UQMSG Experiments for TRECVID 2011

Heng Tao Shen, Jie Shao, Zi Huang, Yang Yang, Jingkuan Song, Jiajun Liu, Xiaofeng Zhu
School of Information Technology and Electrical Engineering
The University of Queensland, Australia
{shenht,huang,yang.yang,jiajun,zhux}@itee.uq.edu.au, {j.shao1,uqjsong1}@uq.edu.au

Abstract

This paper describes the experimental framework of the University of Queensland's Multimedia Search Group (UQMSG) at TRECVID 2011. We participated in two tasks this year, both for the first time.

For the semantic indexing task, we submitted four lite runs: L_A_UQMSG1_1, L_A_UQMSG2_2, L_A_UQMSG3_3 and L_A_UQMSG4_4. They are all of training type A (actually we only used IACC.1.tv10.training data), but with different parameter settings in our keyframe-based Laplacian Joint Group Lasso (LJGL) algorithm with Local Binary Patterns (LBP) feature.

For the content-based copy detection task, we submitted two runs: UQMSG.m.nofa.mfh and UQMSG.m.balanced.mfh. They used only the video modality information of keyframes and were both based on our Multiple Feature Hashing (MFH) algorithm that fuses local (LBP) and global (HSV) visual features, with different application profiles (reducing the false alarm rate v.s. balancing false alarms and misses).

Due to time constraint, we were not able to improve the performance of our systems adequately on all the available training data this year for these tasks. Evaluation results suggest that more efforts need to be made to well tune system parameters. In addition, sophisticated techniques beyond applying keyframe-level semantic concept propagation and near-duplicate detection are required for achieving better performance in video tasks.

1 Introduction

In 2011, the Multimedia Search Group within the Data and Knowledge Engineering Research Division at the University of Queensland participated two tasks at TREC Video Retrieval Evaluation (TRECVID) [5] for the first time. They are the semantic indexing and content-based copy detection. Our main intension is to test, with minimum modifications, how related algorithms we developed recently for image tagging [7] and near-duplicate retrieval [6] can perform on a large and highly heterogeneous dataset of Internet Archive Creative Commons videos (referred to as IACC.1) used in TRECVID 2011.

2 Semantic Indexing

The objective of the semantic indexing task is to automatically assign semantic tags (representing high-level features or concepts) to relevant video segments. Specifically, it requires to submit for each semantic feature/concept a list of up to 2000 shot IDs in the test dataset (IACC.1.B), ranked according to the possibility of its occurrence. In this section, we provide a brief description of our experiment for this task.

2.1 Our Approach

The approach we applied here is similar with that described in [7] for local image tagging. In our algorithm, a Laplacian Joint Group Lasso (LJGL) model can collaboratively assign appropriate semantic tags to different regions within each keyframe at the same time. Different from many other methods that take video frames as a whole in the process of semantic indexing, we tried to propagate semantic concepts to specific regions at a more fine-grained level.

All keyframes in the training dataset are preliminarily segmented into regions using the well-known normalized cuts clustering [4], and a reconstruction dictionary is formed by relevant keyframes (segmented k nearest neighbors). Given a keyframe in the test dataset, we also segment it into several regions and extract visual features for each individual region. To construct the dictionary, we first search for its k nearest neighbors in the training dataset, and then concatenate their segmented regions. Subsequently, all these test regions are simultaneously reconstructed based on the dictionary by a Laplacian Joint Group Lasso (LJGL) model. This LJGL model considers the robust encoding ability of joint group lasso [7], and also leverages a Laplacian prior to preserve the local manifold structure among reconstructed regions. Finally, reconstruction coefficients are used to propagate semantic concepts from training regions to test regions.

In our implementation, Local Binary Patterns (LBP) [3] is adopted as the visual feature. LBP assigns each pixel with a value by comparing its eight neighbor pixels with the center pixel and transforming the result to a binary value. The histogram of the values is then accumulated as a local descriptor.

As a new participating group to this task, we submitted four "lite" runs (50 selected concepts) due to limited time for developing our system: L_A_UQMSG1_1, L_A_UQMSG2_2, L_A_UQMSG3_3 and L_A_UQMSG4_4. They have different parameters in our Laplacian Joint Group Lasso (LJGL) algorithm based on keyframes. Because of late registration, we missed the deadline to sign up for collaborative annotation [2] of 2011 IACC training videos (IACC.1.A), so actually we only could use 2010 IACC training videos (IACC.1.tv.traning) which have common feature annotation publicly downloadable from [2] for our system development. According to the TRECVID 2011 guidelines [1], our runs belong to training type A.

