Correlation-based Video Semantic Concept Detection using Multiple Correspondence Analysis

Lin Lin, Guy Ravitz, Mei-Ling Shyu
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
l.lin2@umiami.edu, {ravitz,shyu}@miami.edu

Shu-Ching Chen
School of Computing and
Information Sciences
Florida International University
Miami, FL 33199, USA
chens@cs.fiu.edu

Abstract

Semantic concept detection has emerged as an intriguing topic in multimedia research recently. The ability to interpret high-level semantics from low-level features has been the long desired goal of many researchers. In this paper, we propose a novel framework that utilizes the ability of multiple correspondence analysis (MCA) to explore the correlation between different items (feature-value pairs) and classes (concepts) to bridge the gap between the extracted low-level features and high-level semantic concepts. Using the concepts and benchmark data identified and provided by the TRECVID project, we have shown that our proposed framework demonstrates promising results and performs better than the Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayesian (NB) classifiers that are commonly applied to the TRECVID datasets.

1 Introduction

Recently, one area that has begun to attract much deserved attention in the research community is the area of *concept (semantics) detection*. The main challenge is how to determine the semantic meaning of a picture or a video shot from extracted information such as low-level visual features (color, texture, etc.). This is referred to as the semantic gap in the research community, and bridging this gap is still considered a great challenge [2, 7, 11].

In [7], Mylonas et al. suggested to utilize mid-level information to narrow the gap between low-level features and high-level concepts. In their work, hierarchical clustering was used to construct a region thesaurus from a small training set. The constructed thesaurus contained all the region types encountered in the training set. These region types were considered as the mid-level information which incorporated both low-level and high-level information. The approach in [2] used several local and global classifiers to compensate the fact that the correlation between low-level features and high-level concepts is too weak to be recovered by a single classifier. In their approach, 9 color momentums, 24 Gabor wavelets components, and 2 spatial coordinates were extracted for 260 overlapped patches of 32×32 pixels (in 352×240 images). Topologic context was used to compute the image level concept confidence based on the confidence of all the different patches. In our previous work [5], we proposed a framework that discovered shot-based semantic concepts from news TV broadcasts using *association rule mining (ARM)*. Pure positive and pure negative association rules were generated for each concept and a separate classifier with a different classification rule ranking strategy was developed for each concept (namely, weather, commercial, and sports).

Another problem that has been identified in the existing work [5, 8] is the poor classification performance with an imbalanced data set. In the case of many investigated concepts, the number of positive data instances available in the training data set is very small, which makes the task of training a model to detect these concepts very difficult. In [11], for example, the precision performance of detecting 101 different concepts is provided. It can be observed that for many of the concepts that suffer from the data imbalance problem, much smaller precision values were recorded. Hence, the data imbalance problem in the context of concept detection is very significant and shall be addressed when designing a concept detection framework. In [6], we have introduced the utilization of *Multiple Correspondence Analysis (MCA)* as a feature selection procedure. We compared its performance to several existing feature selection methods and obtained very promising results. The proposed framework was able to help the classifiers detect more positive data instances in the testing data set without misclassify-

978-0-7695-3454-1/08 $25.00 © 2008 IEEE
DOI 10.1109/ISM.2008.111

ing too many negative data instances by identifying the best feature subset for each of the investigated concepts. In addition, the proposed framework was able to reduce the feature space by 50%, which is considered significant. The encouraging performance of MCA in learning the correlation between attributes (features) and the different investigated classes helped us further realize the great impact it can have on the process of concept (semantics) detection and its ability to help narrow the semantic gap and handle the imbalanced data problem better.

In this paper, we propose a novel framework which utilizes MCA to evaluate each of the extracted low-level features and identify the items (i.e., feature-value pairs) as the rules that better represent each one of the investigated concepts. Unlike in our previous work [6] where we used MCA to perform feature selection, in this current work, we utilize MCA to generate rules for classification regardless of what the feature set is. Note that unlike the approach in [9] that used Independent Component Analysis (ICA) to automatically extract features that are closely related to the human perceptual processing, MCA is applied to the indicator matrix in order to learn the correlations between the items and the class labels (to be introduced in details in Section 2.2). In [13], the authors' main desire was to model the relationship between the investigated concepts, which was achieved by using the mutual information. In our proposed framework, we treat the concept and non-concept to the opposite direction and focus on modeling the relationship between the concepts and the features.

