The University of Queensland at TRECVID 2007 Search Task

Heng Tao Shen, Xiaofang Zhou, Jie Shao, and Zi Huang
School of Information Technology and Electrical Engineering
The University of Queensland, Australia
{shenht, zxf, jshao, huang}@itee.uq.edu.au

Abstract

This paper describes our first participation in TRECVID. We took part in the search task and submitted two interactive runs. Both of them are of Type c, and use no ASR/MT output information: run id I_c_2_UQ1_1 uses 64-dimensional visual features and run id I_c_2_UQ2_2 uses 32-dimensional visual features for retrieval. Based on the evaluation results with the benchmark video data, we observed that there is no significant difference between these two runs in terms of what measured. Meanwhile, utilizing visual information only seems not easy to capture the high-level semantic similarity required by the search topics of this year, so more sophisticated approaches such as multi-modality fusion are needed to improve the system performance.

1 Introduction

The rapid increase in the generation and dissemination of digital videos has attracted the Data and Knowledge Engineering research group at the University of Queensland to develop effective and efficient solutions of analyzing, indexing and searching very large video databases [10, 6, 9]. This year is the first time for us to participate in NIST TREC Video Retrieval Evaluation (TRECVID), and for the search task we submitted the results of two interactive runs, with emphasis on testing how the visual features can contribute to search effectiveness in a video retrieval system we developed recently [11].

Our search model is mainly based on the statistical summarization of individual frame features to capture the dominating content and content changing trends of a video shot in vector space. We have not incorporated the text transcripts extracted from the common Automatic Speech Recognition (ASR) engine [4] and Machine Translation (MT) output provided yet, as TRECVID requires a baseline search run treating the videos as if no ASR/MT for the languages used existed, to understand the visually encoded information. No training data (shared or private) specific to any sound and vision data has been used in the construction or running of our system, so according to guidelines of TRECVID 2007 [2], our search system belongs to Type c. During the interactive search, in an effort to obtain the initial answer set for providing an entrance for user interaction, the known positively classified examples of shots from the development video data collection are used as the query input in the first iteration for each search topic at the beginning. A dynamic browsing interface then can provide the user with an easy-to-use interaction functionality. In the rest of this paper, we provide brief descriptions of our experimental system, and give comments of some results returned from the evaluations by NIST with the standard TRECVID 2007 video collection.

2 Search Process

2.1 Overview

The retrieval methodology we applied to TRECVID is functionally same with that described in [11], i.e.,

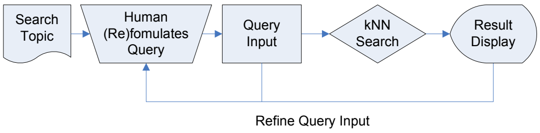


Figure 1: Framework of interactive search system.

the search process still follows a query-by-example style. We measure some proposed distance value in feature space and apply k Nearest Neighbor (kNN) search method for score ranking. Meanwhile, a graphical user interface is needed here to realize the interactive search for incorporating relevance feedback. As can be seen from the entire flow of framework displayed in Figure 1, for all the 24 search topics (Topic 197 to Topic 220) of this year, the initial query examples used in our system are obtained from the provided video examples which come from the development data (for consistency reason the provided external image examples are all discarded). For each of these search topics, the video shots in the test dataset can be ordered by the probability of satisfying the search criterion, and the top listed video shots will be displayed simultaneously in a page as the search result. From the relevant videos returned from the system, a user can choose arbitrary one as the new query example to be used in the next iteration to refine the initial search process during the course of interaction, so the output of ranked video shot list can be changed.

This year, our main aim is to test the existing video retrieval system we developed to understand the effect of visual features on search effectiveness in the context of semantic similarity query, so both the two runs we submitted are purely based on visual information. Our system is basically not trained on the development sound and vision data, so it belongs to the search system of Type c, according to the isolation definition specified in TRECVID's guidelines [2, 12].

