Video Semantic Event/Concept Detection Using a Subspace-Based Multimedia Data Mining Framework

Mei-Ling Shyu, *Senior Member, IEEE*, Zongxing Xie, *Student Member, IEEE*, Min Chen, *Member, IEEE*, and Shu-Ching Chen, *Senior Member, IEEE*

***Abstract*—In this paper, a subspace-based multimedia data mining framework is proposed for video semantic analysis, specifically video event/concept detection, by addressing two basic issues, i.e., *semantic gap* and *rare event/concept detection*. The proposed framework achieves full automation via multimodal content analysis and intelligent integration of distance-based and rule-based data mining techniques. The content analysis process facilitates the comprehensive video analysis by extracting low-level and middle-level features from audio/visual channels. The integrated data mining techniques effectively address these two basic issues by alleviating the class imbalance issue along the process and by reconstructing and refining the feature dimension automatically. The promising experimental performance on goal/corner event detection and sports/commercials/building concepts extraction from soccer videos and TRECVID news collections demonstrates the effectiveness of the proposed framework. Furthermore, its unique domain-free characteristic indicates the great potential of extending the proposed multimedia data mining framework to a wide range of different application domains.**

***Index Terms*—Data mining, eigenspace, eigenvalue, event/concept detection, principal component, video semantics analysis.**

I. Introduction

WITH the proliferation of video data and growing requests for video applications, there is an increasing need of advanced technologies for indexing, filtering, searching, and mining the vast amount of videos such as event detection and concept extraction. Here, video events are normally defined as the interesting events which capture users' attentions (i.e., goal events, traffic accidents, etc.); whereas the concepts refer to high-level semantic features, like "commercials," "sports," etc. [22].

In the literature, most of the state-of-the-art event detection frameworks were conduced toward the videos with loose structures or without story units, such as sports videos, surveillance videos, or medical videos [27]. In contrast, the concept-extraction schemes were largely carried out on the news videos which have content structures. One of the typical driven forces is the creation of the TRECVID benchmark by the National Institute of Standards and Technology, which aims to boost the researches in semantic media analysis by offering a common video corpus and a common evaluation procedure. Most of such studies are conducted in a two-stage procedure [11]. We name the first stage as video content processing, where the video clip is segmented into certain analysis units and their representative features are extracted. The second stage is called the decision-making process that extracts the semantic index from the feature descriptors to improve the framework robustness. An overview of the related work is described in Section II.

Despite the numerous amounts of efforts, this area is still far from maturity and many challenges exist.

- Bridging the semantic gap between low-level video features and high-level semantic concepts: most current researches for event/concept extraction rely heavily on certain artifacts such as domain-knowledge and *a priori* models [12], which largely limit their extensibility in handling other application domains and/or video sources.
- Class-imbalance (or rare event/concept detection) problem: the events/concepts of interests are often infrequent.

In this paper, we target addressing the aforementioned challenges, to some extent, with the adoption of multimodal content analysis and the intelligent integration of distance-based and rule-based data mining techniques. The major contributions of this study are summarized as follows.

- The proposed framework offers a unique solution for both event detection and concept extraction. To fully demonstrate its effectiveness, a large set of soccer videos and 2004/2005 TRECVID broadcast news videos are applied as the testbed for performance evaluation.
- The proposed framework greatly eliminates the dependency on domain knowledge and automates the process of event/concept detection. A common set of low-level and middle-level multimodal features is extracted and fed to the data mining component without the need of domain knowledge. Though the proposed framework adopts the idea of two-stage procedure, its uniqueness lies in the fact that a subspace-based (a special distance-based) data mining technique is used in the decision-making stage to prune the data and alleviate the class imbalance issue.

The remainder of the paper is organized as follows. An overview of the related work is given in Section II. The proposed framework is discussed in Section III. Section IV

Manuscript received February 10, 2007; revised August 30, 2007. The work of M.-L. Shyu was supported in part by NSF ITR (Medium) IIS-0325260. The work of M. Chen was supported in part by the Florida International University Dissertation Year Fellowship. The work of S.-C. Chen was supported in part by NSF EIA-0220562 and NSF HRD-0317692. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhongfei (Mark) Zhang.

M.-L. Shyu and Z. Xie are with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124 USA (e-mail: shyu@miami.edu; zxie@umsis.miami.edu).

M. Chen is with the Department of Computer Science, University of Montana, Missoula, MT 59812 USA (e-mail: chen@cs.umt.edu).

S.-C. Chen is with the Distributed Multimedia Information System Laboratory, School of Computing and Information Sciences, Florida International University, Miami, FL, 33199 USA (e-mail: chens@cs.fiu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2007.911830

presents the experimental settings and performance evaluation. Section V concludes this paper.

II. Related Work

Sports video analysis, especially sports events detection, has received a great deal of attention [6], [7] owing to its great commercial potentials. For video content processing, many earlier studies adopted unimodal approaches that studied the respective role of visual, audio, and texture mode in the corresponding domain. Recently, the multimodal approach attracts growing attention as it captures the video content in a more comprehensive manner. In [6], a multimodal framework using combined audio, visual, and textual features was proposed. A maximum entropy method was proposed in [10] to integrate image and audio cues to extract highlights from baseball video.

News videos are another video source which receives great attentions from the research community. News have a rather definite structure [26] which has been exploited for content analysis [25]. Especially, the idea of defining a set of semantic concepts for which detectors could be built ahead of search time has generated great interests to the researchers, including TRECVID participants. In terms of media-based features, multimodal approaches are widely adopted [14], which explore visual features, audio features, automatic speech recognition (ASR) transcript-based features, metadata, etc.