2.2 Evaluation and Discussion

The primary performance measure evaluated by NIST is called *mean inferred average precision* [8] per run. The inferred average precisions for the set of concepts officially evaluated in TRECVID 2011 (23 concepts in total) from the submitted concepts for our four runs are given in Table 1.

A general observation is that there are significant differences across the results of different concepts that were evaluated. For example, while some concepts (such as "Adult") have relatively higher inferred average precisions, many concepts that involve certain kind of events (object + action, such as "Dancing", "Demonstration_Or_Protest", "Explosion_Fire", "Running", "Singing", "Sitting_Down" and "Walking") are very hard for our current system. This is because in our system essentially we applied an image tagging approach straightforwardly on keyframes (a main keyframe has been prescribed for each video shot in the master shot reference in TRECVID). While an object can be described in a single frame, this kind of actions can hardly be judged by semantic concept propagation of single frame alone. Also, the numbers of positive samples for the concepts from the training data we used being not sufficient could be another important reason.

Overall, this experiment suggests that more sophisticated techniques especially (i) analysis of multiple keyframes per shot for spatial-temporal feature and (ii) investigation of ontology relationships of concepts ("imply" and "exclude") are needed for better performance in the future. Meanwhile, our current system is purely based on the visually encoded information, and the available donor-supplied metadata (e.g., title, keywords and description) and automatic speech recognition for the English speech may also be used in the construction or running of our future system for semantic indexing.

3 Content-based Copy Detection

The objective of the content-based copy detection task is to automatically find video segments derived

Concept	TV11 ID	Run 1	Run 2	Run 3	Run 4
Adult	2	0.0011	0.0009	0.0017	0.0020
Car	21	0.0003	0.0003	0.0003	0.0003
Cheering	27	0.0001	0.0006	0.0000	0.0002
Dancing	38	0.0000	0.0000	0.0001	0.0000
Demonstration_Or_Protest	41	0.0000	0.0000	0.0000	0.0000
Doorway	44	0.0003	0.0002	0.0004	0.0004
Explosion_Fire	49	0.0000	0.0000	0.0000	0.0000
Female_Person	51	0.0000	0.0001	0.0001	0.0002
Female-Human-Face-Closeup	52	0.0002	0.0003	0.0002	0.0001
Flowers	53	0.0000	0.0000	0.0000	0.0000
Hand	59	0.0004	0.0008	0.0004	0.0003
Indoor	67	0.0008	0.0007	0.0010	0.0009
Male_Person	75	0.0002	0.0001	0.0003	0.0002
Mountain	81	0.0004	0.0006	0.0007	0.0006
News_Studio	83	0.0001	0.0002	0.0002	0.0002
Nighttime	84	0.0001	0.0001	0.0001	0.0001
Old_People	86	0.0001	0.0004	0.0002	0.0006
Running	100	0.0000	0.0000	0.0000	0.0000
Scene_Text	101	0.0004	0.0005	0.0006	0.0005
Singing	105	0.0001	0.0001	0.0001	0.0001
Sitting_Down	107	0.0000	0.0000	0.0000	0.0003
Walking	127	0.0000	0.0000	0.0000	0.0001
Walking_Running	128	0.0001	0.0001	0.0001	0.0002
Mean inferred average precision		0.0002	0.0003	0.0003	0.0003

Table 1: Inferred average precisions for the set of concepts that were evaluated, for all four runs.

from another video by means of various audio/visual transformations. Specifically, given the test collection of videos (IACC.1.A+IACC.1.tv10.training) and a set of 11256 queries it requires to determine for each query the place, if any, that some part of the query occurs, with possible transformations, in the test collection. In this section, we provide a brief description of our experiment for this task.

