Scalable Mobile Video Question-Answering System with Locally Aggregated Descriptors and Random Projection

Guan-Long Wu[†], Yu-Chuan Su[⊙], Tzu-Hsuan Chiu[⋆], Liang-Chi Hsieh[⊕], Winston H. Hsu[‡]
{garywgl[†], r96098[⋆], winston[‡]}@csie.ntu.edu.tw, spooky@cmlab.csie.ntu.edu.tw[⊙],
viirya@gmail.com[⊕]
National Taiwan University, Taipei, Taiwan

ABSTRACT

We present a scalable mobile video Question-Answering system with locally aggregated descriptors and random projection using user-generated videos all around the world for ACM Multimedia 2011 Technicolor challenge: "precise event recognition and description from video excerpts." Our proposed system takes a video excerpt as a query and explores its canonical semantics of that. We collect a public events video dataset containing 7 topics with 1963 YouTube videos to evaluate our proposed system. The experiment results show that our proposed video feature representation outperforms a state-of-the-art near-duplicate retrieval based on color histogram. The signature generated by random projection not only ensures real-time efficiency but also achieves a competitive MAP performance with original feature.

Categories and Subject Descriptors: H.3.3 [Information Search And Retrieval]: Search process; H.3.1 [Content Analysis and Indexing]: Abstracting methods

General Terms: Experimentation, Performance

1. INTRODUCTION

Targeting on the topic of "precise event recognition and description from video excerpts" for ACM Multimedia 2011 Technicolor Challenge, we present a novel and promising mobile video Question-Answering (QA) system. Comparing to the traditional text search interface, our proposed video QA system based on video excerpts helps users to investigate the semantics in a video effectively. Users do not need to come up with the appropriate keywords as query, and the QA system returns the top 5 ranked semantics once they upload a query video. The snapshot of our proposed system is illustrated in Figure 1.

There are two main challenges for the proposed system. The first is how a system processes the explosive growing user-generated videos efficiently. Take YouTube as an example, according to Wikipedia, YouTube receives more than 3 billion views per day and users upload more than 24 hours of videos per minute. The second is how to explore the multi-

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

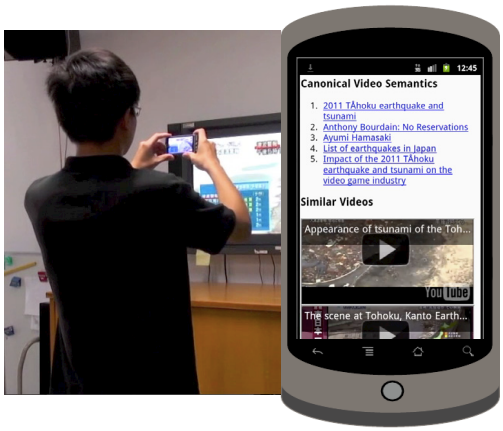

Figure 1: The snapshot of mobile video QA system.

lingual semantics in a video excerpt. YouTube reported that around three quarters of data are uploaded outside the US. However, people in different countries use different languages to annotate videos. To tackle these challenges, searching video content directly by video excerpts can retrieve multilingual information more intuitively than keywords, because public event videos are similar in visual appearances. Therefore, we can explore the video semantics in multiple languages based on near-duplicate video retrieval.

In this paper, we present a scalable video QA system with locally aggregated descriptors and random projection. We review the related work briefly in Section 2. Description of our proposed work is presented in Section 3. We show the experiment results and discuss them in Section 4, and conclusion is given in Section 5.

2. RELATED WORK

2.1 Question-Answering System

A QA system requires meeting precise information need of user by natural language or media. A typical text-based QA system answers questions written in natural language by text automatically. Nowadays, it is an important issue to leverage the resources of the community websites which have a large amount of archives to answer user's question by finding the most related question. Wang *et al.* [7] propose a new retrieval framework for community-based QA to tackle the similar question matching problem based on syntactic tree structure. However, the knowledge base of the work

is Yahoo! Answers and it cannot be applied on YouTube easily due to the lack of structural data for text semantics. Therefore, the effectiveness of typical text-based QA system is limited.

For multimedia QA, Multimedia contents are suitable for answers in several types of questions. For example, user can easily learn to tie shoelaces by watching a demo video clip instead of reading a paragraph describing the steps. Li *et al.* [5] propose a multimedia QA system named *Video Reference* where the input is a sentence of question and the output is the most relevant video on YouTube. Although it leverages the video contents on YouTube, it still uses text as queries only.

2.2 Near-duplicate Video Retrieval

Near-duplicate video retrieval has been receiving much attention in recent years due to the rapid growth of online video services. Most existing works on duplicate-video retrieval focus either on videos that are very similar in visual content or video copy detection.