In the decision-making stage, data mining has been increasingly adopted. For instance, [23] proposed a hybrid classification method called CBROA which integrates the decision tree and association rule mining methods in an adaptive manner. However, its performance is restricted by a segmentation process and a pre-defined confidence threshold. Moghaddam and Pentland are pioneers in the introduction of principal component analysis (PCA) to the face recognition domain, and have popularized the use of PCA in supervised classification in this domain [13]. As far as video semantic analysis is concerned, support vector machines (SVM) is a well-known algorithm adopted for event detection [17] in sports videos and concept extraction [1], [19] in TRECVID videos. Although SVM presents promising generalization performance, its training process does not scale well as the size of the training data increases [9]. C4.5 [15] is a matured representative data mining method, which was also applied in sports video analysis [5].

Generally speaking, there exist diverse measures to organize the data mining procedure like distance-based, rule-based, instance-based, statistic-based, etc. Among them, distance-based and rule-based are the two basic and widely used classification measures. Different data mining measures have different merits and applicable domains. As the video event/concept detection application is inherently challenging, any existing individual data mining measure can hardly fulfill the task well without the support of certain artifacts as shown in most of the current researches. Though some generalized video event/concept extraction approaches have been conducted, their detection capability is limited [21] due to the well-known *semantic gap* and *rare event/concept detection* issues [4]. The *rare event/concept detection* (also known as *imbalance data set*) issue occurs when there are a very small percentage of positive instances while the large number of negative instances dominate the detection model training process. This issue usually results in a undesirable degradation of the detection performance. Combining different measures in a new framework may offer a potential solution as it utilizes multiple merits and extends applicable domains.

In our proposed framework, we aim at automating the video event/concept detection procedure via the combination of distance-based and rule-based data mining techniques. Specifically, our previously proposed distance-based RSPM algorithm [16] is improved to perform the rough classification including the noise/outlier filtering and feature combination and selection. Then, the well-known rule-based algorithm C4.5 decision tree [15] is employed for further classification. In essence, one of the unique characteristics of the proposed framework is its capability of addressing the *rare event/concept detection* and *semantic gap* issues without relying on the artifacts or domain knowledge.

III. The Proposed Framework

As shown in Fig. 1, our proposed framework consists of three major components.

A. Video Parsing and Feature Extraction

Video parsing, or called syntactic segmentation, involves temporal partitioning of the video sequence into meaningful units which then serve as the basis for descriptor extraction and semantic annotation. In this study, shots are adopted as the basic syntactic unit as they are widely accepted as a self-contained and well-defined unit, where our shot-boundary detection algorithm [3] consisting of pixel-histogram comparison, segmentation map comparison, and object tracking is employed. In essence, the differences between consecutive frames are compared in terms of their pixel/histogram values, segmented regions characteristics and foreground objects' size/location, and shot boundary is detected when the difference reaches a certain threshold. Here, the segmentation map and object information are extracted using the simultaneous partition and class parameter estimation (SPCPE) unsupervised object segmentation method [2].

In terms of feature extraction, multimodal features (visual and audio) are extracted for each shot based on the detected shot boundaries. Totally, five visual features are extracted for each shot, namely *pixel_change*, *histo_change*, *background_mean*, *background_var*, and *dominant_color_ratio*. Here, *pixel_change* denotes the average percentage of the changed pixels between the consecutive frames within a shot and *histo_change* represents the mean value of the frame-to-frame histogram differences in a shot. Another visual feature is the *dominant_color_ratio* [4] that represents the ratio of dominant color in the frame based on histogram analysis and is widely used for shot classification. Then region-level analysis is conducted based on segmentation results (the background and foreground regions identified by SPCPE). The features *background_mean* and *background_var* are therefore used to capture shot-level standard deviation and mean color values for each segmented frame, respectively.

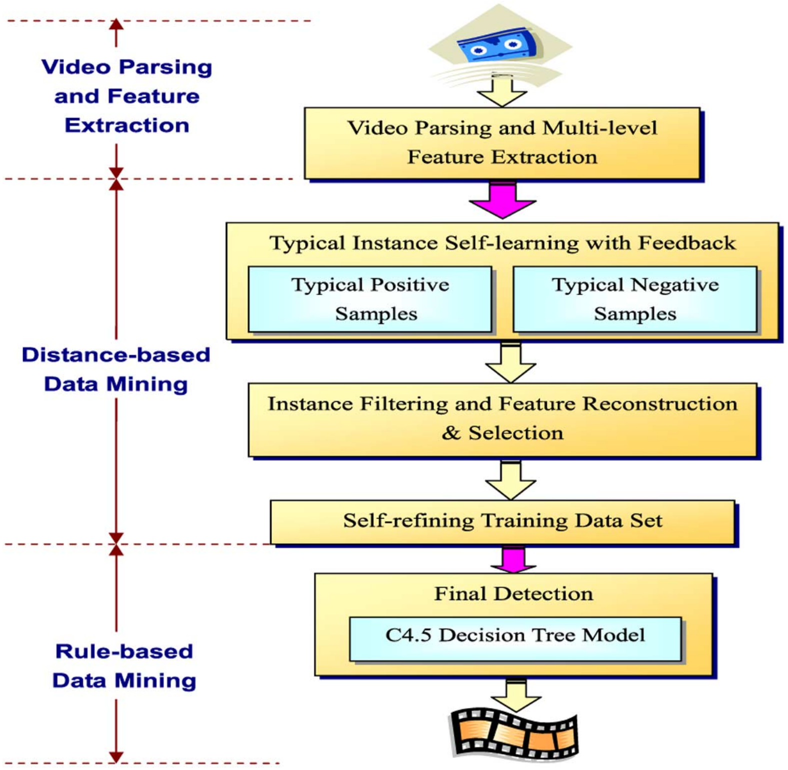

Video Parsing and Feature Extraction
Video Parsing and Multi-level Feature Extraction
Typical Instance Self-learning with Feedback
Typical Positive Samples
Typical Negative Samples
Distance-based Data Mining
Instance Filtering and Feature Reconstruction & Selection
Self-refining Training Data Set
Final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

Fig. 1. Architecture of the proposed framework.

