Florida International University and University of Miami TRECVID 2008 - High Level Feature Extraction

Guy Ravitz, Lin Lin, 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

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

Abstract

This paper describes the FIU-UM group TRECVID 2008 high level feature extraction task submission. We have used a correlation based video semantic concept detection system for this task submission. This system first extracts shot based low-level audiovisual features from the raw data source (audio and video files). The resulting numerical feature set is then discretized. Multiple correspondence analysis (MCA) is then used to explore the correlation between items, which are the feature-value pairs generated by the discretization process, and the different concepts. This process generates both positive and negative rules. During the classification process each instance (shot) is tested against each rule. The score for each instance determines the final classification. We have conducted two runs using two different predetermined values as the score threshold for classification:

- A_FIU-UM-FE1_1*: train on partial TRECVID2008 development data (all TRECVID2007 development data + partial TRECVID2007 test data) using* -2 *as the instance score for final classification*
- A_FIU-UM-FE2_2*: train on partial TRECVID2008 development data (all TRECVID2007 development data + partial TRECVID2007 test data) using 0 as the instance score for final classification (simple majority)*

We observed a slight improvement in the A_FIU-UM-FE2_2 *run over the* A_FIU-UM-FE1_1 *run. Initially it appeared from the training data that using a score of* -2 *could potentially provide a better performance, however; in order to test a true majority voting concept we have conducted the second run (*A_FIU-UM-FE2_2*) using 0 as our threshold. Based on the submitted results and our results produced in some of our previous work [6] we believe that the MCA process has the capability to learn the correlation between low-level features such as color, volume, texture etc. and high level features (concepts) and by that help narrow the semantic gap. One of the biggest challenges of this year's high level feature extraction task was the fact that the target high-level feature list has been changed. This year we have used the same low-level features that we used in 2007. We believe that this low level feature set might have not been the best candidate to represent new high level feature list. Therefore, We believe that extracting additional audio-visual features which are a little more relevant to the new concept list would have improved our observed performance. Finally we observed that the problem of imbalanced data is still a major challenge that our system is having difficulties to address. In this paper we will provide more details regarding our system, discuss our observations, and provide some thoughts regarding the future to which this system is heading.*

1 Introduction

The high level feature extraction task in the TRECVID project [10] addresses one of the greatest challenges in the area of multimedia content-based analysis. The goal is to automatically detect high level features (concepts) such as hand, mountain, harbor, cityscape and more, using raw data extracted from video, audio, and text. Many researchers refer to this problem as the semantic gap, and still consider bridging this gap as a great and of high priority challenge [1, 7, 11]

In [1] Ayache et al. claimed that the correlation between low-level features and high-level concepts is too week to be recovered by a single classifier. As a solution the authors proposed a system where several features were extracted for 260 overlapped patches of 32×32 pixels, and an image level concept confidence was computed using topologic context based on the confidence of all the different patches.

In [7], Mylonas et al. suggested to narrow this so called semantic gap by utilizing mid level features. A region thesaurus was constructed by applying hierarchical clustering to a small training set. This thesaurus contained all the region types that have been encountered in the training data set, and was considered as the mid-level information. A model vector was formed for each image based on the region thesaurus. This model vector semantically described the visual content of the image. Finally, a neural network-based classifier was trained for each concept. Given the model vector as an input, the classifier provided the confidence of the existence of the concept within the investigated image.

In [12], Zha et al. discovered an effective way to facilitate semantic video concept detection by using ontology. The presented ontology was built comprehensively in three steps, i.e., concept selection, property selection, and relation selection. First, the concepts were organized into six categories, namely programs, locations, people, objects, activities, and graphics. Second, the property was described as the weights of different modalities obtained by either data-driven approaches or manually. Last, the semantic linkages among concepts were captured by using pairwise correlation (e.g., the relation between the concepts "road" and "car") and hierarchical relation (e.g., the hierarchical structure from concepts "outdoor" to "building" and to "office"). The authors adopted a propagation strategy for the pairwise relation and a Bayesian hierarchical combination strategy for the hierarchical relation.

