VIDEO SEMANTIC CONCEPT DISCOVERY USING MULTIMODAL-BASED ASSOCIATION CLASSIFICATION

Lin Lin[1], *Guy Ravitz*[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
l.lin2@umiami.edu, {ravitz,shyu}@miami.edu, chens@cs.fiu.edu

ABSTRACT

Digital audio and video have recently taken a center stage in the communication world, which highlights the importance of digital media information management and indexing. It is of great interest for the multimedia research community to find methods and solutions that could help bridge the semantic gap that exists between the low-level features extracted from the audio or video data and the actual semantics of the data. In this paper, we propose a novel framework that works towards reducing this semantic gap. The proposed framework uses the apriori algorithm and association rule mining to find frequent itemsets in the feature data set and generate classification rules to classify video shots to different concepts (semantics). We also introduce a novel pre-filtering architecture which reduces the high positive to negative instances ratio in the classifier training step. This helps reduce the amount of misclassification errors. Our proposed framework shows promising results in classifying multiple concepts.

1. INTRODUCTION

In recent years, digital audio and video have established its dominance in the communication world. With the advancements in the area of communications, digital media became an integrated part of our everyday life. Digital media fused itself into the broadcasting world in the form of technologies such as digital cable and satellite television, digital video recorders, and the highly recently demanded HDTV (High Definition Television) broadcasting. In the area of telephony, audio and video streaming ability became the standard requirements for any cellular device and network by today's average user. In the Internet, digital media has realized in the form of audio and video streaming, which plays a major role in many websites online today. This impressive dominance of digital media gave birth to the need of technology that can accommodate a convenient and efficient viewing and browsing of digital audio and video. The overwhelming popularity of technology such as portable audio/video players, audio and video online libraries, HDTV, and audio video conferencing stresses the dominant and important role digital media plays in our society's life.

The major challenge involved with the design and implementation of such technologies is to grant the system the ability to understand the semantics of the audio and/or video so it can provide a more efficient and convenient viewing and browsing experience to the user. Currently, most audio/video retrieval, browsing, and summarization systems attempt to understand the semantics of the data using different low-level features like color, texture, energy, pitch, etc. This introduces the well-known semantic gap problem. As mentioned in [1], bridging the semantic gap may be the biggest challenge that the multimedia community faces in order to support multimedia data retrieval, and it has recently received much attention. Lin et al. used SVM and hybrid Bayesian networks to construct a semantic model for astrocytoma malignant degrees in MRI images [2]. In their work, low-level features were extracted using medical image processing techniques and used to train the Bayesian network in order to create a conditional probability table. Following that, a stream of low-level features and mid-level semantics from the test data was piped into the BN-SVM-based inference engine, and the benign and malignant states were rendered. Finally, the system declared the image content to be of either benign or malignant tumor. In [3], another solution, namely user feedback, for bridging the semantic gap was explored. In their proposed system, the authors used the so-called user log feedback to record the users' operations in the feedback process. These logged operations were used by the system to learn the users' semantic concepts. Semantic correlation was used to reflect the semantic relevance between images. Their proposed image retrieval system calculated the similarity between the query image and database image using both the similarity between the visual

*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.

features and the learned semantic similarity. In our previous study [4], we have shown that data mining can be effectively used to address the semantic gap issue, where low-level features were first extracted from both audio and video information of digitized soccer broadcasts, followed by temporal analysis to identify temporal patterns and to reduce the data, and finally, a multi-modal based decision tree was constructed to detect goal events in soccer broadcast.

In this paper, a novel framework that uses association rule mining (ARM) technique to discover shot-based semantic concepts from video sequences is proposed. The news broadcast video sequences are used as the testbed, from where different semantic concepts news broadcast video sequences are extracted. We first extract audio and visual features from the broadcast and use association classification to classify the different shots so as to bridge the gap between the low-level features and the high-level concepts. The experimental results demonstrate that our proposed framework can achieve promising precision and recall performance in classifying the concepts such as weather, sports, and commercial.