The representations of the audio features in this framework are exploited in both time-domain and frequency-domain, dividing into three distinct groups (volume related features, energy related features, and spectrum flux related features). Totally, tengeneric audio features are utilized (one volume, five energy, and four spectrum flux features). Here, volume is an indication of the loudness of sound; short time energy means the average waveform amplitude defined over a specific time window; and Spectral Flux is defined as the 2-norm of the frame-to-frame spectral amplitude difference vector. For each category, some statistical attributes such as mean and standard deviation are captured as the corresponding features. A complete list of all the audio features as well as their feature descriptions is presented in [5]. In addition, due to the reason that the audio track can be continuous even around the shot boundary, the volume statistics information (mean and max) is captured for the duration of 3 s around the shot boundary to explore the audio track information. In summary, the constructed feature set F consists of 17 features (ten audio features, five visual features, and two temporal features) [4].

B. *Distance-Based Data Mining*

It is frequently observed that the *rare event/concept detection* issue arises since the video data amount is typically huge and the ratio of the event/concept instances to the negative instances is typically very small (e.g., only less than 1:100 in our goal event detection empirical studies). Accordingly, it would be difficult for a typical detection process to capture such a small portion of targeted instances from the huge amount of data especially with the existence of the noisy and irrelevant information introduced during the video production and feature extraction processes. Therefore, before performing the actual detection process, a pre-filtering process is needed to trim as many negative instances as possible.

Motivated by the powerfulness and robustness of our previously proposed distance-based anomaly detection algorithm called representative subspace projection modeling (RSPM) [16], a series of novel automatic distance-based data mining schemes are proposed in this paper to eliminate a great portion of negative instances and thus to overcome the *rare event/concept detection* issue. In brief, it contains three automatic schemes: typical instance self-learning with feedback; instance filtering and feature reconstruction and selection; and self-refining training data set.

1) Typical Instance Self-Learning With Feedback: It is well acknowledged that noisy or irrelevant data contained in the original data set would adversely impact the detection performance. Especially, when the number of interesting event/concept instances is much less than that of the negative instances, the positive instances might be over-shadowed by their counterpart during the training process. While most of the other researchers used domain-based information to address this issue, we propose the use of two basic concepts (**typical positive instances** and **typical negative instances**) in the proposed feedback based self-learning typical instances selector.

Assume that the original data set have p features and N instances. Thus, let $\mathbf{X} = \{\mathrm{x}_{ij}, i = 1,2,\ldots,p \text{ and } j = 1,2,\ldots,N\}$, which is a $p \times N$-dimensional matrix containing N p-dimensional column vectors $\mathsf{X}_j = (\mathrm{x}_{1j}, \mathrm{x}_{2j}, \ldots, \mathrm{x}_{pj})'$,

$j = 1, 2, \ldots, N$, which represent N original instances after video parsing and features extraction process. Let $\bar{\mathsf{X}} = (1/N)\sum_{j=1}^{N}\mathsf{X}_j$, the robust correlation matrix is defined as $\mathbf{S} = (1/N-1)\sum_{j=1}^{N}(\mathsf{X}_j - \bar{\mathsf{X}})(\mathsf{X}_j - \bar{\mathsf{X}})'$.

The normalization of the training data set is required to minimize the feature scale effects for multivariate data. Let $\bar{\mu}_i$ and s_{ii} be the sample mean and variance of the ith row of the matrix $\mathbf{X}$, that is, of the ith feature of the data set, and x_{ij} $(i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, N)$ be the elements in the matrix $\mathbf{X}$. Thus, we can define the normalized training data set with p features as $\mathbf{Z} = \{\mathrm{z}_{ij}, i = 1, 2, \ldots, p \text{ and } j = 1, 2, \ldots, N\}$ with the corresponding column vectors $\mathsf{Z}_j = (\mathrm{z}_{1j}, \mathrm{z}_{2j}, \ldots, \mathrm{z}_{pj})'$, $j = 1, 2, \ldots, N$, representing the normalized data instances, by utilizing the normalization function given in (1)