In our previous work we have been able to demonstrate the effectiveness of MCA in learning the correlation between low-level features and high-level concepts. In [6] we have proposed an effective feature space reduction algorithm which improved the accuracy of semantic concept detection. The proposed system utilized MCA to discover the correlation between low-level features and high level-concepts. We used data and concepts from TRECVID2007 to demonstrate the effectiveness of the proposed system. Finally we compared the performance of our algorithm with some well known feature selection algorithms and were able to demonstrate the superiority of our proposed algorithm especially in the case of imbalanced data sets. In [5] We used MCA to learn the correlation between items (feature-value pairs) and classes in order to generate classification rules for different high level concepts. TRECVID2007 data was used to evaluate this proposed system as well, and the performance was compared to some well known classification algorithms. Our results were promising and demonstrated superior performance in some cases of imbalanced data sets.

We used our work described in [5], with several modifications, to produce the results submitted to the TRECVID2008 high level feature extraction task. We will describe the framework and the different modifications later on in this paper.

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

The multimedia content based concept detection system we used to produce the results submitted to the TRECVID2008 high level feature extraction task is based on our previous work presented in [5]. This framework consists of 4 stages namely, feature extraction and normalization, discretization, MCA based rules generation, and classification. The MCA based stage we used here is slightly different than the one we described in [5]. More details regarding these differences will be given in the following sections. Before we jump into a detailed discussion of the framework we used, we will provide a short discussion regarding MCA and introduce some important concepts related to it.

2.1 Multiple Correspondence Analysis (MCA)

Correspondence Analysis (CA) refers to a technique which is designed to analyze simple two-way and multi-way tables which contain 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 [9]. MCA analyzes set of observations described by a set of nominal (categorical) variables. Each of these variables comprises several levels which are coded by MCA. Each level is coded to a binary column. For each nominal variable, only one of the columns (levels) can get a value of 1. MCA analyzes the product of such coded matrix, which results in the generation of the Burt matrix. The functionality of MCA motivated us to explore its utilization to analyze labeled instances described by a set of low-level features to capture the correspondence between items (feature-value pairs) and the investigated concepts (classes).

Assuming that we have a nominal feature set of K features (including the classes) that characterizes I data instances in a multimedia database. 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. We can denote the $I \times J$ indicator matrix by

X and the $J \times J$ Burt matrix by $Y = X^T X$ We 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 then produce the principle components from the following singular value decomposition (SVD) taken from [5]:

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

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

This allows us to project our multimedia data set into a new space by using the first and second principle components in the 2-d space. The similarity of every items and every concept (class) can be represented by the inner product of each item and class which is calculated by the cosine of the angle between each item and class. The relationship could be described by saying that the smaller the angle is, the more correlated the item and the concept are.

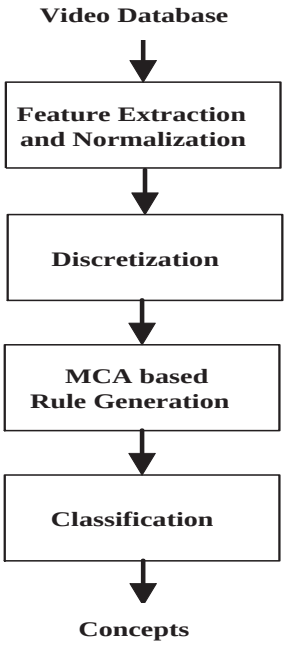

Figure 1. The Proposed Framework.

2.2 Framework Architecture

The concept detection framework we have used is shown in Figure 1. During the first stage 28 shot based low level audiovisual features are extracted from the video corpus. 15 of the audio features and 5 of the video features were introduced in [2]. The rest of the features were added in [4, 6]. The low level audio features were used to extract information such as average energy, dominant frequencies, dynamic range, and zero crossing rate. The visual features were used to extract information such as dominant color, motion estimation and more. The result of the first stage is a shot based numerical feature set. In

order to use MCA the feature set had to be discretized. This was done using the method described in [3], using the information gain for the disparity measure. The training set was discretized first, and then the partitions generated by this process were used to discretize the testing set. In this work we refer to these partitions as *items*. The result of the second stage is that each feature would have several possible items, and each data instance (shot) in the training data set would have one item per each feature.

