Florida International University - University of Miami TRECVID 2014

Miguel Gavidia[3], Tarek Sayed[1], Yilin Yan[1], Quisha Zhu[1], Mei-Ling Shyu[1],
Shu-Ching Chen[2], Hsin-Yu Ha[2], Ming Ma[1], Winnie Chen[4], Tiffany Chen[5]

[1]Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33146, USA

[2]School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA

[3]Department of Computer Science
University of Miami, Coral Gables, FL 33124, USA

[4]School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907, USA

[5]Miami Palmetto Senior High School, Pinecrest, FL 33156, USA

m.gavidia@miami.edu, t.sayed@miami.edu, y.yan4@umiami.edu, q.zhu2@umiami.edu, shyu@miami.edu, chens@cs.fiu.edu, hha001@cs.fiu.edu, m.ma6@umiami.edu,

Abstract

This paper demonstrates the framework and results from the team "Florida International University - University of Miami (FIU-UM)" in TRECVID 2014 Semantic Indexing (SIN) task [Smeaton09]. Two runs were submitted, and the summary of these four runs are as follows:

- 2B_M_A_FIU-UM.14_3: RSPM (Collateral Representative Subspace Projection Modeling) - RSPM based ranking using key frame (KF) features.
- 2B_M_A_FIU-UM.14_4: RSPM - RSPM based ranking using KF features.

In Runs 1 and 2, the same baseline RSPM-based model is applied. We utilized the approach that was used in the 2012 SIN task to see if we could continue to improve the gains. However as a result, the performance was poor and provided no discernible gains.

1. Introduction

In the TRECVID 2014 projects [Over14], the semantic indexing (SIN) task aims to recognize the semantic concept contained within a video shot. The SIN task needs to address several challenges such as data imbalance, scalability, and semantic gap [Chen06, Chen07, Lin07, Lin08]. The automatic annotation of semantic concepts in video shots can be an essential technology for retrieval, categorization, and other video exploitations. The semantic concept retrieval research directions consist of (i) developing robust learning approaches that adjust to the increasing size and the diversity

of the videos, (ii) fusing information from other sources such as audio and text, and (iii) detecting the low-level and mid-level features that have a high discriminability.

The size of the high-level semantic concepts remains the same as that of last year's SIN task (i.e., 60 semantic concepts). For each of the 60 semantic concepts, the participants are allowed to submit a maximum of 2,000 possible shots, and the submission result is rated by using mean inferred average precision (mean xinfAP) [Yilmaz08].

This paper is organized as follows. Section 2 describes our proposed framework and the specific approaches utilized for each run. Section 3 shows the submission results in details. Section 4 summarizes the whole paper and proposes some future directions to pursue.

2. The Proposed Framework

Our proposed framework for the TRECVID 2014 SIN task is shown in Figure 1. The key frame level features (KF) are extracted and normalized. For both Run 1 and Run 2, the RSPM (Collateral Representative Subspace Projection Modeling) model is applied but with different KF feature sets. The xinfAP values are calculated from the models trained on the TRECVID 2012 training data and evaluated on the TRECVID 2012 testing data.

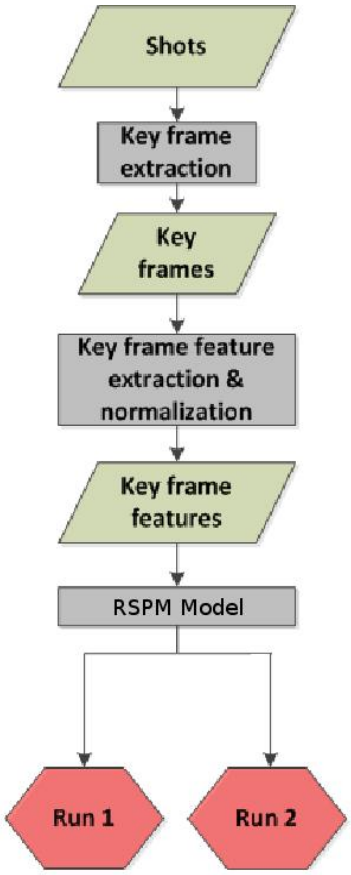

Figure 1. Our proposed framework for semantic indexing

2.1 Data Pre-processing and Feature Extraction

A key frame for each shot is provided to the SIN task participants in both the training and testing videos. Ten kinds of KF features are extracted from each frame in the training and testing data, including the color histogram in the HSV space and the canny edge histogram. Before extracting the features, histogram equalization is employed to regulate the contrast of the frames.