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

Among the original data set, assume that there are $N1$ labeled positive instances $\mathbf{X}^1 = \{\mathrm{x}_{ij}\}$ and $N2$ labeled negative instances $\mathbf{X}^2 = \{\mathrm{x}_{ij}\}$. The proposed **typical positive instances** are defined as selected $T1$ $(T1 < N1)$ positive instances in $\mathbf{X}^e = \{\mathrm{x}_{ij}\}$. The normalized matrix $\mathbf{Z}^e = \{\mathrm{z}_{ij}\}$ of $\mathbf{X}^e$ can be obtained via (1), where $\bar{\mu}_i$ and s_{ii} are the sample mean and variance of the ith row in $\mathbf{X}^e$. Let $(\lambda_1^e, \mathbf{E}_1^e), (\lambda_2^e, \mathbf{E}_2^e), \ldots, (\lambda_p^e, \mathbf{E}_p^e)$ be the p typical positive eigenvalue-eigenvector pairs of the robust correlation matrix $\mathbf{S}$ for $\mathbf{Z}^e$, composing a sample $p \times p$ typical correlation matrix, where $\lambda_1^e \geq \lambda_2^e \geq \cdots \geq \lambda_p^e \geq 0$. Also, let $\mathbf{Y}^e = \{\mathrm{y}_{ij}, i = 1, 2, \ldots, p \text{ and } j = 1, 2, \ldots, N\}$ be the projection of $\mathbf{Z}^e$ onto the p-dimensional typical positive eigenspace. $\mathbf{Y}^e$ is called the typical instance score matrix and is composed of score column vectors $\mathsf{Y}_j = (\mathrm{y}_{1j}, \mathrm{y}_{2j}, \ldots, \mathrm{y}_{pj})'$, $j = 1, 2, \ldots, N$, that correspond to the projection of each of the N normalized typical positive instances onto the typical positive eigenspace. The ith sample score value of the jth normalized observation vector Z_j is given by:

$$\mathrm{y}_{ij} = \mathbf{E}_i^{e\prime}\mathsf{Z}_j = e_{i1}\mathrm{z}_{1j} + e_{i2}\mathrm{z}_{2j} + \cdots + e_{ip}\mathrm{z}_{pj} \tag{2}$$

where $\mathbf{E}_i^e = (e_{i1}, e_{i2}, \ldots, e_{ip})'$ is the ith typical positive eigenvector.

We define the p score row vectors of $\mathbf{Y}^e$, representing the distribution of the p eigenspace features of all the normalized and projected typical positive instances, as $\mathbf{R}_i^e = (\mathrm{y}_{i1}, \mathrm{y}_{i2}, \ldots, \mathrm{y}_{iN})$, $i = 1, 2, \ldots, p$. The representative components used to model the similarity [16] of the typical positive instances are selected based on the generally used standard deviation concept with (3).

$$\mathrm{STD}(\mathsf{R}_m^e) < a, \text{ where} \tag{3}$$

- $\mathrm{STD}(\mathsf{R}_m^e)$ is the standard deviation of the score row vectors satisfying (3);
- a is the arithmetic mean of the standard deviation values from all $\mathbf{R}_i^e$; and
- $m \in \mathbf{M}$ is defined as the selected typical positive principal component space.

A class-deviation equation ((4)) is designed to differentiate the normal and anomaly instances in the view of typical positive instances.

$$\mathrm{c}_j^e = \sum_{m \in M} \frac{(\mathrm{y}_{mj})^2}{\lambda_m^e}, \text{ where} \tag{4}$$

- $m \in \mathbf{M}$ is the index of the mth principal component satisfying the condition in (3);
- λ_m^e is the eigenvalue of the corresponding mth principal component;
- y_{mj} is the score value of the mth feature in the selected principal component space for the jth typical positive instance, and;
- c_j^e, $j = 1, 2, \ldots, N$, is the threshold value for each observation in the typical positive data set.

Thus, the maximum value $\mathrm{c}_{\max}^e$ of all 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. Assume that the score matrix $\mathbf{Y} = \{\mathrm{y}_{ij}\}$ is the projection of matrix $\mathbf{Z}$ (the normalized matrix of $\mathbf{X}$ via (1)) onto the typical positive eigenspace, the decision rules to classify each of the observations X_j are established naturally based on $\mathrm{c}_{\max}^e$:

Classify the jth instance as abnormal to the typical positive instances if $\mathrm{c}_j > \mathrm{c}_{\max}^e$; and

Classify the jth instance as normal to the typical positive instances if $\mathrm{c}_j \leqslant \mathrm{c}_{\max}^e$.

Accordingly, if we randomly select $T1$ instances from $\mathbf{X}^1$ for K times and for 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, the best group can be found which can recognize maximum instances in $\mathbf{X}^1$ and at the same time reject the maximal percentage of instances in $\mathbf{X}^2$. If we set $T1 = N1$, a 100% recognizing percentage can be reached. Accordingly, if K and $T1$ are big enough, the best group can be defined as $\mathbf{X}^e$, the one that can recognize 100% instances in $\mathbf{X}^1$ and at the same time reject the maximal percentage of instances in $\mathbf{X}^2$.

Similarly, as the negative instances are very huge and thus provide more selection space for typical negative instances, we can define the proposed **typical negative instances** $\mathbf{X}^n$ 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 $\mathbf{X}^2$. Accordingly, the typical negative eigenvalue-eigenvector pairs $(\lambda_1^n, \mathbf{E}_1^n), (\lambda_2^n, \mathbf{E}_2^n), \ldots, (\lambda_p^n, \mathbf{E}_p^n)$ and their typical negative eigenspace and the index vector $\mathbf{M}'$ for typical negative instances can also be obtained automatically.

2) Instance Filtering and Feature Reconstruction & Selection: 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). Further assume that $\mathbf{X}^{Ae}$ and $\mathbf{X}^{An}$ are the labeled positive and negative instances in $\mathbf{X}^A$, respectively. Then $\mathbf{X}^e$ and $\mathbf{X}^n$ are trained sequentially to pre-filter negative instances in both training and testing data sets to address the *rare event/concept detection* issue.