3.1 Our Approach

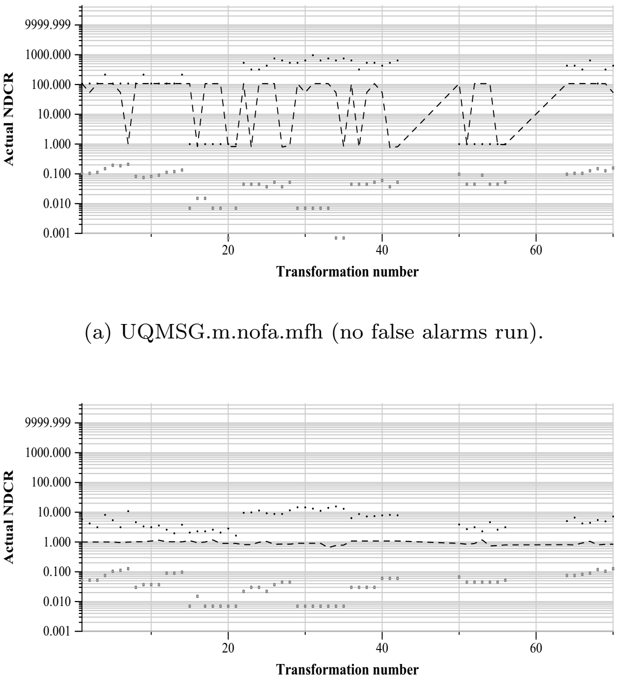
Since the test videos have already been divided into shots with corresponding keyframes prescribed, we also segmented the query videos into shots and extracted their keyframes (one or multiple keyframes may be generated per shot by the method we adopted). Then, we could apply the Multiple Feature Hashing (MFH) algorithm [6] for near-duplicate frame retrieval. Different from many other methods that use a single local or global feature, our basic idea is that the multiple features of videos, each of which reflects the specific information of video data, are often complementary to each other. The algorithm preserves the local structure information of each individual feature and also globally considers the local structures for all the features to learn a group of hash functions which map the test video keyframes into the Hamming space and generate a series of binary codes to represent them.

The MFH algorithm comprises two phases. In the first phase (which is processed offline), a series of s hash functions $\{h_1(\cdot), \ldots, h_s(\cdot)\}$ is learnt, each of which generates one bit hash code for a keyframe according to the given multiple features. Each keyframe has s bits. Using the derived hash functions $\{h_1(\cdot), \ldots, h_s(\cdot)\}$, each keyframe in the test videos can be represented by the generated s-sized hash codes in linear time. In the second phase (which is processed online), the keyframes of the query videos are also represented by s-sized hash codes mapped from the s hash functions. Video copies are expected to have the same or similar hash codes. Thus, with the obtained hash codes only bit XOR operation in the Hamming space is performed to compute the distance. In this way, near-duplicate frames can be detected very efficiently. MFH is more capable of characterizing the video content, and suitable for detection in large scale video datasets. Detailed technical information about our approach can be found in [6].

In our implementation, we used the information of video modality only. Each keyframe is represented by the local feature LBP and the global feature HSV color histogram respectively. These two visual features are fused in the joint framework of MFH for computing a score of confidence measure.

As required by the TRECVID 2011 guidelines [1], we submitted two runs (one for each of two application profiles): UQMSG.m.nofa.mfh is to reduce the false alarm rate to 0 and then optimize the probability of miss (in our system, threshold of the confidence measure is set as 0.75), while UQMSG.m.balanced.mfh is to balance false alarms and misses (in our system, threshold of the confidence measure is set as 0.5). Due to limited time we did not use the standard training video dataset (tv9.sv.test+tv7.sv.devel+tv7.sv.test) in the system development. Instead, some other web video dataset in [6] is reused for learning hash functions[1].

Our approach is essentially based on keyframes, and thus by nature it could not achieve best results for localization of a asserted video copy. When a video copy is detected, how accurately the system locates the exact extent of the copy in the reference video can be very challenging (in TRECVID 2011, copy location accuracy is seen as a secondary, diagnostic performance measure).



(a) UQMSG.m.nofa.mfh (no false alarms run).

(b) UQMSG.m.balanced.mfh (balanced run).

Figure 1: Actual NDCR results of submitted runs, showing comparison of our run score (dot) versus median (- - -) versus best (box) of all submissions by each transformation.

3.2 Evaluation and Discussion

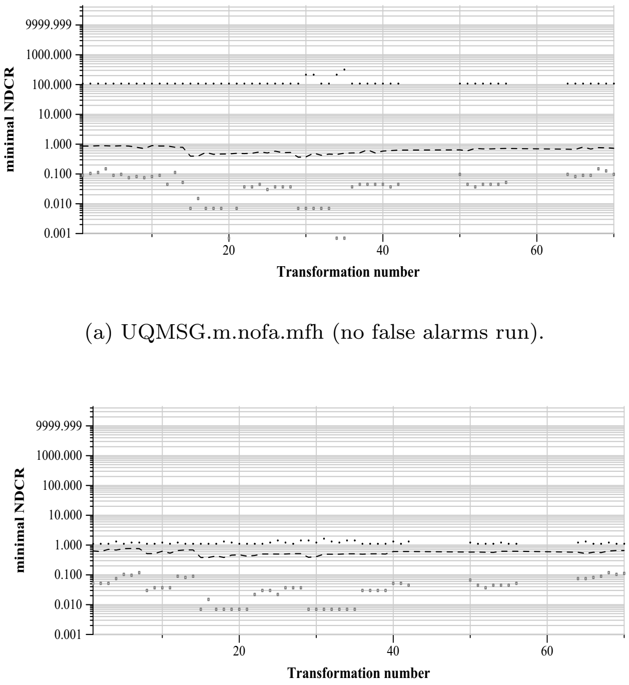
The primary performance measure evaluated by NIST is called *Normalized Detection Cost Rate* (NDCR), which is calculated separately for different transformations for each run. The Detection Cost Rate (DCR) involves a copy target rate R_{Target} and combines two error rates: the probability of miss error (P_{Miss}) and the false alarm rate (R_{FA}). Specifically, it is defined as