To evaluate our proposed framework, we use the concepts and data from TRECVID 2007 [1], and make performance comparison between our proposed framework and the well-known *decision tree* classifier, *support vector machine* classifier, and *Naive Bayesian* classifier. Overall, our proposed framework outperforms all of the three classifiers. Furthermore, it is important to mention that the proposed framework significantly outperforms the support vector machine classifier, which is one of the most commonly used classifiers in the research community for performance evaluation using TRECVID datasets.

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

2 The Proposed Video Semantic Concept Detection Framework

In this paper, we propose a novel video semantic concept detection framework that utilizes MCA to detect semantic concepts from video shots taken from different shows and movies. Our proposed framework consists of the following steps. First, a set of 28 low-level shot-based audio-visual features [3, 5, 6] is extracted from the video data and normalized. Next, we discretize the data in order to be able to properly use MCA. We proceed by evaluating the different items generated by the discretization stage and select the best ones to be used for classification. Finally, we perform the classification using the rules generated by the selected items. 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. For our experiments, we have used the shot boundary ground truth provided to the participants of the TRECVID project.

2.1 Semantic Concepts

To evaluate the proposed framework, we used the news broadcast and movies provided by TRECVID [1], which provides hours of audio-visual test data that can be used by different research groups. According to [1], 54 leading research groups from around the world participated in the TRECVID project in 2007. According to the same source, 43 out of the 54 groups participated in the high level feature extraction (concept detection) task. TRECVID data is considered benchmark data among the information retrieval research community, which provides a great opportunity for different research groups to demonstrate the efficiency of their proposed frameworks. Using this data as our testbed, we have chosen the following concepts, namely vegetation, sky, waterscape, office, road, building, crowd, urban, outdoor, and face. These concepts were taken from the list of concepts provided for the TRECVID 2007 high level feature extraction task [1].

The reasons for selecting these concepts included the facts that (i) there were sufficient amounts of data instances to build useful training and testing data sets for these concepts, and (ii) these concepts represented both balanced and imbalanced data sets which would allow us to demonstrate the robustness of our framework. In addition to the imbalanced data set and the semantic gap challenges, it is important to mention that a shot in the TRECVID data is labeled as containing a specific concept even if only one frame of this shot includes the investigated concept. This furthermore complicates the task of shot-based event detection.

2.2 Multiple Correspondence Analysis (MCA)

Correspondence Analysis (CA) is an exploratory/descriptive data analytic technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. Multiple correspondence analysis (MCA) is an extension of the standard correspondence analysis to more than two variables [10]. MCA is used to analyze a set of observations described by a set of nominal (categorical) variables, and each nominal variable comprises several levels. It codes data by creating a binary column for each level with the constraint that one and only one of the columns gets the value 1 for each nominal variable. Usually, the inner product of such a matrix, called the Burt matrix, is analyzed. Note that the matrix is symmetrical, and that the sum of the diagonal elements in each partition representing the cross-tabulation of a variable against itself must be the same. MCA can also accommodate quantitative variables by recoding them as different bins. Motivated by the functionality of MCA, we explore the utilization of MCA to analyze the data instances described by a set of low-level features to capture the correspondence between items (feature-value pairs) and classes (concepts).