First, $\mathbf{X}^e$ is trained to reject negative instances in $\mathbf{X}^{An}$ and $\mathbf{X}^B$. Here, no positive instances would be removed in both training data set and testing data set since $\mathbf{X}^e$ is defined to recognize 100% instances in $\mathbf{X}^1$. Second, the recognized normal data instances are passed to the second classifier trained with the selected $\mathbf{X}^n$. 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. Similarly, none of the positive instances would be removed since $\mathbf{X}^n$ is defined to reject 100% instances in $\mathbf{X}^1$. Accordingly, a great part of negative instances are refined via suitable selection of $\mathbf{X}^e$ and $\mathbf{X}^n$ while all the positive instances are kept, which greatly alleviates the *rare event/concept detection* issue as will be demonstrated in the experiments.

Finally, all the remaining $\mathbf{X}^{An}$, $\mathbf{X}^{Ae}$, and $\mathbf{X}^B$ are projected onto the typical negative eigenspace and the score values corresponding to the index vector $\mathbf{M}'$ are extracted to replace the original feature set to simplify the next rule-based training model setup with the extra benefits of feature reduction (fewer features). In other words, the original feature set F are reconstructed and automatic filtered to be F' with the dimension p' $(p' < p)$. For further explanation, we can define the remaining and reconstructed data matrix as follows:

- $\mathbf{Y}^{Ae} = \{\mathtt{Y}_{ij}\}$ represents the positive instances in the training data set and R_1^{Ae} corresponds to the first selected principal component of $\mathbf{Y}^{Ae}$;
- $\mathbf{Y}^{An} = \{\mathtt{Y}_{ij}\}$ represents the negative instances in the training data set and R_1^{An} corresponds to the first selected principal component of $\mathbf{Y}^{An}$; and
- $\mathbf{Y}^B = \{\mathtt{Y}_{ij}\}$ represents all testing data, $i = 1, 2, \ldots, p'$.

3) Self-Refining Training Data Set: Since the training data set has a great influence on the later rule-based data mining process, the training data set is further refined in this phase. As the number of positive instances is very limited in video event/concept detection domain, several mislabeled or unfriendly training data may degrade the training model greatly. Therefor, a linear analysis method is proposed to refine these instances. It is a self-refining process for the training data set based on the first dimension of typical negative eigenspace which is always selected automatically as it presents the most data information. In this dimension, in most situations, the major parts of R_1^{Ae} and R_1^{An} values are approximately self-clustered with certain interlaced areas as they possess different class labels. However, for those that are interlaced, they might greatly degrade the performance of the rule-based classifier. For example, in our experiment study in goal event detection, less than 30 positive instances are used as training data, and thus one or two misclassified instances may greatly change the rule construction and impact the final detection performance. Thus, it is promising to filter these heavily interlaced instances before the rule-based data mining process.

As we use the typical negative eigenspace for obtaining the projected scores, most of R_1^{Ae} (the relative anomaly ones) would have a higher value than that of R_1^{An}. Accordingly, the following two self-learning rules are proposed to refine the heavily interlaced instances in the training data set.

- Any instance whose corresponding value in R_1^{An} is larger than the average value of R_1^{An} will be removed;
- Any instance whose corresponding value in R_1^{Ae} is smaller than half of the average value of R_1^{Ae} will be removed.

C. Rule-Based Data Mining

The C4.5 decision tree [15] is a classifier in the form of a tree structure, where each node is either a leaf node, indicating the value of the target class from observations, or a decision node, which specifies certain tests to be carried out on a single attribute-value and which is proceeded by either a branch or a sub-tree for each of the possible outcomes of the test [15]. Its main classification procedure is first to construct a model for each class in terms of the attribute-value pairs and thus use the induced model to categorize any incoming testing instance. The construction of a decision tree is performed through the so-called "divide-and-conquer" approach, i.e., recursively partition the training set with respect to certain criteria until all the instances in a partition have the same class label, or no more attributes can be used for further partitioning. The derived model summarizes all given information from training data set but express it in a more concise and perspicuous manner. The testing process of the decision tree is in the form of traversing a path in the built tree from the root to a certain leaf node and the corresponding class label is assigned to the instance when it reaches a leaf node.

In our proposed framework, given the resulting training data set from the distance-based data mining process, the C4.5 decision tree algorithm [15] is adopted to learn a classifier and the induced classification rules are represented in the form of a decision tree. In the constructed tree, each data entry consists of audio and visual features as well as the class label. The multimodal features are extracted in the feature extraction stage as discussed earlier. A "yes" or "non" class label is assigned to each shot manually, showing whether there is an interested event/concept or not.

IV. Empirical Study

The effectiveness of the proposed framework was rigorously tested upon a large experimental data set, which contains 27 soccer videos and 6 TRECVID videos. The total duration of the soccer videos is about 540 minutes, and these videos were obtained from a variety of sources with various production styles such as European Cup 1998, World Cup 2002, and FIFA 2003. In the data set, the numbers of goal (or corner) shots and non-goal (or non-corner) shots are 41 (95) and 4,844 (4790), respectively, which means that the interesting events account for only around 1% to 2% of the total data set. As mentioned earlier, it is quite challenging to detect these kinds of rare events.

In addition, six TRECVID videos from the broadcast news domain (CNN, ABC) are used as the testbed for concept detection. Three concepts, building, commercials, and sports are selected as the target concepts since they vary greatly from each others in terms of production styles, occurrence frequency (2%, 39% and 3%, respectively, of the entire data set), etc. In the experiments, they are extracted to demonstrate both the effectiveness and generalization of the proposed framework.

A. Experimental Setup

For performance evaluation, two comparative experiments are conducted and the results are analyzed in details.