2.2 Collateral Representative Subspace Projection Modeling (C-RSPM)

The C-RSPM model includes the Classification module and Ambiguity_Solver module [Quirino06]. The Classification module is composed of an array of deviation classifiers which are executed collaterally. In other words, each of the classifiers receives and classifies the same testing instance simultaneously. Its basic idea is that each classifier in the Classification module is trained with a different type of known-class data from a training data set, through the employment of the RSPM technique. Thus, training the C-RSPM classifier consists of training each individual classifier to recognize instances of a specific class type. The challenge of using the RSPM technique in multi-class supervised classification is in keeping a high true detection rate with an enough low false alarm rate. Theoretically, a testing instance normal to a certain classifier's training data should either be entirely rejected by all the other classifiers or be classified into no more than one class group. However, the generated result does not always give such an outcome. Hence, in this context, two common issues should be carefully considered and approached if the proposed collateral classification architecture is to function properly:

1) Globally unrecognized instance issue: A testing instance is rejected by all classifiers and cannot be assigned a class label.
2) Classification ambiguity issue: More than one classifier accepts a testing instance as statistically normal to their training data.

With a low programmed false alarm rate, our proposed C-RSPM addresses the first issue by assigning the rejected testing instances as an “Unknown” class label. To approach the second issue, the Ambiguity_Solver module was developed to coordinate and capture classification conflicts. This module defines a class attachment measure A_k called the Attaching_Proportion for each of the k ambiguous classes, i.e., for all classes that during the classification phase have recognized the testing instance as statistically normal to their training data set. In our proposed framework, the class label of the classifier with the lowest A_k value is assigned to the ambiguous testing instance. The basic idea is that the lowest attaching proportion measure should reflect a closer resemblance of the testing instance to the specific class type. In other words, we consider A_k as a measure of the degree of normality of an instance with respect to a class type. The lower the A_k value is, the lower the distance of the testing instance in question to those instances lying close to the center of the spatial distribution of class k is. Therefore, employing the Ambiguity_Solver module enables one single class label for any testing instance.

3. Experimental Results

The overall framework of TRECVID 2014 SIN task contains three stages:

1. Model training: using TRECVID 2012 training videos as the training data.
2. Model evaluation: using TRECVID 2012 testing videos as the testing data to evaluate the framework and tune the parameters of the models.
3. Model testing: using TRECVID 2012 training + TRECVID 2012 testing videos as the TRECVID 2014 training data, and TRECVID 2014 testing videos as the testing data to generate the ranking results for submission.

Figure 2 and Figure 3 present the performance of our semantic indexing results. The x-axis is the concept number; while the y-axis is the inferred average precision value. More clearly, Table 1 shows the inferred mean average precision (MAP) values of the first 10, 100, 1000, and 2000 shots. The inferred true shots and mean xinfAP are shown in Table 2.

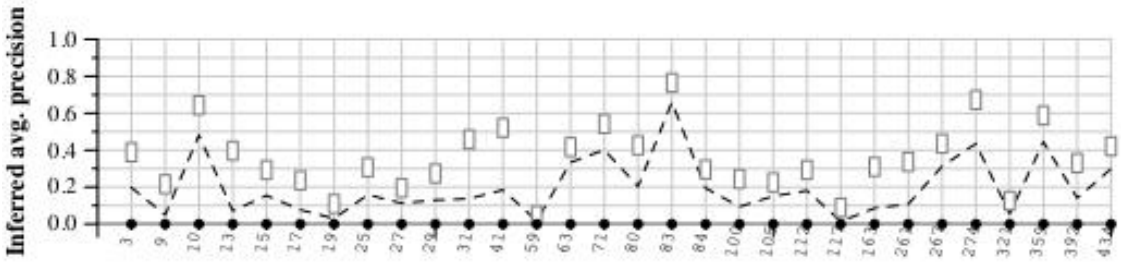

Figure 2. Run scores (dot) versus median (—) versus best (box) for 2B_M_A_FIU-UM.14_3