This paper is organized as follows. In Section 2, we present our proposed framework, and provide detailed discussions on its different components. The experiments and observations are given in Section 3. The paper is concluded in Section 4.

2. THE PROPOSED VIDEO SEMANTIC CONCEPT DISCOVERY FRAMEWORK

In this paper, we propose a video semantic concept discovery framework that utilizes multimodal content analysis and association rule mining (ARM) technique to discover semantic concepts from video data. Our proposed framework consists of the following steps. First, the multimodal audiovisual feature set is extracted from the video data. Next, the training data set is used to generate association classification rules which will later be used to classify the testing data instances. The system assumes that the shot boundary information is known ahead of time, and hence video shot boundary detection is beyond the scope of this paper.

2.1. Semantic Concepts

Using news broadcast video sequences as the testbed, the following three different semantic concepts, namely weather, sports and commercial, are discovered. These concepts are characterized as follows.

- Weather - These are shots containing weather related news or forecast. Usually, there will be a single person in the shot with some graphics displayed in the background. The audio track of such shots would be speech dominated;
- Sports - These shots will display some sports activity such as highlights of a soccer, basketball, or American football game. This concept does not include shots of anchors speaking about a sports news or of athletes being interviewed. The sports shots usually include high amounts of motion in them, and the audio track has more background noise (usually crowd cheer) than the other concepts;
- Commercial - The shots which belong to this group contain a commercial of some type. These shots include all the information that is broadcasted during the breaks of the news broadcast. This group of shots is usually characterized by being very short in durations and having a high audio energy.

To try and capture the aforementioned semantic concepts, we have extracted 25 different features, including 16 audio features, 8 visual features, and 1 feature that represents the length of the shot. We have chosen these concepts as a first step towards a future framework which will have the ability to segment news videos to different stories using the different concepts detected by the proposed shot-based association classification system. To bridge the gap between the features and concepts, association rule mining (ARM) is utilized to learn the concepts and classify the testing data accordingly.

2.2. Association Rule Mining (ARM)

The association rule mining technique is adopted to bridge the semantic gap between low-level multi-modal features and the concepts of interest. As mentioned in [5, 6], association rules are generated by considering different itemsets of different sizes. Given a database *D* of transactions where $T \in D$, a rule such as $A \Rightarrow B$ signifies that when a transaction T has A in it most probably contains B as well. In order to create association rules from a list of itemsets, two measures are considered:

- Support - defined by $P(A \cap B)$ which is the proportion of transactions in *D* that have both *A* and *B*;
- Confidence - defined by $P(B|A)$ which is a measure of the accuracy of the rule.

The process of mining association rules could be summarized to two main steps: (i) finding all the frequent itemsets, and (ii) generating rules using the frequent itemsets which satisfy the minimum support and minimum confidence thresholds. In the case of this paper, the items are the low-level features extracted from the data. Due to the fact that all the feature values are continuous, we had to discretize the data in some way to different intervals. After an empirical study of the feature set, we have decided to use the mean value of each feature as the splitting threshold. This resulted in splitting each feature to two feature-value pairs such that (I_{1i} of f_i) $\in$ [min(f_i), mean(f_i)) and (I_{2i} of f_i) $\in$ [mean(f_i), max(f_i)], where I_{1i} stands for the first partition of the i^{th}

feature. Since the data used in this research originated from different broadcasting networks, we normalized all our features in such way that all values range between 0 and 1 in order to minimize the differences that might occur in the data between the different broadcasts. Due to the normalization, it can be seen that $max(f_i) = 1$. Since the features are split, we now have 50 feature-value pairs from the extracted features, and 2 for the class label which all total to 52 different feature-value pairs. Since this number of feature-value pairs created a large search space, we have decided to use the apriori algorithm which is known to reduces the search space by taking advantage of the a priori property. This property states that if an itemset *Z* is not frequent then adding another item *A* to *Z* will not make *Z* more frequent [6].

