ISSN:0975-9646

A Comparison of Rule based and Distance Based Semantic Video Mining

D.Raghu, K.sandeep , CH. Raja Jacob
Dept. of Computer Science and Engineering,
NOVA College of Engineering&Tech

***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 socier 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**

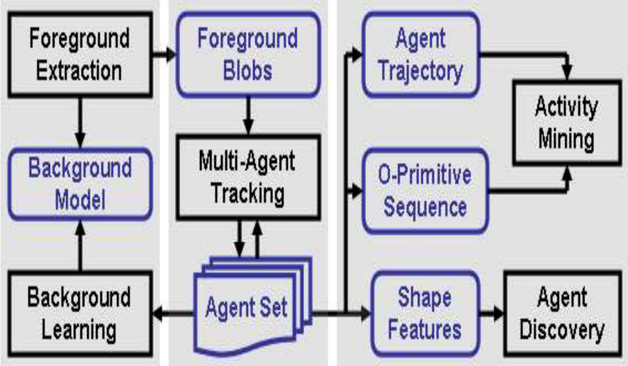
Key words: Data mining, eigenspace, eigenvalue, event/conceptdetection, principal component, video semantics analysis. *Video Parsing and Feature Extraction, Distance-Based Data Mining.*

1. INTRODUCTION:

We believe that the categorization of videos can be achieved by exploring the concepts and meanings of the videos. This task requires bridging the gap between low-level contents and high-level concepts. Once a relationship is developed between the computable features of the video and its semantics, the user would be allowed to navigate through videos by ideas instead of the rigid approach of content matching. However, this relationship must follow the norms of human perception and abide by the rules that are most often adhered to by the creators (directors) of these videos. These rules are generally known as *Film Grammar* in video production literature. Like any natural language, this grammar also has several dialects, but is fortunately, more or less universal. For example, most television game shows share a common pattern of transitions among the shots of host and guests, governed by the grammar of the show. Similarly, a different set of rules may be used to film a dialogue between two actors as compared to an action scene in a feature movie.

1.1Surveillance Video Mining:

Systems that attempt to learn conceptual categories from image schema can be constituted into two classes. First, *Visiononly systems* that use visual priors to construct models for object classes and second, *Multi-modal systems* that combine audio/text and video streams to match image categories with commentaries/annotations [13, 12]. Vision-only systems construct models that require supervised input in the form of object/activity priors. In the second class, co-occurrent language streams are used to enhance the feature space used for categorization, and correlation is often mediated by attentive focus The main thrust of the work is image sequence mining using agent tracking, with special emphasis on merging, de-merging, and other occlusion events. We attempt to learn agent and activity universals with neither supervisory visual input nor any parallel linguistic tokenization. The system combines online background learning and visual motion detection to characterize agents based on their color distribution, shape, as well as trajectories. All features are initialized from the foreground blobs at their very _rst instant of isolated identi_cation. These are tracked further in the image sequences with motion based prediction initialized mean-shift trackers [3]. The agent position information are then fed back to the online background learning system for selective adaptation of the background model.



The functional block diagram of the proposed system. Background modeling is performed in the lowest layer (left column) of processing with feedbacks from multi-agent tracking at the mid-level. The agent discovery and activity mining are performed in the higher layer (right column) by processing agent features obtained in the mid layer.

The agent feature set are tracked across the scenes and the events are characterized in terms of a set of occlusion primitives (O-primitives, henceforth): *isolation* (not connected to any other agents), *partial occlusion* (by background objects), *merging* (with other agents resulting to partial or full occlusions), *disappearance* (due to complete occlusion by background objects or track loss), and *exit* (occlusion by screen boundary) *entry* (the initial emergence of the agent). Additionally, we use a window of attention around each agent in attentive focus and trace the appearance and disappearance of other agents in this window. This set of attention window objects constitutes an important feature for discovering (homo) heterogeneous multi-agent interactions, and this proximity also aids in characterizing the occlusion when it occurs.

2. EXISTING SYSTEM:

In the literature, most of the state-of-the-art event detection frameworks were conducted toward the videos with loose structures or without story units, such as sports videos, surveillance videos, or medical videos. 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. 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. For video content processing, many existing approaches adopted uni modal approaches, its not fulfill the upcoming technique in visual as well as audio. News videos are another video source which receives great attentions from the research community. News have a rather definite structure which has been exploited for content analysis . 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, 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 . As far as video semantic analysis is concerned, support vector machines (SVM) is a well-known algorithm adopted for event detection 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 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 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 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 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.