Assume that there are I data instances in a multimedia database, and these data instances are characterized by a set of low-level features. After discretization (i.e., converting the numerical features into nominal ones), there are K nominal features (including classes), each feature has J_k items (feature-value pairs), and the total number of items (i.e., the sum of all J_k) is equal to J. The $I \times J$ indicator matrix is denoted by X, and the $J \times J$ Burt matrix is denoted by $Y = X^T X$. Then, let the grand total of the Burt matrix be N and the probability matrix be $Z = Y/N$. The vector of the column totals of Z is a $1 \times J$ mass matrix M, and $D = diag(M)$. MCA will provide the principle components from the following singular value decomposition (SVD):

$$D^{-\frac{1}{2}}(Z - MM^T)(D^T)^{-\frac{1}{2}} = P\Delta Q^T, \qquad (1)$$

where Δ is the diagonal matrix of the singular values, and $\Lambda = \Delta^2$ is the matrix of the eigenvalues. The columns of P are the left singular vectors (gene coefficient vectors), and the rows of Q^T are the right singular vectors (expression level vectors) in SVD.

Now we can project the multimedia data into a new space by using those components (the first and second principle components in the 2-d space). The similarity of every item and every class shows the correlation between them. Such similarity could be the inner product of each item and class, i.e., the cosine of the angle between each item and class. The smaller the angle is, the more correlated the item and class are.

2.3 Framework Architecture

Our proposed framework is composed of four main stages, namely feature extraction and normalization, data discretization, MCA based rule generation, and classification.

First, we extract our audio-visual feature set based on the shot boundaries of each video. 15 of the audio features and 5 of the visual features that we have used were introduced in [3]. In [5, 6], we have added 1 additional audio feature, 6 additional visual features, and the shot length feature. The audio feature we have added is the *average zero crossing rate (ZCR)*. This feature was introduced to attempt and detect shots which have audio characteristics that resemble pure speech. This feature can help the framework detect concepts such as face and office.

The additional 6 visual features can be divided into two groups, one attempting to extract more color information from the video, and the other attempting to estimate the motion intensity of the video shot. To capture additional color information, we have calculated the dominant red, green, and blue (RGB) values. This group was introduced to aid the detection of concepts such as vegetation, sky, building, and outdoor. The motion estimation was measured based on intra-frame pixel difference in 9 different regions of the video frame. This group of features was introduced in order to help the detection of concepts such as crowd and face. Once all the audio and visual features are extracted, we normalize each feature set for each video to reduce the effects caused by the fact that different videos were broadcasted differently. For example, two shows could have been broadcasted at different volumes, in which case the use of features such as average audio energy could bias the classifier towards the louder videos for some of the concepts.

The feature extraction process mentioned above generates a set of continuous numerical features. Due to the fact that MCA requires the input data to be nominal, all the extracted features are discretized. The method of discretization we used is described in [4], using the information gain for the disparity measure. We discretized the training data set first, and then used the same partitions for discretizing the testing data set. We call the partitions that the discretization process created *items*. In other words, after the discretization process, each feature would have several possible items, and each data instance (shot) in the training data set would have one item per each feature.

Since the ultimate goal of this framework is to use the selected item-class pairs for classification, we need to identify, for each concept, the items whose existence best signifies the existence of the investigated concept in a shot. In other words, we need to find the items that have the highest correlation with each concept. As observed in [6], the correlation between an item and a class appeared to be quantified by the measured angle between the projection of the item and the respective class. Therefore, in the next stage, the angle between each item and each class is calculated by applying MCA to the discretized training data set. For example, for the concept face (labeled 19 in the TRECVID 2007 data), one feature, named *center to corner ratio*, is discretized into 3 partitions labeled 10241, 10242, and 10243, respectively. The projection of that feature and its corresponding three items is shown in Figure 1. The absolute values of the angles between each item and the face class are 126.59, 24.12, and 122.34 degrees, respectively. In this example, the second item (10242) appears to be a better representation for the positive/face class (19), and the rest (10241 and 10243) could be good representations for the negative/non-face class (0).