The goal of the the third stage of our framework is to generate rules for classification. 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 this stage we are looking for the items that have the highest correlation (smallest angel) with the existence or non existence of the concept in the shot. The next example taken from [5] illustrates this idea.

We use the concept face (labeled 19 in the TRECVID 2007 data) here as an example. One of the extracted features named *center to corner ratio*, is discretized into 3 partitions which are labled 10241, 10242, and 10243, respectively. The projection generated by MCA of that feature and its corresponding items is shown in Figure 2 also taken from[5]. The absolute values of the angles between each item and the face concept (high level feature) are 126.59, 24.12, and 122.34 degrees, respectively. As can be seen from Figure 2 the second item (10242) appears to better represent the positive/face concept (19), and the others (10241 and 10243) could be used as good representations for the negative/non-face class (0).

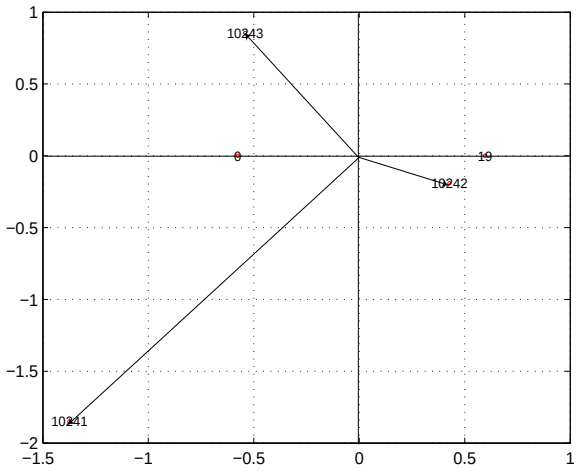

Figure 2. The projection on the first two dimensions in MCA taken from [5]

Next we had to realize the proper angle threshold value to be used in order to decide whether an item is ‘close’ enough to the class in order to justify the generation of a classification rule. In [5] we sorted the angles calculated by MCA and used the average of the first big gap in the angel value before 90-degrees. During our current experiments we have observed that this method does not always provide us with the best classification results (of the training data). For this reason we have decided to sort the angles and iteratively use each value (starting with the highest angle under 90-degrees) to generate rules and classify the training set. Finally for each concept we selected the angel that generated the best precision and recall values, using the training set, as a threshold to be used to generate the final classification rules.

Finally, we used the selected item-class pairs as the one-item rule for classification as follows:

- Each testing instance is checked to see if it includes the selected items.
- Per each item in the instance that matches a rule we give it a score of 1 if it matches a positive rule and -1 if it matches a negative rule.
- Once no more items are found for the investigated testing instance the score for this instance is calculated by summing the scores provided in the second step.
- A class label (concept / high level feature) is given to an instance based on the score based on some threshold value. For example for simple majority voting the threshold is set to 0. Meaning, if an instance matches more positive rules than negative ones, then it is declared positive.

3 Experiments and Results

Our concept detection framework was trained using the majority of the TRECVID2007 development video and some of the TRECVID2007 test videos. The runs we have submitted were produced using part of the TRECVID2008 test data. Our system has been trained and tested as follows. We have used the shot boundary information provided by [8]. Based on this information we have extracted 28 low-level features from the majority of the video corpus. We have populated the resulting feature sets along with the proper ground truth information provided by TRECVID in two separate databases labeled training and testing respectively. The training database included feature and class labeling information for the majority of the TRECVID2007 development data and part of the TRECVID2007 test data. The testing database included a portion of the TRECVID2008 test data.

Due to the fact that some concepts had a very low number of positive instances we had to sample the data in order to properly train our system. For each concept (high level feature) we chose all the positive instances we had from the training database, (e.g 128 positive), and randomly chose equal number of negative instances (128 negative) from the same database. In cases where the number of positive instances was too small, e.g. 50, we duplicated the positive instances to 100, and randonly chose 100 negative instances. Based on some previous studies we have made, we have learned that if the ratio of positive to negative instances is smaller than 10%, MCA would always produce angles around 90 degree (very low correlation). So this sampling method would make sure that MCA produces some small values of angles for both the positive and negative classes. The main problem of this sampling method is that the sampling size is very small and probably does not represent the entire data set well enough.