3. PROPOSEDWORK

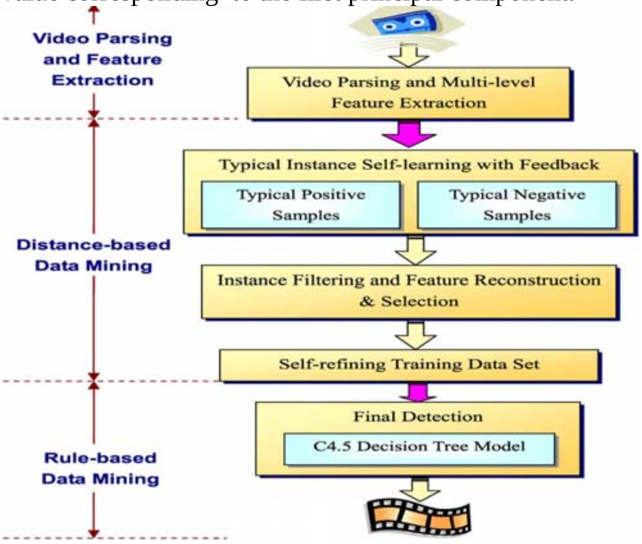
Video parsing, or called syntactic segmentation, involves temporal partitioning of the video sequence into meaningful units which then serve as the basis for descriptor extraction and semantic annotation. In this study, shots are adopted as the basic syntactic unit as they are widely accepted as a self-contained 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.

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 profiteering 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.

Distance-based DataMining

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 distancebased 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.


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

4. EXPERIMENTAL RESULTS

In order to implement the required Bayesian Network algorithm (HAR), we selected to use c# language in .Net environment. For this purpose, a text matrix file of association rules on confidence values in percentages is used. The following are the results of Bayesian Network algorithm implementation.

The following Figures (Figure 4.1 and Figure 4.2) show the comparison of performance of Rule based and Distance based Comparison

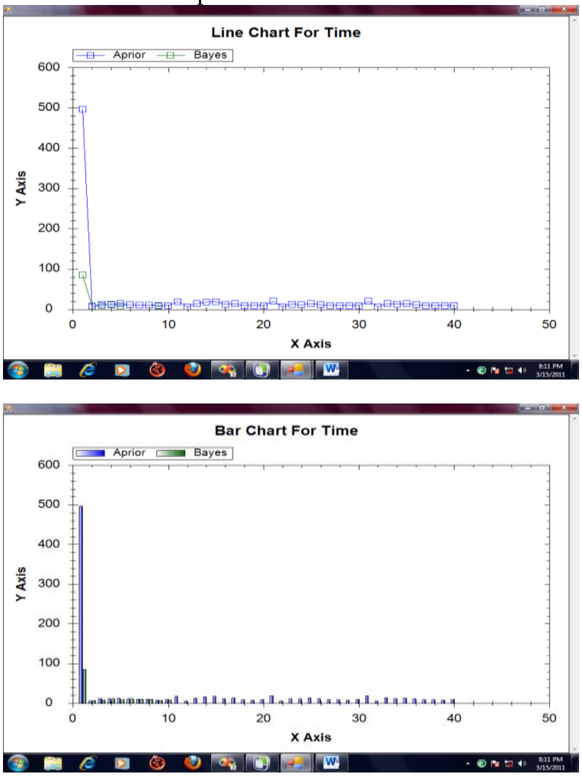


Figure 4.1

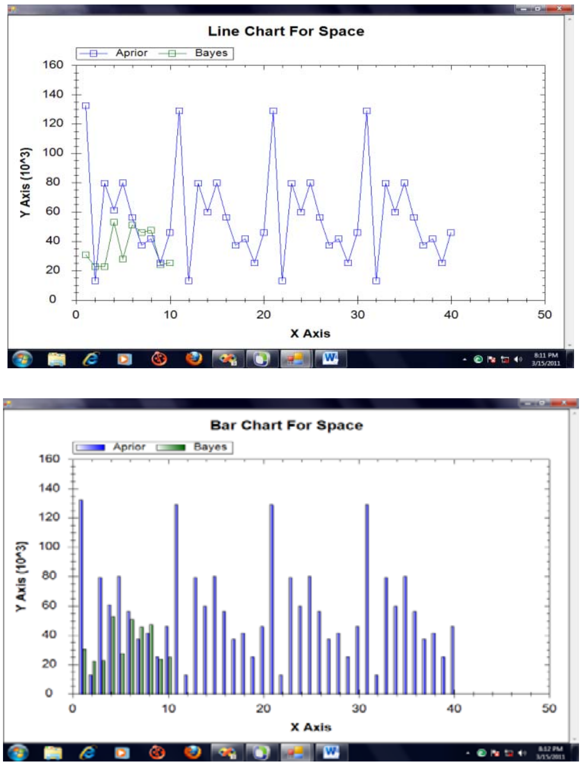


Figure 4.2

5. CONCLUSIONS AND FUTURE WORK

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 adaptively of the proposed framework for concept/event detection.

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. ACMInt. Conf. Multimedia*, Juan les Pins, France, 2002, pp. 347–350. [11] R. Leonardi, P. Migliorati, and M. Prandini, “Semantic indexing ofsoccer 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.