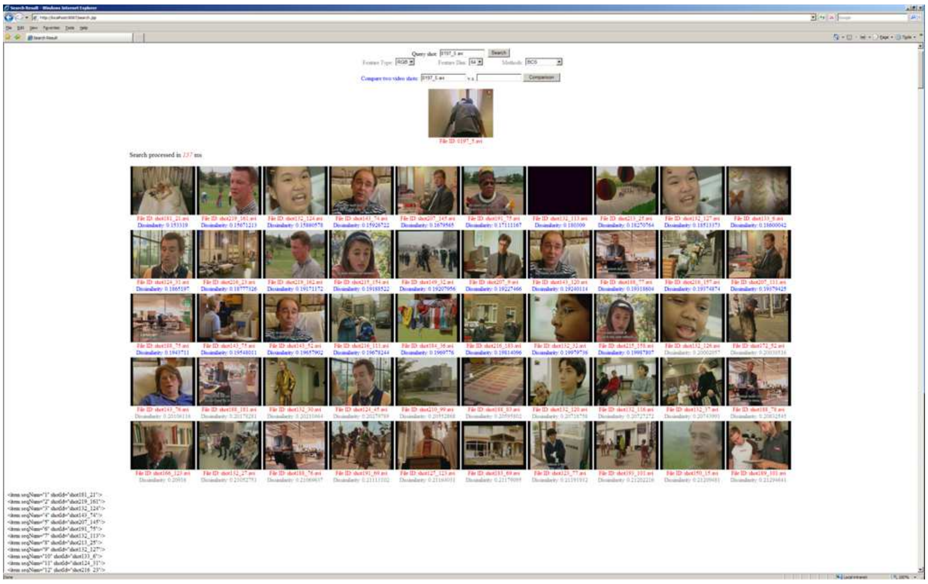
2.2 Shot Representation and Search

Since NIST has prescribed the shot boundaries of both development and test videos, the basic granularity for retrieval is shot. Based on the common master shot reference provided by [7], 18,120 and 18,142 video shots are segmented for the development and test video data, respectively.

In our system, each shot is not represented by its key-frame(s), but represented by continuous video frames. We first convert all the video shots from MPEG format to AVI format and resize them to the smaller resolution with VirtualDub [3] in a batch mode for extracting d-dimensional global visual features of video frames. Many kinds of image features can be employed, such as RGB color histograms, HSV color histograms, etc. Next, a compact summarization technique which tries to capture the dominating content and content changing trends of a video segment is applied here for retrieval [11]. By exploiting the frame feature point correlation existed in vector space with Principal Component Analysis (PCA), the proposed Bounded Coordinate System (BCS) descriptor can represent each video shot by a mean of all its frame points (origin) and d orientations and ranges (bounded principle components). Since the mean and bounded principle component are both d-dimensional vectors, a BCS descriptor actually consists of $(d+1)$ d-dimensional vector points. In short, the video content is analyzed in an offline process that involves visual feature extraction and statistical modelling for each shot of test video collection.

Unlike many key-frame based representations, the search process here is based on the statistical information derived from all the video frames within each shot. In order to compare the similarity of video shots, we compute the distance of their corresponding BCS descriptors. The query processing is performed with the kNN method based on our shot representation specific distance metric in a d-dimensional feature space, which indicates the degree of visual similarity to query.