For videos similar in visual content, it turns out that simple color histogram may provide reasonable result. Using 24 bins color histogram in HSV color space, Wu *et al.* achieve 0.891 MAP on CC_WEB_VIDEO dataset [8]. However, color histogram is insufficient for "semantic duplicate" video detection, and it is also insufficient for partial duplicate video detection, which is a common situation in video QA system. "Semantic duplicate" means the videos share a same semantic definition (e.g. public events) although they are not very similar in visual appearance. To detect partial duplicate videos, temporal network that utilizes the temporal relation in video segments are introduced [6]. Despite its success, the method requires more storage and computational cost than signature-based methods, which limits the scalability of the system.

Video copy detection aims to find videos that come from the same source video but undergo different distortions. To solve the problem, Douze *et al.* [3] proposed a complicated frame matching method. Although such method provides high precision for duplicate video detection even the video suffers from severe degradation, the complex computation hinders the scalability and is thus not suitable for the proposed video QA system.

3. DESCRIPTION OF PROPOSED WORK

3.1 System Overview

Our proposed system is composed by an online part and an offline part. The illustration is shown at Figure 2. In the offline part, videos from YouTube or other video sharing websites are processed, and video feature vectors are extracted by the video feature extraction component. The contextual data of video (e.g. title, tags, description) is used in video profile generation and every video has its own profile which is constructed by leveraging Google Search and Wikipedia.

In the online part, the system compares the query feature vector from a video except uploaded by a user to the ones in database, and finds the top ranking database videos as candidates. Then, the canonical semantics selection ranks the items in candidate profiles and returns the top ranked semantic items to the user. The technical detail of video representation extraction is described in Section 3.2 and Section 3.3, and the descriptions of video profile generation and canonical semantics selection are shown in Section 3.4.

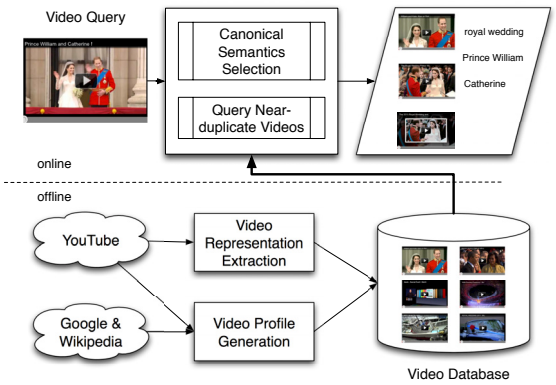

Figure 2: The architecture of our proposed system.

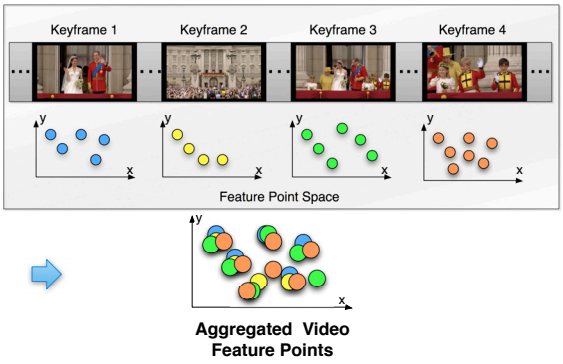

Figure 3: A sample video representation for a video of "Wedding of Prince William and Catherine Middleton."

3.2 Video Representation

Local features have been shown to be very discriminative visual features. Several methods have been proposed to transform local features to an image-wise vector representation. Among them, the vector of locally aggregated descriptors (VLAD) [4] was proposed to remedy the quantization errors suffered by the popular bag-of-word method. Based on VLAD, the system generates a vector which aggregates difference of the SIFT descriptors and the codebook centers in all keyframes. Note that keyframes are extracted by abrupt-shot detection to represent a video. The illustration is shown at Figure 3.

In our implementation, the number of center is set to 64 and the 128 dimensions of Hessian-Affine SIFT points are extracted from video keyframes so that the dimension of the feature vector is 8192. The distance metric is L_2 distance.

3.3 Dimension Reduction for Scalable Computation

In order to reduce the storage and computation time for comparing videos in database, we use a random projection in the dimension reduction process. Random projection (RP) is an effective dimension reduction method for wide range of domains. Bingham *et al.* [2] show the performance com-

parison between the random projection and other dimension reduction techniques (e.g. principle component analysis and discrete cosine transform) in image and text data. RP has a comparable performance with others but the computation cost is greatly reduced.

RP uses random matrix $R(i,j) = r_{ij}$ which follows the Gaussian distribution but it is usually not met in practice due to the higher computation cost. Achlioptas [1] presents a simpler random matrix following a probability distribution:

$$r_{ij} = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2} \end{cases} \quad (1)$$

In practice, the above equation has a lower computation cost because the integer mathematics operation is faster than floating point mathematics operation.