- In the first experiment, the proposed framework is applied to detect the events/concepts from the experimental data

TABLE I
PARAMETERS OF THE EXPERIMENTS

Event	N	N1	N2	T1	T2	K
Corner	4885	95	4790	30	17	50
Goal	4885	41	4844	30	30	50
Build	2304	48	2256	30	17	50
Commercial	2304	898	1406	20	20	50
Sports	2304	63	2241	17	17	50

set following the procedure discussed earlier. For performance evaluation, three approaches discussed in related papers [19], [20] are implemented and tested with the same experimental setup for comparison.

- To demonstrate the capability of the subspace based data mining approach as an independent component in tackling the class imbalance issue, in the second experiment, a set of well-acknowledged classification algorithms are applied with and without it for event/concept detection and their corresponding results are compared.

B. Comparative Experiment #1

Table I shows important parameters used by our proposed framework in the empirical study. Among them, N, $N1$ and $N2$ denote the number of total, positive, and negative instances, respectively, in the data set. The parameters $T1$, $T2$, and K are set based on empirical studies for self-learning as discussed in Section III-B1. In addition, after 50 times random selection and comparison, the selected typical instances possess the following characters: the selected typical goal, sports, commercial, building, and corner instances can recognize 100% corresponding event/concept (positive) instances and reject about 85%, 74%, 36%, 49%, and 72% non-event/non-concept (negative) instances respectively; while the selected typical negative instances can recognize about 80%, 83%, 67%, 71%, and 72% negative instances and reject 100% positive instances.

In order to better evaluate our proposed framework, five-fold cross-validation are used. That is, the 2/3 of the video data are randomly selected for training and the rest are adopted for testing. Accordingly, for each empirical study, totally five decision models are constructed and tested with the corresponding testing data sets.

Table II demonstrate the performance of our proposed distance-based data mining phase for addressing the *rare event/concept detection* issue. "Pre-p." and "Post-P." denote the percentage of the positive instances before and after this phase. p and p' are the parameters discussed in Section III-B2. From the table, we can see that the percentage of positive instances increases greatly to solve the *rare event/concept detection* issue and thus provides a suitable platform for the rule-based algorithm. Furthermore, the dimension of the features is reduced to about 50%, which brings several 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."

Table III shows the event/concept detection performance of corner, goal, building, commercial, and sports of our proposed framework, respectively. In these tables, "RC," "PR," "Cor," "Build," and "Comm" denote recall, precision, corner event, building, and commercial concepts, respectively.

TABLE II
PERFORMANCE OF THE DISTANCE-BASED DATA MINING PHASE

Event	Pre-P. (%)	Post-P. (%)	p	p'
Corner	1.94	21.94	17	9
Goal	0.84	26.81	17	9
Build	2.08	55.17	17	9
Commercial	38.98	74.65	17	9
Sports	2.73	37.72	17	8

TABLE III
CROSS VALIDATION RESULTS FOR EVENT/CONCEPT DETECTION

	Cor	Goal	Build	Comm	Sports
RC1 %	53.9	84.6	61.5	88.2	85.2
PR1 %	25.5	91.7	50.0	78.4	74.2
RC2 %	49.5	92.3	53.3	91.8	78.6
PR2 %	29.6	92.3	61.5	83.8	78.6
RC3 %	69.4	85.7	71.4	89.2	86.4
PR3 %	22.7	85.7	62.5	81.6	73.1
RC4 %	62.5	85.7	57.1	87.1	81.5
PR4 %	25.9	70.6	53.3	77.4	88.0
RC5 %	65.2	78.6	75.0	90.0	80.5
PR5 %	33.4	84.6	53.0	81.8	89.2
RC_Avg. %	60.1	85.4	63.7	89.3	82.4
PR_Avg. %	27.4	85.0	56.1	80.6	80.6

In many existing work, the average precision results for sports and building concepts detection from TRECVID videos are around 30% [8], [19]. However, performance of such frameworks is generally associated with both the representativeness of low-level features and the capability of mapping low-level features to high-level semantic meanings. Since many details of low-level feature representations, such as the number of bins for color histogram, are not clearly stated in most existing work, it is difficult to reproduce the exact feature set. To avoid the bias, three mapping approaches, namely "ESVM," "SVML," and "LKNN" discussed in related papers, are implemented and are tested upon the features we extracted. Here, in "ESVM" [19], early fusion is performed to integrate the unimodal streams (i.e., audio or visual) into a multimedia feature representation, which serves as the inputs for support vector machines (SVMs). In contrast, "SVML" [19] performs SVM classifications on unimodal features and then the learned unimodal concept detection scores are combined for final decision [18]. Finally, "LKNN" [20] uses K-nearest neighbor's method to cluster shots in the development set and to classify testing data. The purpose of this experiment is to illustrate the effectiveness of the subspace based data mining approach, which is essentially the focus of this paper.