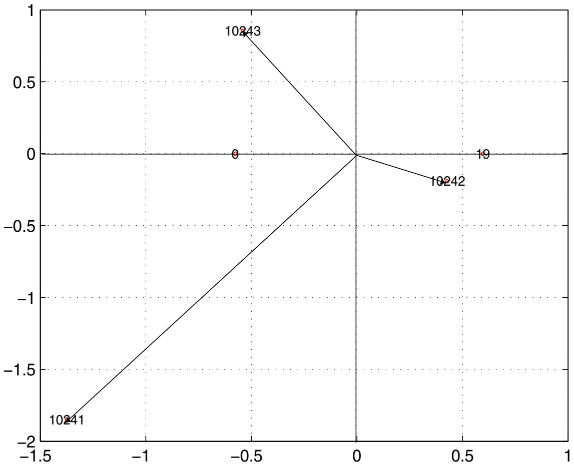

Figure 1. The projection on the first two dimensions in MCA

The next challenge we faced was to realize the proper threshold value to be used in order to decide whether an item is 'close' enough to the class to justify the generation of a classification rule. In order to select the proper angle threshold value, the angles generated by MCA for each concept are sorted. The distribution of the angles generated for one of the concepts is displayed in Figure 2. In this figure, the x-axis is the number of the sorted angles, and the y-axis is the corresponding degree values. To produce the proper threshold value, we use the first big gap from the distribution of the sorted angles before 90-degree as the lower boundary and 90-degree as the upper boundary, and calculated the average value between these two boundary values. Based on this threshold value, the items whose corresponding angle values are smaller than the threshold are deemed to have a high correlation with the class and therefore are used as rules for classification.

Finally, the selected item-class pairs are used as the one-item rule for classification as follows. Each testing data instance is checked to see if it includes the selected items. Per each item in the data instance that matches a rule, the respective data instance is classified using the class label of that rule. After no more items are found for the specific data instance, the sum of the numbers of the positive and negative classifications for that data instance is calculated, and the majority class is assigned. For example, if one data instance matched 7 positive rules and 3 negative rules, it will be labeled as positive by the framework. We then move to the next data instance until all the data instances in the testing data set are processed. This procedure is repeated for each concept, each time using the proper item-class pairs as the classification rules.

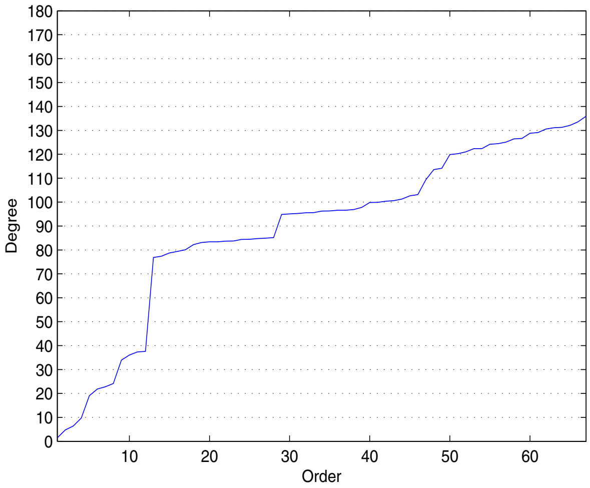

Figure 2. The distribution of the angles between items and one concept calculated by MCA using the testing data set

3 Experiments and Results

Our proposed framework is evaluated using the videos taken from the video collection made available to the TRECVID 2007 high level feature extraction task participants. To train the model, the whole database was sampled to obtain a data set which includes 4,841 shots for the experiments. 28 low-level features were extracted from each video and there was 1 class for each concept. In Table 3, we present the numbers of positive and negative data instances for each concept that exist in our entire data set. The last column of this table shows the ratio between the positive and negative data instances in our data. This illustrates how balanced or imbalanced each of the data sets that we have used for each of our experiments is.

We evaluated our system using the precision, recall, and

Concepts	Positive	Negative	P/N Ratio
Vegetation	534	4307	12.40%
Sky	606	4235	14.31%
Waterscape	685	4156	16.48%
Office	913	3928	23.24%
Road	962	3879	24.80%
Building	1044	3797	27.50%
Crowd	1047	3794	27.60%
Urban	1085	3756	28.89%
Outdoor	1631	3210	50.81%
Face	2375	2466	96.31%

Table 1. The numbers of positive and negative data instances, and the ratio (P/N Ratio) for each concept

F1-score metrics under the 3-fold cross validation approach, i.e., three different random sets of training and testing data sets were constructed for each concept. In each case, the training data set includes two third of all the data instances, while the testing data set includes the remaining one third of the data instances. To show the efficiency of our proposed framework, we compared its performance to the performances of the Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayesian (NB) classifiers available in WEKA [12] using the same evaluation metrics and the same data sets.