As we know, the computation cost for video ranking of large scale video dataset is usually very large. This problem becomes more critical for a retrieval system that runs interactive search. The compact summarization can be an efficient solution to tackle this challenge. The similarity measure of BCS descriptor only involves some translation, rotation and scaling operations, and the computational complexity



Query shot: 0197_5.avi Search

Feature Type: RGB Feature Dim: 64 Methods: BCS

Compare two video shots: 0197_5.avi v.s. Comparison

File ID: 0197_5.avi

Search processed in 237 ms

```
<item seqName="1" shotId="shot181_21"/>
<item seqName="2" shotId="shot219_161"/>
<item seqName="3" shotId="shot132_124"/>
<item seqName="4" shotId="shot143_74"/>
<item seqName="5" shotId="shot207_145"/>
<item seqName="6" shotId="shot191_75"/>
<item seqName="7" shotId="shot132_113"/>
<item seqName="8" shotId="shot213_25"/>
<item seqName="9" shotId="shot132_127"/>
<item seqName="10" shotId="shot133_6"/>
<item seqName="11" shotId="shot124_31"/>
<item seqName="12" shotId="shot216_23"/>
```

Figure 2: Interface for interactive search.

is linear in the dimensionality of feature space and independent of video length [11]. Therefore, it greatly improves the search efficiency and is especially suitable for interactive search system with the real-time response requirement for feedback.

2.3 Interactive Search Interface

A convenient and efficient interface is important for a video search system to facilitate user interaction. Our application is implemented in JSP with Apache Tomcat web server so it can be viewed with any standard web browser online, as shown in Figure 2. Once a query is submitted, the server is able to retrieve the most similar shots from the database of test videos. This interface can result in an impressive number of video shots being reviewed within the 15 minutes time limit during the course of interactive search. For each iteration, any relevant video shot can be selected as the new query example to be used next. The strategy is that when a shot is marked as relevant by user, the neighbors of this shot (e.g., the nearby video shots which are sharing a same sequence number) are inserted at the top of the ranked list of shots to be presented to users, so some further refinement could be conducted. We then can get a new ranking of video shots according to the feedback and a new layout of the set of top ranked videos can be displayed for browsing on the screen.

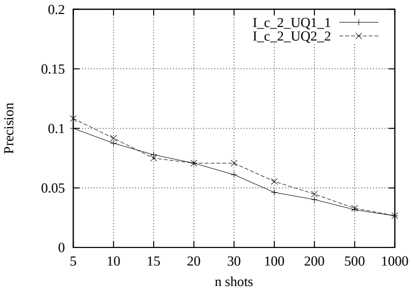


Figure 3: Precision at n shots.

3 Evaluations and Discussions

We submitted two interactive runs to NIST for evaluating performance this year. They are:

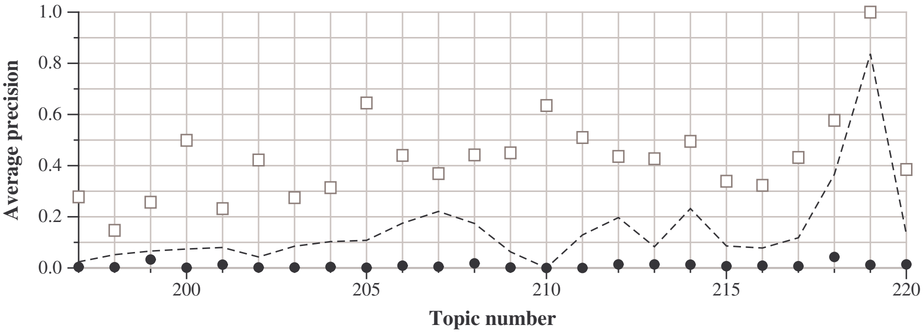


Figure 4: Run score (dot) versus median (- - -) versus best (box) by topic for I_c_2_UQ2_2.

- I_c_2_UQ1_1: no ASR/MT output used, but 64-dimensional features of visual information
- I_c_2_UQ2_2: no ASR/MT output used, but 32-dimensional features of visual information

Experimental results of our search runs by NIST official evaluations are shown in the Figure 3. While may not providing statistical significance, the search effectiveness differs not much between two runs of different dimensionalities according to the numerical results. This can be explained as our method applied mainly summarizes each video shot into a single and small representative (BCS) which exploits the frame *correlation* to capture the dominating content and content changing trends, while the detailed features of individual frames become less important. The Mean Average Precision (MAP) of our search system is 0.010. In the next, we just use the second run, whose run id is I_c_2_UQ2_2 to analyze some lessons learned.