Next the data is discretized as mentioned above, and passed to the rule generation stage where we use MCA to generate the classification rules. For each concept we first use the sample training set to generate the proper angles using MCA. We then sort these angles and generate rules based on each angle. Next, we use the generated rules by each of the angles to classify the entire training set (complete training database) and calculate precision, recall, and F1- measure values for each such classification. The classification is done based on simple majority voting. We then take the angle that generates the best results using the training data and use the rules created by this angle as the final classification rules. Next, we tested the final classification rule set on the entire training data again this time using different score summation thresholds namely, -2, -1, 0, 1, and 2. We have observed that for each concepts the best results on the entire training data was obtained by using -2, and 0, and therefore we have decided to use -2 as the threshold for the first run (*A_FIU-UM-FE1_1*) and 0 (simple majority voiting) for the second run (*A_FIU-UM-FE1_2*). Finally, for each run we sorted the classified shots by their summed

score from highest to lowest and submitted the top 2000 shots for each of the runs. The TRECVID evaluations for our runs can be seen in Figure 3 and Figure 4 respectively.

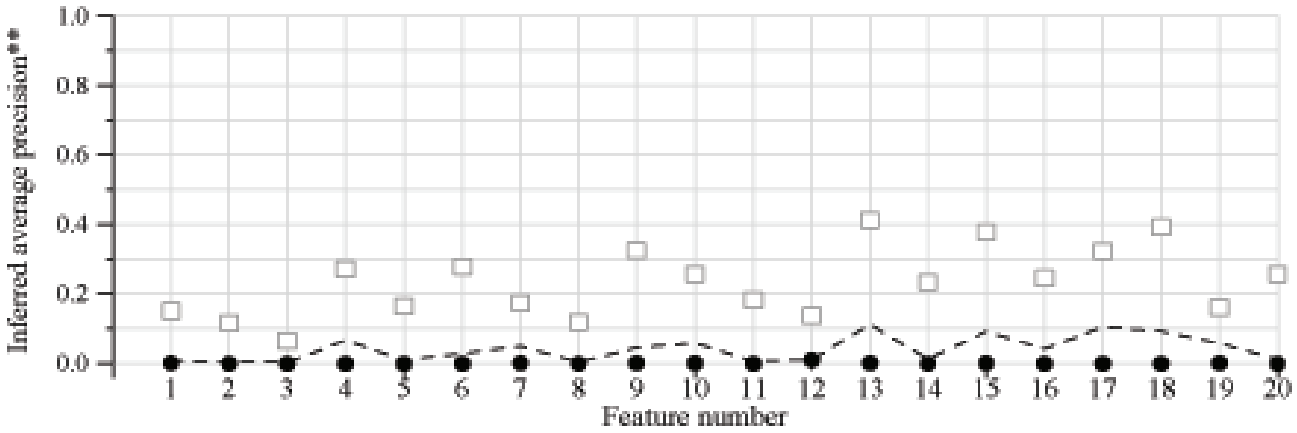

Figure 3. Run score (dot) versus median (- - -) versus best (box) by feature for *A_FIU-UM-FE1_1*