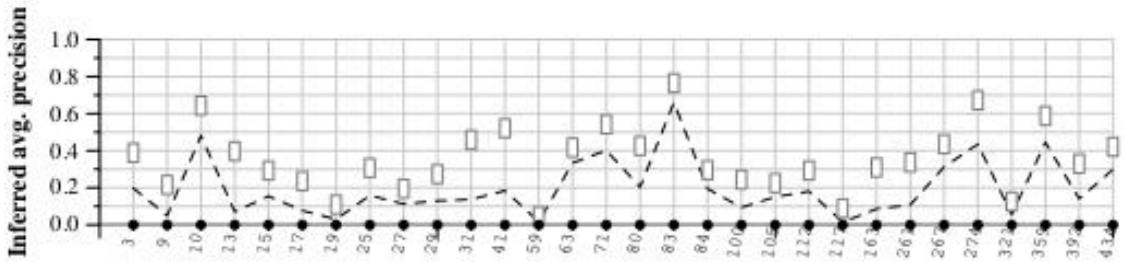

Figure 3. Run scores (dot) versus median (—) versus best (box) for 2B_M_A_FIU-UM.14_4

Table 1: The MAP values at first n shots for all 2 runs

Framework	10	100	1000	2000
2B_M_A_FIU-UM.14_3	0.3%	0.2%	0.2%	0.2%
2B_M_A_FIU-UM.14_4	0.0%	0.4%	0.4%	0.4%

Table 2: Inferred true shots and mean xinfAP

Framework	Inferred true shots	Mean xinfAP
2B_M_A_FIU-UM.14_3	117	0.00
2B_M_A_FIU-UM.14_4	218	0.00

4. Conclusion and Future Work

In this notebook paper, the framework and results of team FIU-UM in the TRECVID 2014 SIN task are summarized. As can be seen from the results, there are still a lot of improvements that need to be conducted. The following important directions will be investigated:

- In our framework, only global features are utilized. Object-level and mid-level features need to be explored.
- The proper re-ranking strategy needs to be explored in depth to further improve the retrieval accuracy.
- The proper filtering strategy needs to be adopted to address the data imbalance issue.

It is also necessary to exchange the ideas and thoughts with other groups to come up with novel approaches to further improve the performance.

References

[Chen06] S.-C. Chen, M.-L. Shyu, C. Zhang, and M. Chen. A Multimedia Data Mining Framework for Soccer Goal Detection based on Decision Tree Logic. International Journal of Computer Applications in Technology, Vol. 27, No. 4, pp. 312-232, 2006.

[Chen07] M. Chen, S.-C. Chen, and M.-L. Shyu. Hierarchical Temporal Association Mining for Video Event Detection in Video Databases. The Second IEEE International Workshop on Multimedia Databases and Data Management (MDDM'07), in conjunction with IEEE International Conference on Data Engineering (ICDE2007), pp. 137-145, Istanbul, Turkey, April 15, 2007.

[Lin07] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Video Semantic Concept Discovery using Multimodal-based Association Classification. In Proceedings of the IEEE International Conference on Multimedia & Expo (ICME), pp. 859-862, Beijing, China, July 2-5, 2007.

[Lin08] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Effective Feature Space Reduction with Imbalanced Data for Semantic Concept Detection. In Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC2008), pp. 262-269, Taichung, Taiwan, June 11-13, 2008.

[Over14] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quéenot. TRECVID 2014 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In Proceedings of TRECVID 2014, NIST, USA.

[Quirino06] T. Quirino, Z. Xie, M.-L. Shyu, S.-C. Chen, and L. W. Chang. Collateral Representative Subspace Projection Modeling for Supervised Classification. In Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), pp. 98-105, November 13-15, Washington D.C., USA.

[Smeaton09] F. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. Springer US, first edition, 2009.

[Yilmaz08] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A Simple and Efficient Sampling Method for Estimating ap and ndcg. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 603–610, New York, NY, USA, 2008.