Assume the original dimension of feature vector is D and the reduced dimension is N, the dimension of random projection matrix is $[D \times N]$. Equation 1 is used to generate the projection matrix in our proposed work.

After applying the random projection to the feature vector, we get a reduced vector V and we quantize the vector V based on the equation:

$$V_i = \begin{cases} 1 & \text{if } V_i \geq 0, \\ 0 & \text{if } V_i < 0 \end{cases} \quad (2)$$

to generate the signature. We use hamming distance as the distance metric.

The effectiveness of using random projection can be discussed from two perspectives: computation cost and storage. For the computation cost issue, it is intuitive that the computation time for hamming distance is less than norm-2 distance because hamming distance can be calculated using XOR operation and counting the bits set after the XOR operation. For the storage issue, an original feature vector requires $D \times 8$ bytes (the storage of double floating point is 8 bytes) to store, and the signature uses $\frac{N}{8}$ bytes only. Thus, the signature-based representation reduces both the storage and the computation cost. We show the mean average precision performance of signature using different number of bits and the original feature representation in section 4.

3.4 Canonical Video Semantics Selection

In this part, the system completes the canonical video semantics selection in two phases. The first phase is the Wikipedia profile generation using Google Search and it is an offline process. For each video, the video title is used as the query keywords in Google Search and the titles of the top T_p returned wikipedia article form the profile of the video.

For instance, if a user uses "President Barack Obama 2009 Inauguration" as a query and Google Search returns the wikipedia articles such as "Inauguration of Barack Obama," "United States presidential inauguration" and "Presidency of Barack Obama," etc. The main reason of using Google Search instead of the default searching mechanism[1] in Wikipedia is that Google Search provides more accurate ranked results in general.

The second phase, it is an online query process, the canonical semantics selection. For a query video excerpt, the feature vector of the query is extracted. Then, the system returns the ranking list according to the similarity between the query video and the videos in the database. The canonical video semantics of the query is voted by the top T_v videos in the returned ranking list and the ranking items in their profiles. The equation 3 lists the scoring function:

$$score(i) = \sum_{r \in T_v} \sum_{\exists i \in profile_r} \frac{1}{\sqrt{rank_r}} \frac{1}{\sqrt{rank_i}} \quad (3)$$

i represents a semantic item in a video profile, $rank_r$ shows the ranking based on distance of video vectors, and $rank_i$ describes the ranking in a video profile.

In our proposed system, the interface shows the top 5 video semantics. We compare the performance of Prec@5 based on the method but using our proposed video representation and a color-based feature [8] in the Section of experiment result.

[1]http://en.wikipedia.org/wiki/Special:Search

4. EXPERIMENT RESULTS

4.1 Video Dataset

To evaluate the performance of our proposed system, we collect a public event video dataset using YouTube API. The 7 public events listed below are selected and the topic is used as the query terms to collect video data:

1. "2011 Tohoku earthquake and tsunami"
2. "apple ipad 2 steve jobs"
3. "Beijing Olympics 2008 Opening ceremony"
4. "President Barack Obama 2009 Inauguration"
5. "September 11 attacks"
6. "Wedding of Prince William and Catherine Middleton"
7. "windows 98 crash bill gates"

For each topic, we utilize YouTube API to download the top 100 videos according to "relevance," "published," "viewCount," and "rating," respectively, and the results are then aggregated together. Hence, there should be no more than 400 videos in a query topic. At the end, the dataset contains 1963 videos with 89618 keyframes. The size of video files (flv format) is about 27 GB. We label the ground truth of these videos and 626 videos are verified belonging to one of the 7 event topics.

4.2 Experiment Setup

In the experiment of this work, the dimension of the proposed local-feature-based video representation is 8192, and the dimension of the color-based feature is 24 following the setting of [8]. The programs are implemented using Python and run at a machine contains an 8-core CPU and 48GB memory. In canonical semantics selection, the top 20 videos are considered in the voting step.

4.3 Performance Comparison

Figure 4 shows the mean average precision (MAP) comparison of our proposed feature and the color-based baseline for the ground truth videos in semantic duplicate retrieval. We use each ground truth videos as a query and calculate average precision using leave-one out method where the query depth is 50 (The query video must be removed from database for each query.) Finally, the mean average precision for each event topic is calculated.

The MAP of our proposed video feature outperforms the color-based baseline, except for the "September 11 attacks" event topic. We conjecture the reason is that the visual

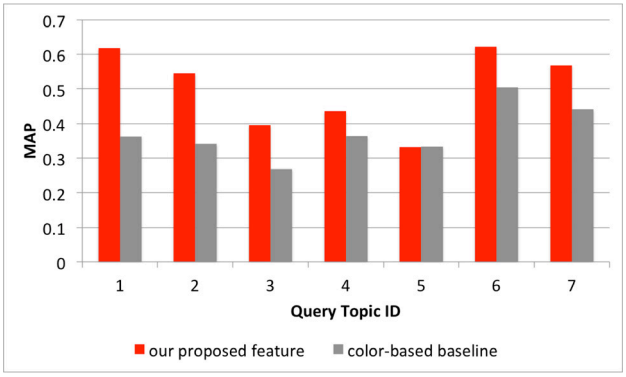

Figure 4: The MAP comparison of our proposed feature and the color-based baseline for the ground truth videos in semantic duplicate retrieval.