2.3. Framework Architecture

Figure 1 shows the architecture of our proposed framework. As can be observed from this figure, the multimodal audiovisual features from the entire data set are extracted. The next step is to split the data set into training and testing sets. We further split the training set to two sets, one containing all the positive instances and the other including the negative instances. For example, in the case of the sports concept, all the shots which belong to the sports concept will form the positive set, and all the rest of the shots will form the negative set. These sets are fed into our proposed pre-filtering mechanism. Our previous studies [7, 4] showed that if the ratio between the positive and negative instances in a data set is too low, it has to be balanced before training and testing the data mining model. Most of the concepts in news video have a small number of shots matching them, which makes the ratio between the positive and negative instances too low. Therefore, a pre-filtering mechanism is proposed to prune down the negative instance set to address the data imbalance issue, which in turns improve the training of the classifier.

Furthermore, it can be seen from Figure 1 that the pre-filtering step (surrounded by a dashed line) begins by generating the positive rule set using the positive training data. As mentioned in Section 2.2, we first generate the frequent itemsets to identify the more significant feature-value pairs which indicate that their corresponding features have higher potentials to identify positive instances. From the positive rule set (i.e., generated by positive instances), all the rules generated by the 6-itemsets (including 5 feature-value pairs and 1 class label) and above are used to prune the negative training instances. We consider rules generated by larger itemsets as better since they include more features and have a lower probability of misclassification. To prune the negative instances, we simply run them through the classifier using the selected positive rules. The classification is simply done by comparing the rules to each instance and when the first match is found between a rule and an instance, it is classified using the class label of the matched rule. If no match to the positive rule is found, the instance is classified as negative by the default rule. The output of this classification is two sets, one with instances that were incorrectly classified as positive (called fuzzy negative set), and the other one including instances that were not matched to any positive rule (called pure negative set). The system discards all the so-called fuzzy negative training instances and uses the so-called pure negative training instances to generate the negative rule set. This is done similarly to the way the positive rules are generated. Finally, the positive and negative rule sets are merged and fed into the classifier along with the testing data set for the classification. Prior to merging the two rule sets together, each set was sorted by 3 ordered criteria, namely the number of feature-value pairs, confidence value, and support value of each rule. Then, the rules are merged together to generate the final classification rule set. Due to the different properties of the data sets representing the semantic concepts such as weather, commercial, and sports, we propose different strategies to merge the rules. The different classifiers are given at the right of Figure 1

- For the weather concept, the number of negative instances is very large compared with the positive instances. Therefore, more negative rules are included after the positive ones to the weather classifier in order to improve its ability to classify the negative instances.
- For the sports concept, the number of positive instances in the training set are not sufficient to represent the high variations of the concept, which caused many missed classifications. In order to solve this issue, more rules were extracted from an additional training set consist-

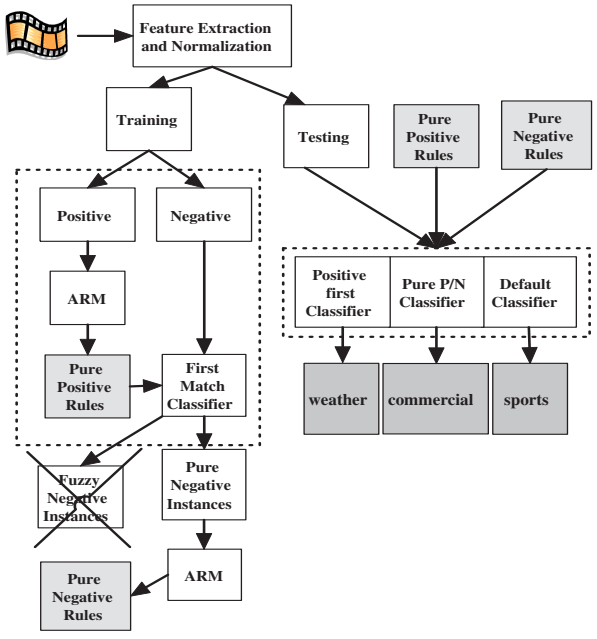

Fig. 1. The architecture of the proposed framework

ing of both positive and negative instances.

- For the commercial concept, many conflict classification situations occur. Therefore, a combination of rules generated from the pure negative instances and pure positive instances.