Similarly, for these three methods, the five-fold cross validation scheme is adopted and the average precision and recall values across these five-fold tests are reported in Table IV. As can be inferred from Tables III and IV, our proposed framework maintains promising performance and outperforms these three mapping approaches in almost all the event/concept detection cases. One exception is for building detection where "ESVM" and "SVML" produce a higher recall value and "LKNN" yields a better result in precision score. However, our approach achieves a better balance between recall and precision metrics. Another observation is that the precision values of corner event detection are generally low for all the approaches. By checking the false positive instances, we found that a large portion of

TABLE IV
PERFORMANCE OF THREE MAPPING APPROACHES

	Cor	**Goal**	**Build**	**Comm**	**Sports**
RC_ESVM %	43.6	61.6	73.6	61.5	53.7
PR_ESVM %	11.3	21.2	11.1	45.1	26.2
RC_SVML %	37.4	72.1	65.0	42.9	61.0
PR_SVML %	11.8	26.4	21.0	42.3	20.7
RC_LKNN %	52.1	62.2	37.0	61.8	19.0
PR_LKNN %	26.3	71.9	65.4	63.1	40.0

TABLE V
PERFORMANCE OF EVENT/CONCEPT DETECTION WITHOUT SUBSPACE-BASED DATA MINING STEP

	Cor	**Goal**	**Build**	**Comm**	**Sports**
RC_OR %	0.0	2.4	0.0	45.7	0.0
PR_OR %	0.0	20.0	0.0	53.9	0.0
RC_RF %	3.2	17.1	0.0	69.2	27.0
PR_RF %	20.0	58.3	0.0	67.3	58.6
RC_LG %	0.0	29.3	0.0	57.5	14.3
PR_LG %	0.0	54.5	0.0	71.3	40.9

TABLE VI
PERFORMANCE OF EVENT/CONCEPT DETECTION WITH SUBSPACE-BASED DATA MINING STEP

	Cor	**Goal**	**Build**	**Comm**	**Sports**
RC_OR %	11.6	48.6	32.6	72.4	58.7
PR_OR %	28.9	72.0	71.4	66.7	80.4
RC_RF %	31.6	67.6	32.6	88.0	60.3
PR_RF %	49.2	75.8	60.0	79.3	74.5
RC_LG %	27.4	54.1	15.2	83.6	54.0
PR_LG %	55.3	74.1	100.0	76.9	81.0

them belong to either goal kicks or line-throws whose broadcast patterns are quite similar to that of corner events.

In summary, the performance of our framework remains reasonably good and consistent without the inference of any domain knowledge. This experiment demonstrates the effectiveness and potential of the proposed framework in the domain of video semantic analysis.

C. Comparative Experiment #2

In the second experiment, a group of well-recognized data classification methods, such as one rule classifier (for short, OR), random forest (RF), and logistic (LG) [24] are adopted for event/concept detection with and without the data pruning process (i.e., subspace data mining step).

As can be seen from Table V, all the data classifiers perform poorly without the data pruning process. In many cases, none of the event/concept unit can be detected and thus the recall values equal to zero. In contrast, Table VI shows the results of conducting data classification after the data pruning process. Clearly, both recall and precision scores increase dramatically in all cases. This experiment thus demonstrates the capability of the subspace based data mining approach as an independent component in tackling the rare event/concept detection issue. On the other hand, we also show that a better detection performance can be achieved by reducing class imbalance situations.

V. CONCLUSIONS

Video event/concept detection is of great importance in video indexing, retrieval, and summarization. However, the *semantic gap* and *rare concept/event detection* issues inhibit the viability of the existing approaches in diverse event/concept detection domains. To address these issues, in this paper, a novel subspace-based multimedia data mining framework is proposed that utilizes the multimodal content analysis and the distance-based and rule-based data mining techniques. One of the unique contributions of the proposed framework is that it is automatic without the need of domain knowledge and thus can be easily extended to various application domains. The relax of the domain knowledge is achieved by adopting several distance-based data mining schemes to alleviate the class imbalance issue and to reconstruct and reduce the feature dimension. Thereafter, the C4.5 decision tree is employed to construct the training model for the final event/concept detection. The experimental results in Section IV demonstrate the effectiveness and adaptivity of the proposed framework for concept/event detection.

In our future work, this framework will be tested and extended in more concept/event detection applications, such as detecting significant events from surveillance videos and other important concepts (indoors, outdoor, landscape, etc.) from the TRECVID videos.

REFERENCES

[1] A. Amir *et al.*, "IBM research TRECVID-2003 video retrieval system," in *NIST TRECVID*, 2003.

[2] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *Int. J. Artific. Intell. Tools*, vol. 10, no. 4, pp. 715–734, Dec. 2001.

[3] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative shot boundary detection for video indexing," in *Video Data Management and Information Retrieval*, S. Deb, Ed. Hershey, PA: Idea Group Publishing, 2005, pp. 217–236.

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

[5] S.-C. Chen, M.-L. Shyu, C. Zhang, and M. Chen, "A multimodal data mining framework for soccer goal detection based on decision tree logic," *Int. J. Comput. Applic. Technol.*, vol. 27, no. 4, pp. 312–323, 2006.

[6] S. Dagtas and M. Abdel-Mottaleb, "Extraction of TV highlights using multimedia features," in *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, Cannes, France, 2001, pp. 91–96.

[7] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796–807, Jul. 2003.

[8] S. Gao, X. Zhu, and Q. Sun, "Exploiting concept association to boost multimedia semantic concept detection," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Honolulu, HI, 2007, vol. 1, pp. 981–984.

[9] B. Han, Support Vector Machines Center for Information Science and Technology, Temple University, Philadelphia, PA, 2003 [Online]. Available: http://www.ist.temple.edu/~vucetic/cis526fall2003/lecture8.doc

[10] M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," in *Proc. ACM Int. Conf. Multimedia*, Juan les Pins, France, 2002, pp. 347–350.

[11] R. Leonardi, P. Migliorati, and M. Prandini, "Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled Markov chains," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 634–643, May 2004.

[12] B. Li and I. Sezan, "Semantic event detection via temporal analysis and multimodal data mining," in *Proc. Int. Conf. Image Processing*, Barcelona, Spain, 2003, vol. 1, pp. 17–20.

[13] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, Jul. 1997.