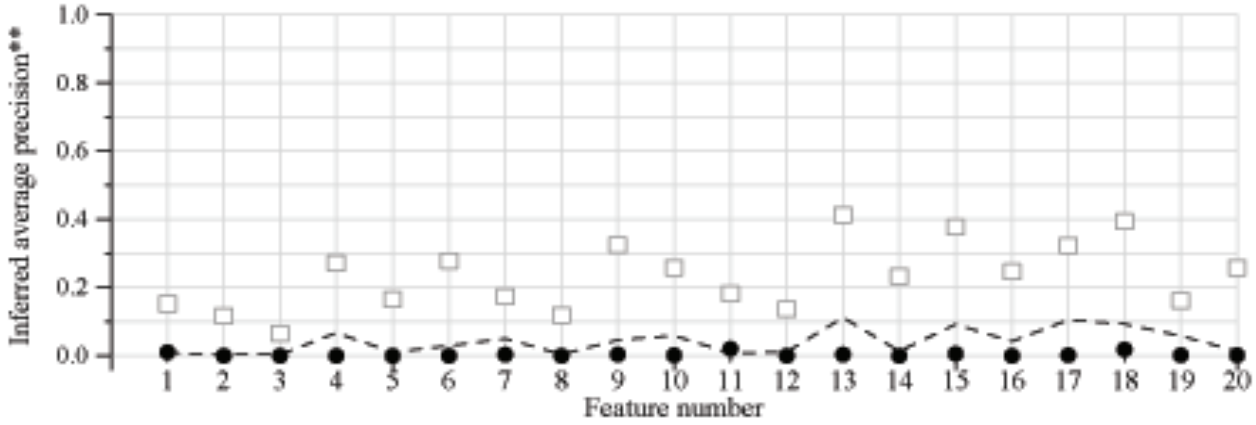

Figure 4. Run score (dot) versus median (- - -) versus best (box) by feature for *A_FIU-UM-FE1_2*

Based on the evaluation information returned to us from TRECVID our second run *A_FIU-UM-FE1_2* performed slightly better than the first one. For all 20 concepts, out of 4670 positive shots *A_FIU-UM-FE1_2* returned 267 positive shots, while *A_FIU-UM-FE1_1* returned 144 positive shots. The mean inferred precision was 0.004 for *A_FIU-UM-FE1_2* and 0.001 for *A_FIU-UM-FE1_1*. finally it can be seen from Figure 3 and Figure 4 that for almost 50% of the concepts we achieved an overall inferred average precision equal to the median and in the rest our system was below the median. In once case (concept 11 of *A_FIU-UM-FE1_2*) we achieved a result slightly higher than the median.

4 Conclusion and Future Work

In this paper we introduced the system we have used to generate the runs we have submitted to the TRECVID2008 high level feature extraction task. The evaluation results of our system are provided and discussed in the previous section. The main challenge this year was the fact that a new concept (high level features) list was created and this made some of the previously used low level features irrelevant. Due to restriction in time and resources we were not able to use all the provided data for development and this resulted in some concepts (high level features) having very little positive examples in our development data set. This made it very difficult to train a reasonable model with the entire data set and forced us to sample the training set. We believe that the sampling was not a good representation of the data and therefore affected the overall performance of our system. Based on good results in our previous studies and some reasonable results obtained for some of the concepts using the training data set we still believe that MCA can be highly useful in the process of high level feature extraction.

As we conclude this year's TRECVID task we have identified a few areas where improvement can be achieved. First, we intend to identify and extract more low-level audiovisual features that will generate a better representation of the new high level feature list. When it comes to low level feature extraction, the ultimate goal would be to find features that could be as general as possible and will be able to represent as many concepts as possible. Second, we plan to improve the feature extraction process so that this stage would consume less time and this way we could finish processing the entire data provided for the task (both development and test sets). Third, we are planing to attempt including other sources, such as YouTube, to our development set to improve the training of the system, especially in the case of those concepts that have a low number of positive examples in the development data. Fourth, we are planing to keep improving our classification algorithms, and finally, we are planing to improve our final ranking mechanism. We believe that all the aforementioned will help improve the efficiency of our concept detection system.

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 and Florida Hurricane Alliance Research Program sponsored by the National Oceanic and Atmospheric Administration. For Guy Ravitz, this research was supported in part by the Graduate Activity Fee Allocation Committee (GAFAC) of the University of Miami.

References

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

[2] 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.

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

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

[5] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Correlation-based video semantic concept detection using multiple correspondence analysis. *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, December 15-17 2008.

[6] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Effective feature space reduction with imbalanced data for semantic concept detection. *Proceedings of the 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. *Proceedings of the Second International Workshop on Semantic Media Adaptation and Personalization*, pages 193–198, December 17-18 2007.

[8] C. Petersohn. Fraunhofer hhi at trecvid 2004: Shot boundary detection system. *TREC Video Retrieval Evaluation Online Proceedings, TRECVID*, 2004.

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

[10] A. F. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In A. Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.

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