Abstract

This paper demonstrates the framework and results from the team “Florida International University - University of Miami (FIU-UM)” in TRECVID 2015 Semantic Indexing (SIN) task [1]. Four runs were submitted, and the summary of these four runs is given as follows:

- **2C_M_A_FIU UM.15.1**: MCA late fusion - Multiple Correspondence Analysis (MCA) based ranking using the MCA scores of all ten key frame (KF) features.
- **2C_M_A_FIU UM.15.2**: MCA early fusion - MCA based ranking using the selected five KF features.
- **2C_M_A_FIU UM.15.3**: Run 1 + Time information - MCA late fusion combined with MCA scores from frames other than key frames.
- **2C_M_A_FIU UM.15.4**: MCA early fusion - MCA based ranking using the selected four KF features.

In Run 1, the MCA scores from ten KF features are combined and re-ranked. For Run 3, the result from the aforementioned run (i.e., run 1) and the time information extracted from frames other than key frames are fused. In this way, we wanted to test whether the time information could help improve the results. In Run 2 and Run 4, different feature sets are combined to feed to the same baseline MCA-based model. As a result, from the submission results, Run 2 outperforms the other three runs.

- **Processing type**: Automatic
- **Class**: M - main, single concepts
- **Training type**: A (only the IACC data)
1 Introduction

In the TRECVID 2015 project [2], the semantic indexing (SIN) task aims to recognize the semantic concept contained within a video shot. This task has several challenges such as data imbalance, scalability, and semantic gap [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. 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 include (1) developing robust learning approaches that adjust to the increasing size and the diversity of the videos, (2) fusing information from other sources such as audio and text, (3) detecting the low-level and mid-level features that have a high discriminating capability, etc. [16, 17, 18, 19].

The size of the high-level semantic concepts remains the same as the SIN task of the previous year, which has 60 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 the mean inferred average precision (mean xinfAP) [20].

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

The proposed framework of the TRECVID 2015 SIN task is shown in Figure 1. The key frame level features (KF) are extracted and normalized. In Run 1, the MCA late fusion model is applied; while for Run 2 and Run 4, the MCA early fusion model is applied. Run 3 is based on Run 1 and also includes the time information. The xinfAP values are calculated from models trained on TRECVID 2015 training data and evaluated on TRECVID 2015 testing data.

2.1 Data Pre-processing and Feature Extraction

A key frame for each shot is provided to the SIN task participates in both training and testing videos. Ten kinds of KF features are extracted from each frame in training and testing data, including CEDD [21], Cooccur [22], Gabor [23], Haar [24], LBP [25], Sobel [26], HoG [27, 28], Canny edge histogram [29], and color histograms in HSV and YCbCr spaces. Before extracting the features, histogram equalization is employed to regulate the contrast of frames [30, 19]. Then, these features are extracted and normalized for each key frame.

2.2 Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) [31, 32, 33, 34, 35] is a data analysis technique for nominal/categorical data, which has been used to detect and represent the underlying structures in a data set. The procedure therefore appears to be the counterpart of principle component analysis for categorical data. MCA extends correspondence analysis (CA) by providing the ability to analyze tables containing some measure of correspondence between the rows and columns with more than two variables. To the best of our knowledge, our team is the first one to apply the MCA technique in the area of multimedia information retrieval.

MCA first projects the features and classes into a 2D space spanned by the first principle component (i.e., the eigenvector with the largest eigenvalue) and the second principle component (i.e., the eigenvector with the second largest eigenvalue). The correlation between different feature-value pairs and different classes can be used as an indication of the similarity between them. If one feature has two feature-value pairs in a binary classification application, the angle of one feature-value pair and one class definitely equals the angle of the other feature-value pair and the other class. MCA can be applied for semantic concept detection. One can use the similarity value from the angle file as the criterion to evaluate the distance of the testing data instance to the semantic concept.
Figure 1. The proposed framework for semantic indexing
2.3 Information Fusion

There are two main categories of the fusion methods, namely early fusion and late fusion. For early fusion, we concentrated four (Run 4) / five (Run 2) low-level features as the feature super vectors. The MCA model is then applied for concept detection. On the other hand, late fusion concentrates on the scores generated by the MCA model for each low-level feature. Another MCA model is used to fuse these scores to have a final score for each shot.

Time information is also very important in multimedia data. Motivated by this fact, we extracted the features from frames other than key frames in order to get the relative time information. A concept may not appear or be clearly shown in the key frame of a shot, while the identical concept may be detected in the other frames in the same shot. Therefore, in Run 3, we fused the results from Run 1 with the MCA scores generated by the other frames that carry the time information.

3 Experimental Results

3.1 Data

Given the test collection (IACC.2.C), master shot reference, and concept definitions, for each target concept, a list of at most 2000 shot IDs from the test collection was returned and ranked according to their likelihood of containing the target concept. TRECVID 2015 test data set (IACC.2.C) contains the 200 hours of videos drawn from the IACC.2 collection using videos with durations between 10 seconds and 6 minutes. The train data set combines the development and test data sets from the 2010 and 2011 issues of the SIN Task, namely the IACC.1.tv10.training, IACC.1.A, IACC.1.B, and IACC.1.C data sets. Each contains about 200 hours of videos drawn from the IACC.1 collection using videos with durations ranging from 10 seconds to a little bit longer than 3.5 minutes.

The overall framework of TRECVID 2015 SIN task contains three stages:

1. Model training: using TRECVID 2014 training videos as the training data.
2. Model evaluation: using TRECVID 2014 testing videos as the testing data to evaluate the framework and tune the parameters of the models.
3. Model testing: using TRECVID 2014 training + TRECVID 2014 testing videos as the TRECVID 2015 training data, and TRECVID 2015 testing videos as the testing data to generate the ranking results for the submission.

3.2 Evaluation

A subset of the submitted concept results (20) announced after the submission date were evaluated by the assessors at NIST pooling and sampling. Measures (indexing) are shown as follows [36].

1. Mean extended inferred average precision (mean xinfAP) [37], which allows the sampling density to vary so that it can be 100% in the top strata. This is the most important one for average precision.
2. As in the past years, other detailed measures based on recall and precision are generated and given by the sample_eval software provided by the TRECVID team.
3.3 Performance

All of the measures below were based on the assessment of a 2-tiered random sampling (1-200@100% and 201-2000@11.1%) of the full submission pools and the sample_eval software was used to infer the measures.

Figure 2 to Figure 5 present the performance of our semantic indexing results. The x-axis is the concept number; while the y-axis is the inferred average precision. 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](image1)

**Figure 2.** Run scores (dot) versus median (—) versus best (box) for 2C_M_A FIU UM 15 1

![Figure 3](image2)

**Figure 3.** Run scores (dot) versus median (—) versus best (box) for 2C_M_A FIU UM 15 2

![Figure 4](image3)

**Figure 4.** Run scores (dot) versus median (—) versus best (box) for 2C_M_A FIU UM 15 3

4 Conclusion and Future Work

In this notebook paper, the framework and results of team FIU-UM in TRECVID 2015 SIN task are summarized. We can tell there are still a lot of improvements need to be done based on the results. Some important
directions are desired to 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.
- Deep learning techniques should be integrated to reach a better performance.

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

References