[14] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proc. 12th ACM Int. Conf. Multimedia*, New York, 2004, pp. 660–667.

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

[16] T. Quirino, Z. Xie, M.-L. Shyu, S.-C. Chen, and L. Chang, "Collateral representative subspace projection modeling for supervised classification," in *Proc. 18th IEEE Int. Conf. Tools with Artificial Intelligence (ICTAI'06)*, Washington, DC, Nov. 13–15, 2006, pp. 98–105.

[17] D. Sadlier and N. E. O'Connor, "Event detection in field-sports video using audio-visual features and a support vector machine," *IEEE Trans. Circuits Syst. Video Technology*, vol. 15, no. 10, pp. 1225–1233, Oct. 2005.

[18] C. G. M. Snoek *et al.*, "Early versus late fusion in semantic video analysis," in *Proc. ACM Int. Conf. Multimedia*, Singapore, 2005, pp. 399–402.

[19] C. G. M. Snoek, M. Worring, J. C. Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 421–430.

[20] M. Srikanth, M. Bowden, and D. Moldovan, "LCC at TRECVID 2005," in *Proc. NIST TRECVID 2005*.

[21] E. A. Tekalp and A. M. Tekalp, "Generic play-break event detection for summarization and hierarchical sports video analysis," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Baltimore, MD, 2003, pp. 169–172.

[22] D. Tjondronegoro, Y.-P. Chen, and B. Pham, "Content-based video indexing for sports analysis," in *Proc. ACM Int. Conf. Multimedia*, Singapore, 2005, pp. 1035–1036.

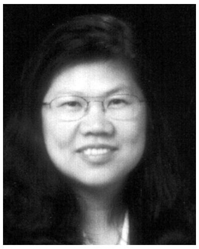
[23] V. S. Tseng, C.-J. Lee, and J.-H. Su, "Classify by representative or associations (CBROA): A hybrid approach for image classification," in *Proc. 6th Int. Workshop on Multimedia Data Mining: Mining Integrated Media and Complex Data*, Chicago, IL, Aug. 2005, pp. 37–53.

[24] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA: Morgan Kaufman, 2005.

[25] X. Wu, C.-W. Ngo, and Q. Li, "Threading and auto-documentary in news videos," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 59–68, Mar. 2006.

[26] Z. Xiong, X. Zhou, Q. Tian, Y. Rui, and T. S. Huang, "Semantic retrieval of video," *IEEE Signal Process. Mag. (Special Issue on Semantic Retrieval of Multimedia)*, vol. 23, no. 2, pp. 18–27, Mar. 2006.

[27] 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 Trans. Knowl. Data Eng.*, vol. 17, no. 5, pp. 665–677, May 2005.



Mei-Ling Shyu (M'95–SM'03) received the Master's degrees in computer science, electrical engineering, and restaurant, hotel, institutional, and tourism management from Purdue University, West Lafayette, IN, in 1992, 1995, and 1997, respectively, and the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, in 1999.

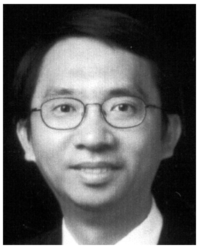
She has been an Associate Professor with the Department of Electrical and Computer Engineering (ECE), University of Miami (UM) Miami, FL, since June 2005. Prior to that, she was an Assistant Professor in ECE at UM dating from January 2000. Her research interests include data mining, multimedia database systems, multimedia networking, database systems, and security. She has authored and co-authored more than 140 technical papers published in various prestigious journals, book chapters, and refereed conference/workshop/symposium proceedings.

Dr. Shyu has been a Guest Editor of several journal special issues and a Program Chair/Co-Chair of several conference/workshop professional meetings.

Zongxing Xie (S'06) received the Master's degree from the Department of Electrical and Computer Engineering, University of Miami, Miami, FL, in May 2007.

Min Chen (M'02) received the Master's and Ph.D. degrees from the School of Computing and Information Sciences, Florida International University (FIU), Miami, in August 2004 and May 2007, respectively.

She is an Assistant Professor in the Department of Computer Science, University of Montana, Missoula. Her research interests include distributed multimedia database systems, image and video database retrieval, and multimedia data mining.



Shu-Ching Chen (M'95–SM'04) received the Master's degrees in computer science, electrical engineering, and civil engineering in 1992, 1995, and 1996, respectively, and the Ph.D. degree in 1998, all from Purdue University, West Lafayette, IN.

In August 1999, he was an Assistant Professor in the School of Computing and Information Sciences (SCIS), Florida International University (FIU), Miami, and since August 2004, has been an Associate Professor. His main research interests include distributed multimedia database management systems and multimedia data mining. He has authored and co-authored more than 170 research papers in journals, refereed conference/symposium/workshop proceedings, and book chapters. He is one of the co-authors of a book entitled *Semantic Models for Multimedia Database Searching and Browsing* (Norwell, MA: Kluwer, 2000).

Dr. Chen received the Best Paper Award from 2006 IEEE International Symposium on Multimedia. He was awarded the IEEE Systems, Man, and Cybernetics Society's Outstanding Contribution Award in 2005 and was co-recipient of the IEEE Most Active SMC Technical Committee Award in 2006. He was also awarded the Excellence in Graduate Mentorship Award from FIU in 2006, the University Outstanding Faculty Research Award from FIU in 2004, the Outstanding Faculty Service Award from SCIS in 2004, and the Outstanding Faculty Research Award from SCIS in 2002. He has been a General Chair or Program Chair for more than 20 conferences, symposiums, and workshops.