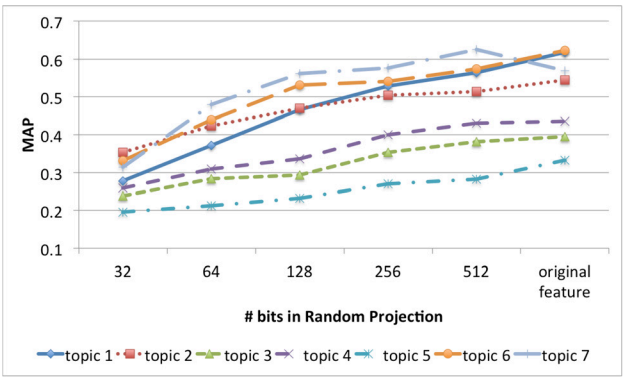

Figure 5: The MAP comparison of different number of bits in Random Projection for the ground truth videos in semantic duplicate retrieval.

pattern (e.g. explosion) in "September 11 attacks" is not captured well by local features.

Figure 5 shows the MAP comparison of different number of bits in RP for the ground truth videos in semantic duplicate retrieval. The MAP performance decreases when using less number of bits to represent a video in general cases. However, the 512 bits configuration achieves a satisfactory performance comparing to the original features and saves lots of storage and computation cost.

Figure 6 shows the precision at 5 of documents retrieved (P@5) comparison between our proposed feature and the color-based baseline for canonical video semantics. We randomly select 20 ground truth videos for each topic. If the number of ground truth videos for a topic is less than 20, we choose all of that. For each video, we let users label whether each of the 5 canonical video semantic is relevant. It's worth noting that the comparison between the performances is more important than the exact value, because the words or sentences well describing a video varies in number.

The P@5 of our proposed feature outperforms the color baseline in 5 topics, but it is worse in "September 11 attacks" and "Windows 98 crash bill gates." It probably contains more than one factor to affect the subjective QA performance. However, in the same video semantics selection mechanism,

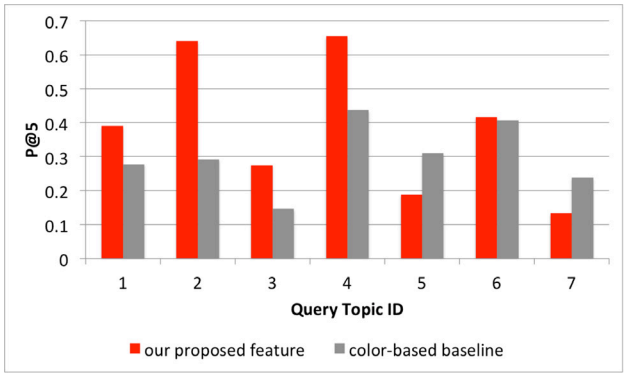

Figure 6: The P@5 comparison of our proposed feature and the color-based baseline for canonical video semantics.

our proposed feature provides higher P@5 than the color-based baseline.

5. CONCLUSION

We propose a scalable mobile video Question-Answering system with locally aggregated descriptors and random projection. User can simply record the interested video excerpt by mobile phone and upload it to our system to get the possible video semantics. Note that users do not need to provide any textual information in this scenario. The snapshot of our proposed system is shown in Figure 1. Thus we consider that it is a promising solution to target at ACM Multimedia Technicolor Challenge.

For scalable computation, our experiment results show the signature-based video representation using random projection to reduce the storage usage and computational cost effectively and reaches a corresponding performance comparing to the original feature representation.

6. REFERENCES

[1] D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66, 2003.

[2] E. Bingham et al. Random projection in dimensionality reduction: applications to image and text data. In *ACM SIGKDD*, 2001.

[3] M. Douze et al. Inria-lear's video copy detection system. In *TRECVID*, 2008.

[4] H. Jégou et al. Aggregating local descriptors into a compact image representation. In *IEEE CVPR*, 2010.

[5] G. Li et al. Video reference: question answering on youtube. In *ACM MM*, 2009.

[6] H.-K. Tan et al. Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *ACM MM*, 2009.

[7] K. Wang et al. A syntactic tree matching approach to finding similar questions in community-based qa services. In *ACM SIGIR*, 2009.

[8] X. Wu et al. Practical elimination of near-duplicates from web video search. In *ACM MM*, 2007.