The average precision (Pre), recall (Rec), and F1-score (F1) values obtained for all the frameworks over all three folds during our different experiments are presented in Table 2, where columns 2 to 4 provide the performance of Weka's DT, SVM, and NB, respectively, and the last column provides the performance of our MCA based proposed framework.

As can be seen from Table 2, our proposed MCA based framework achieves promising results compared to the decision tree, support vector machine, and naive bayesian classifiers. Furthermore, it can be seen that the proposed framework demonstrates significant improvement over the other classifiers in the cases of the imbalanced data sets. For example, when comparing the F1 measure obtained by using the proposed MCA based method to the highest F1 measure recorded for the other 3 classifiers in each of the experiments, we can see that there is an average improvement of 11 percents in the case of Vegetation, Sky, and Waterscape concepts, and an average improvement of 12.2 percents in the case of Office, Road, Building, Crowd, and Urban concepts. It can be noticed from Table 1 that the positive to negative data instance ratios of the first 3 concepts are between 10 and 20 percents, and the ratios of the next 5 concepts are between 20 and 30 percents. This fact is extremely encouraging as it shows that the proposed framework handles

Vegetation	**DT**	**SVM**	**NB**	**MCA**
Pre	0.41	0.00	0.34	0.26
Rec	0.09	0.00	0.14	0.82
F1	0.15	0.00	0.20	0.38
Sky	**DT**	**SVM**	**NB**	**MCA**
Pre	0.51	0.00	0.34	0.43
Rec	0.19	0.00	0.49	0.52
F1	0.28	0.00	0.40	0.47
Waterscape	**DT**	**SVM**	**NB**	**MCA**
Pre	0.62	0.60	0.41	0.41
Rec	0.42	0.23	0.50	0.99
F1	0.50	0.33	0.45	0.58
Office	**DT**	**SVM**	**NB**	**MCA**
Pre	0.70	0.69	0.49	0.59
Rec	0.57	0.51	0.67	0.97
F1	0.63	0.58	0.57	0.73
Road	**DT**	**SVM**	**NB**	**MCA**
Pre	0.65	0.60	0.45	0.50
Rec	0.46	0.36	0.49	0.96
F1	0.54	0.45	0.47	0.66
Building	**DT**	**SVM**	**NB**	**MCA**
Pre	0.60	0.55	0.51	0.51
Rec	0.40	0.39	0.40	0.93
F1	0.48	0.45	0.44	0.66
Crowd	**DT**	**SVM**	**NB**	**MCA**
Pre	0.56	0.79	0.46	0.39
Rec	0.28	0.08	0.46	0.77
F1	0.37	0.15	0.46	0.52
Urban	**DT**	**SVM**	**NB**	**MCA**
Pre	0.66	0.58	0.50	0.49
Rec	0.39	0.37	0.47	0.92
F1	0.49	0.45	0.49	0.64
Outdoor	**DT**	**SVM**	**NB**	**MCA**
Pre	0.59	0.61	0.53	0.50
Rec	0.51	0.41	0.52	0.59
F1	0.54	0.49	0.53	0.54
Face	**DT**	**SVM**	**NB**	**MCA**
Pre	0.65	0.68	0.65	0.57
Rec	0.64	0.64	0.65	0.81
F1	0.64	0.66	0.65	0.67

Table 2. Performance evaluation for the concepts

the imbalanced data sets better than the other classifiers. As previously mentioned, the imbalanced data set problem is considered one of the major challenges in detecting high level concepts using the raw low level content of the videos. We also observe that SVM always yields zero precision and recall values in the extremely imbalanced concepts, namely Vegetation and Sky. This can be explained by the fact that when the class distribution is too skewed, SVM will generate a trivial model by predicting everything to the majority class, i.e., the negative class.