$$C_{Miss} \cdot P_{Miss} \cdot R_{target} + C_{FA} \cdot R_{FA},$$

where C_{Miss} is the cost of a miss, C_{FA} is the cost of a false alarm, and R_{target} is the priori copy target rate. The costs and copy target rate depend on the application. For TRECVID 2011, $R_{target} = 0.005/\text{hr}^2$, C_{Miss}=10, and C_{FA}=1000 (for the "no false alarm" profile) or C_{FA}=1 (for the "balanced" profile). In order to compare the detection cost rate values across a range of parameter values, DCR is normalized as

$$NDCR = \frac{DCR}{C_{Miss} \cdot R_{target}} + \beta \cdot R_{FA} = P_{Miss} + \beta \cdot R_{FA},$$

[1] It is worthwhile to mention that both the Sound and Vision (S&V) training video data and our training video data differ markedly from the 2011 IACC test videos.



(a) UQMSG.m.nofa.mfh (no false alarms run).

(b) UQMSG.m.balanced.mfh (balanced run).

Figure 2: Minimal NDCR results of submitted runs, showing comparison of our run score (dot) versus median (- - -) versus best (box) of all submissions by each transformation.

where $\beta = C_{FA}/(C_{Miss} \cdot R_{target})$. The minimal NDCR (as a function of the decision threshold) and associated decision threshold are calculated for each transformation, for each run. The smaller the minimal NDCR, the better the performance. The actual NDCR is also calculated for each transformation and run using the submitted decision threshold value for optimal performance. Figure 1 and Figure 2 show the officially released actual and minimal NDCR results of our runs respectively, along with comparisons versus median and best among all submissions of this year.

From the figures, it can be observed that while our no false alarms run can achieve reasonably good actual NDCR results for certain transformations (for transformations 15∼21 and 50∼56 which correspond to video transformation categories T3 "insertions of pattern" and T8 "post production" respectively, our results are better than the medium results), the overall performance of our system is still unsatisfactory. There are some reasons. First, the system was not well trained yet as the learning of hash functions was largely unoptimized. Second, an appropriate choice of the threshold used for the two different application profiles plays an important role in the detection system because it has direct and significant influence on what are measured here (particularly, the calculation of NDCRs). However, due to time constraint, we were not able to well tune our system adequately and study the threshold value for optimal performance. In addition, we also found later that in the implementation many black frames were extracted from the query videos by the software package we used, and unfortunately these spurious keyframes affected the final detection results of our runs quite severely.

This experiment also suggests that in the future, we mainly need to (i) resolve keyframe sampling issue that causes many true video copies are missed (due to different keyframes extracted for comparison) and makes copy location accuracy a defect in the algorithm, and (ii) conduct video+audio detection for the task by further incorporating the audio modality information into our experimental framework.

4 Summary

This paper provides an overview of our experimental framework at TRECVID 2011 semantic indexing and content-based copy detection tasks and gives comments of our experience. For our first year of participation, the main objective was to put in place our test environment and deliver results with some algorithms that were not necessarily optimal for the exact TRECVID tasks. Thus, we were not expecting to perform very well as our systems were initially designed for some other problems. Nonetheless we have clearly identified the major challenges and our future work directions for TRECVID benchmark.

Acknowledgement

We are grateful to Dr. Yi Yang (Carnegie Mellon University) for his contributions made in [7, 6] when he was previously with our group.

References

[1] Guidelines for TRECVID 2011. http://www-nlpir.nist.gov/projects/tv2011/tv2011.html.

[2] TRECVID 2011 Collaborative Annotation. http://mrim.imag.fr/tvca.

[3] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[4] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[5] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[6] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *ACM Multimedia*, 2011.

[7] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, and F. Nie. Tag localization with spatial correlations and joint group sparsity. In *CVPR*, pages 881–888, 2011.

[8] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM*, pages 102–111, 2006.