3. EXPERIMENTS AND RESULTS

We tested our proposed framework using 3 different local news broadcasts that were taped from two different local TV stations. Each video is approximately 30 minutes long. These videos include a total of 1,599 shots. In this paper, we are interested in the *commercial*, *weather*, and *sports* semantic concepts, since we were not able to collect a sufficient amount of instances to build useful training and testing sets for other semantic concepts. Out of the 1,599 shots, 853 are commercial, 44 are weather, and 95 are sports. For each concept, 5-fold cross validation experiments were performed, i.e., 5 different random pairs of testing and training sets were constructed. In each experiment, the instances that matched the tested concept were labeled as the positive instances and the rest as the negative ones. We evaluated each classifier using the precision and recall metrics as shown in Table 1 for positive and negative weather (W+ and W-), positive and negative commercial (C+ and C-), and positive and negative sports (S+ and S-).

Concept	W+	W-	C+	C-	S+	S-
Precision (%)	61	98	72	81	72	97
Recall (%)	81	96	81	72	82	95

Table 1. Average precision and recall for the three classifiers

As can be seen from this table, our proposed framework achieves promising results with above 70% of the precision and recall values for all three semantic concepts, except for the weather concept. The good positive recall results show that our framework has the ability to identify a high percentage of positive instances in each concept, which is more desirable, despite paying the price of a slightly lower precision. The reason of the low precision value for weather is the data imbalance problem, i.e., there are very few positive instances (16) compared to the negative instances (234) in the testing data set. For example, in one of the experiments for the weather concept, though 13 out of 16 positive instances were correctly classified and only 9 out of 234 non-weather instances were mis-classified, the precision value was still greatly affected due to the data imbalance problem. These results demonstrated that the proposed framework was able to somewhat bridge the gap between low-level features and high-level concepts. Furthermore, a different set of significant features can be observed for each concept, e.g., the volume dynamic range and background mean features for *commercial*, the volume standard deviation of the frame-to-frame difference and the mean root-mean-square energy of the subband features for *sports*, and the length of the shot and histogram change features for *weather*.

4. CONCLUSION

In this paper, a multi-modal classification framework based on association classification is proposed. The news broadcast videos are used as the testbed to validate the performance of our proposed framework using the sports, weather, and commercials concepts. In order to aid the proposed framework in performing the task, a novel data pre-filtering mechanism that uses the itemset concept in ARM is developed to reduce the fuzzy negative instances from the training set to improve the training of the classifier. The experimental results demonstrate that our proposed framework is capable of discovering the semantic concepts of interest with good recall and acceptable precision which most of them achieve 70% and above.

5. REFERENCES

[1] J. Fan, H. Luo, and A. K. Elmagarmid, "Concept-oriented indexing of video databases: Toward semantic sensitive retrieval and browsing," *IEEE Transaction on Image Processing*, vol. 13, no. 7, pp. 974–992, July 2004.

[2] C.-Y. Lin, J.-X. Yin, X. Gao, J.-Y. Chen, and P. Qin, "A semantic modeling approach for medical image semantic retrieval using hybrid bayesian newtworks," *Proceedings of International conference on Intelligent Systems Design*, pp. 482–487, October 2006.

[3] J. Han, K. N. Ngan, M. Li, and H. Zhang, "Learning semantic concepts from user feedback log for image retrieval," *Proceedings of International conference on Multimedia and Expo*, vol. 2, pp. 995–998, June 2004.

[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 Magazine, Special Issue on Semantic Retrieval of Multimedia*, vol. 23, no. 2, pp. 38–46, March 2006.

[5] J. Hipp, U. Guntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining - a general survey and comparison," *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 1, pp. 58–64, 2000.

[6] D. T. Larose, *Discovering Knowledge in Data*, Wiley-Interscience, 2005.

[7] S.-C. Chen, M.-L. Shyu, M. Chen, and C. Zhang, "A decision tree-based multimodal data mimimg framework for soccer goal detection," *Proceedings of International conference on Multimedia and Expo*, pp. 265–268, June 2004.