Finally, it is observed that the F1-scores obtained for the proposed framework are either equal to or higher than the ones obtained for the other classifiers. Furthermore, the recall values of the proposed framework are always higher than the ones achieved by the other classifiers. Our observation demonstrates that the proposed concept detection framework can effectively handle the data imbalance issue and bridge the semantic gap with promising results.

4 Conclusion

In this paper, a multimodal correlation-based video semantic concept detection framework using MCA is proposed. Data taken from the TRECVID 2007 video corpus is used to validate the detection performance of our proposed framework by detecting the Vegetation, Sky, Waterscape, Office, Road, Building, Crowd, Urban, Outdoor, and Face concepts. We utilize the functionality of MCA to measure the correlation between different items and classes to infer the high-level concepts from the extracted low-level audio-visual features. The experimental results demonstrate that our proposed framework performs well in detecting the selected concepts from the TRECVID benchmark data. The results show significant performance increases in the case of imbalanced data sets with an average performance increase of 11.25% in the F1 scores of the proposed framework in comparison to the highest F1 score of all 3 classifiers per each concept. Furthermore, we are able to show an increased overall recall and F1 performance over the well-known decision tree, support vector machine, and naive bayesian classifiers.

5 Acknowledgment

For Mei-Ling Shyu, this research was carried out in part under the auspices of the Cooperative Institute for Marine and Atmospheric Studies (CIMAS), a Joint Institute of the University of Miami and the National Oceanic and Atmospheric Administration, cooperative agreement #NA17RJ1226. “The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency.” For Shu-Ching Chen, this research was supported in part by NSF HRD-0317692, NSF HRD-0833093, and Florida Hurricane Alliance Research Program sponsored by the National Oceanic and Atmospheric Administration.

References

[1] *Guidelines for the TRECVID 2007 Evaluation.* http://www-nlpir.nist.gov/projects/tv2007/tv2007.html.

[2] S. Ayache, G. Quenot, and S. Satoh. Context-based conceptual image indexing. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, 2:II–II, May 14-19 2006.

[3] 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. *International Journal of Computer Applications in Technology, Special Issue on Data Mining Applications*, 27(4):312–323, 2006.

[4] U. M. Fayyad and K. B. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8:87–102, 1992.

[5] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Video semantic concept discovery using multimodal-based association classification. *IEEE International Conference on Multimedia and Expo*, pages 859–862, July 2-5 2007.

[6] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Effective feature space reduction with imbalanced data for semantic concept detection. *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, pages 262–269, June 11-13 2008.

[7] P. Mylonas, E. Spyrou, and Y. Avrithis. High-level concept detection based on mid-level semantic information and contextual adaptation. *Second International Workshop on Semantic Media Adaptation and Personalization*, pages 193–198, December 17-18 2007.

[8] A. Natsev, M. R. Naphade, and J. Tesic. Learning the semantics of multimedia queries and concepts from a small number of examples. *MULTIMEDIA ’05: Proceedings of the 13th ANNUAL ACM conference on Multimedia*, pages 598–607, 2005.

[9] J.-Y. Pan and F. C. Videocube: a novel tool for video mining and classification. *Proceedings of the Fifth International Conference on Asian Digital Libraries*, December 11-14 2002.

[10] N. J. E. Salkind. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage Publications, Inc, 2007.

[11] C. G. M. Sonek, M. Worring, J. C. Van Gemert, J.-M. Geuseborek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA ’06: Proceedings of the 14th ANNUAL ACM conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM Press.

[12] WEKA. *http://www.cs.waikato.ac.nz/ml/weka/*.

[13] Z.-J. Zha, T. Mei, Z. Wang, and X.-S. Hua. Building a comprehensive ontology to refine video concept detection. *Proceedings of ACM SIGMM International Conference Workshop on Multimedia Information Retrieval(MIR)*, pages 227–236, September 2007.