The overall performance of our system is not very satisfactory, as shown in Figure 4 by the comparison with other TRECVID submissions of this year. The main reason is due to its intrinsic design philosophy and our choice to focus on the effect of visual information only rather than combining multiple information sources for the first year. Although this kind of compact shot representation can facilitate retrieval of near-duplicate videos which are visually similar, it is generally not tolerant to the semantic concept similarity. As demonstrated by the search topics of this year, the trend is focusing more on event (object + action), instead of finding an object that can be described in a single frame [2]. Moreover, we have not incorporated the ASR/MT transcripts yet. As a consequence, more improvements are needed to detect high-level similarity in the future, by utilizing multi-modality fusion to couple with other concept features and text features as well.

Another consideration is that, as pointed out in [8], the interactive run in TRECVID models a search scenario where someone else searches the video database on behalf of the people who finally judges the results, which is somewhat different from the widely used model of relevance feedback in where the searcher is the people who judges the system performance. For a normal user, he/she is usually more content with the top ranked good results rather than retrieving all the relevant items in the database. Subjectively the top ranked results are good on many of search topics, which can be indicated by the precision at n shots.

4 Conclusions

This year we primarily applied our proposed statistical summarization method to TRECVID benchmark test video data. Our search effort mainly relies on a system that exploits the frame feature correlation within each shot for detecting low-level visual similarity. In the future, we will pay more attention on semantic based video retrieval to improve the search performance at TRECVID. We will utilize the visual concepts which are detected in the high-level feature extraction task, e.g., measuring concept similarities in multimedia ontologies [5]. Meanwhile, the speech text information provided by the ASR engine and MT output can also be utilized in the system, e.g. extracting named-entities from the text query topics and searching against text transcripts with the Lemur toolkit [1]. With these visual, semantic and speech modalities, the result can be expected to be better.

Acknowledgement

The work reported in this paper is supported in part by Australian Research Council (ARC) under grant DP0663272.

References

[1] The lemur toolkit for language modeling and information retrieval. http://www.lemurproject.org.

[2] Trecvid 2007 guidelines. http://www-nlpir.nist.gov/projects/tv2007/tv2007.html.

[3] Virtualdub. http://www.virtualdub.org.

[4] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of Second International Conference on Semantics and Digital Media Technologies (SAMT)*, 2007.

[5] M. Koskela, A. F. Smeaton, and J. Laaksonen. Measuring concept similarities in multimedia measuring concept similarities in multimedia. *IEEE Transactions on Multimedia*, 9(5):912–922, 2007.

[6] H. Lu, B. C. Ooi, H. T. Shen, and X. Xue. Hierarchical indexing structure for efficient similarity search in video retrieval. *IEEE Trans. Knowl. Data Eng.*, 18(11):1544–1559, 2006.

[7] C. Petersohn. Fraunhofer hhi at trecvid 2004: Shot boundary detection system. In *TRECVID*, 2004.

[8] M. J. Pickering, D. Heesch, S. M. Rüger, R. O'Callaghan, and D. R. Bull. Video retrieval using global features in keyframes. In *TREC*, 2002.

[9] J. Shao, Z. Huang, H. T. Shen, X. Zhou, and Y. Li. Dynamic batch nearest neighbor search in video retrieval. In *ICDE*, pages 1395–1399, 2007.

[10] H. T. Shen, B. C. Ooi, X. Zhou, and Z. Huang. Towards effective indexing for very large video sequence database. In *SIGMOD Conference*, pages 730–741, 2005.

[11] H. T. Shen, X. Zhou, Z. Huang, and J. Shao. Statistical summarization of content features for fast near-duplicate video detection. In *ACM Multimedia*, pages 164–165, 2007.

[12] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Multimedia Information Retrieval*, pages 321–330, 2006.