for training and the rest (9 videos) are adopted for testing. 10 such groups are formed randomly for 10-fold cross-validation, and thus totally 10 decision models are constructed and tested with the corresponding testing data sets. In the experiments, the parameters for self-learning are set to $T1$=$T2$=30 and K=50 as discussed in Section 2.2.1 based on empirical studies.

3.2. Performance Evaluation

After 50 times random selection and comparison, the selected typical goal instances can recognize 100% goal instances and reject about 85% non-goal instances; while the selected typical non-goal instances can recognize about 80% non-goal instances and reject 100% goal instances. As a consequence, after executing the instance and feature filtering step discussed in Section 2.2.2, only less than 180 non-goal instances remain in both training data set and testing data set without any goal instance being removed, and only 9 reconstructed features ($p' = 9$) are automatically selected. The step in Section 2.2.3 refines a small number of heavily interlaced goal instances (less than 4) and several heavily interlaced non-goal instances (less than 8) utilizing the proposed two rules. Therefore, when passing the filtered data to the C4.5 classifier, less than 130 non-goal instances and about 25 goal instances are used to construct the decision tree model.

Table 1. Cross validation results for goal detection

No.	**Goal**	**Ident.**	**Missed**	**Mis-Ident.**	**RC (%)**	**PR (%)**
1	13	13	0	8	100	61.9
2	13	11	2	1	84.6	91.7
3	13	10	3	0	76.9	100
4	13	12	1	1	92.3	92.3
5	13	10	3	1	76.9	90.9
6	14	12	2	2	85.7	85.7
7	14	11	3	1	78.6	91.7
8	14	11	3	0	78.6	100
9	14	12	2	5	85.7	70.6
10	14	11	3	2	78.6	84.6
Avg.					**83.8**	**86.9**

Table 1 shows the goal event detection performance of our proposed data mining framework, where "RC", "PR", and "Ident." denote "Recall", "Precision", and "Identified". The "Missed" column indicates the number of goal instances that are misclassified as non-goal, and the "Mis-Ident." column indicates the number of non-goal instances that are misclassified as goal instances. The Recall and Precision for the goal event can be defined as:

$$Recall = Identified/(Identified + Missed);$$
$$Precision = Identified/(Identified + Mis{-}identified).$$

As can be inferred from the table, though the goal events account for less than 1% of the total data set, the average recall and precision values reach 83.8% and 86.9%, respectively, indicating the effectiveness and potential of the proposed framework. Additionally, one extra benefit of the proposed framework should be noted that the feature set has been reduced to almost one half, which brings operational benefits such as less storage requirement for multimedia database, less training time, less testing time, simplified tree model, and avoidance of "curse of dimensionality".

4. CONCLUSIONS

In this paper, an advanced framework that utilizes both the distance-based and rule-based data mining techniques for domain independent video event detection to address the *rare event detection* issue is proposed. The proposed framework is fully automatic without the need of any domain knowledge, which is achieved by data pre-processing including increasing the percentage of positive instances and reducing the feature dimension, and a decision tree classifier for event detection. The experimental results in goal event detection from multiple broadcast video data show the viability and effectiveness of the proposed framework for general event detection.

5. REFERENCES

[1] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden markov models," *Proc. IEEE IEEE International Conference on Acoustics, Speech, And Signal Processing (ICASSP)*, vol. 4, pp. 4096–4099, May 13-17, 2002.

[2] D. Tjondronegoro, Y.-P. Chen, and B. Pham, "Content-based video indexing for sports analysis," *Proc. of ACM Multimedia*, pp. 1035–1036, November 6-11, 2005.

[3] X. Yu et al., "A robust and accumulator-free ellipse hough transform," *Proc. of ACM Multimedia*, pp. 256–259, October 10-16, 2004.

[4] E.A. Tekalp and A.M. Tekalp, "Generic play-break event detection for summarization and hierarchical sports video analysis," *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 169–172, July 6-9, 2003.

[5] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna, "Semantic event detection via temporal analysis and multimodal data mining," *IEEE Signal Processing Magazine, Special Issue on Semantic Retrieval of Multimedia*, vol. 23, no. 2, pp. 38–46, March 2006.

[6] B. Li and I. Sezan, "Semantic sports video analysis: approaches and new applications," *Proc. of International Conference on Image Processing*, vol. 1, pp. 17–20, September 14-17, 2003.

[7] T. Quirino, Z. Xie, M.-L. Shyu, S.-C. Chen, and L. Chang, "Collateral representative subspace projection modeling for supervised classification," *Proc. of the 18th IEEE International Conference on Tools with Artificial Intelligence*, pp. 98–105, November 13-15, 2006.

[8] J.R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA:Morgan Kaufmann, 1993.