A MULTIMODAL AND MULTILEVEL RANKING FRAMEWORK FOR CONTENT-BASED VIDEO RETRIEVAL

Steven C.H. Hoi and Michael R. Lyu

Department of Computer Science and Engineering and the Shun Hing Institute of Advanced Engineering
The Chinese University of Hong Kong, Shatin, Hong Kong
E-mail:{chhoi, lyu}@cse.cuhk.edu.hk

ABSTRACT

One critical task in content-based video retrieval is to rank search results with combinations of multimodal resources effectively. This paper proposes a novel multimodal and multilevel ranking framework for content-based video retrieval. The main idea of our approach is to represent videos by graphs and learn harmonic ranking functions through fusing multimodal resources over these graphs smoothly. We further tackle the efficiency issue by a multilevel learning scheme, which makes the semi-supervised ranking method practical for large-scale applications. Our empirical evaluations on TRECVID 2005 dataset show that the proposed multimodal and multilevel ranking framework is effective and promising for content-based video retrieval.

Index Terms— video retrieval, multimodal fusion, multilevel ranking, semi-supervised learning, performance evaluation

1. INTRODUCTION

With the rapid growth of digital devices, Internet infrastructures, and Web technologies, videos nowadays can be easily captured, stored, uploaded, and delivered over the Web. Although general Web search engines have achieved many successes, searching video content over the Web is still challenging. Most commercial Web search engines usually only index meta data of videos and search them by texts. Without understanding the media contents, traditional search engines may be limited in video retrieval. There should be a great room in improving traditional search engines for video retrieval through exploiting the rich media contents. This makes content-based video retrieval (CBVR) a promising direction for developing future video search engines.

In the past decades, content-based image retrieval has been actively studied in signal processing and multimedia communities [1]. Recently content-based video retrieval has attracted more and more research attention. From 2001, TREC Video Retrieval (TRECVID) evaluation has been set up for benchmark evaluations of video retrieval [2]. In general, a content-based video search engine can be built upon a traditional text-based search engine with extracted video contents, such as speech recognition scripts, close captions, and video Optical Character Recognition (OCR) texts. Although such video search engines receive benefits from mature text search engine techniques, some nature of video data, such as noisy text transcripts, makes the content-based video search tasks much more difficult than the traditional search tasks of text documents. Therefore, it is clearly not enough to directly apply text-based search engine solutions on video retrieval tasks.

In the past several years, some previous research efforts in content-based video retrieval have shown that the combination of resources from multiple modalities is able to improve the retrieval performance of traditional text-based approaches on video search tasks [3, 4, 5]. Despite promising improvements have been achieved in the past several years, until now, it remains a very challenging task for conducting content-based retrieval on large-scale video databases. Many difficult open issues are not yet tackled. One of the most challenging and essential issues is how to develop an effective learning scheme in combining resources from multiple modalities for ranking search results and balancing the retrieval performance, while achieving computational efficiency for large-scale applications. To attack this challenge, we propose a multimodal and multilevel learning framework for ranking search results effectively, which not only can improve the retrieval performance of traditional approaches, but also can be efficient for large-scale video retrieval tasks.

The rest of this paper is organized as follows. Section 2 presents our multimodal and multilevel ranking framework and describes our methodology in detail. Section 3 discusses our experimental evaluations on TRECVID 2005 dataset. Section 4 sets out our conclusion.

2. MULTIMODAL AND MULTILEVEL RANKING FRAMEWORK

2.1. Overview

In this section we present a multimodal and multilevel learning framework and discuss the engaged methodology. First of all, we describe how to represent videos by graphs. Based on

the graph representations, we suggest to learn harmonic ranking functions over the graphs by Gaussian field and harmonic functions. We then discuss how to fuse multiple resources for ranking through the graphs. Finally, we present the architecture of our multimodal and multilevel learning framework, which is able to reach a good tradeoff between retrieval performance and computational efficiency.

2.2. Video Representation and Graph Based Modeling

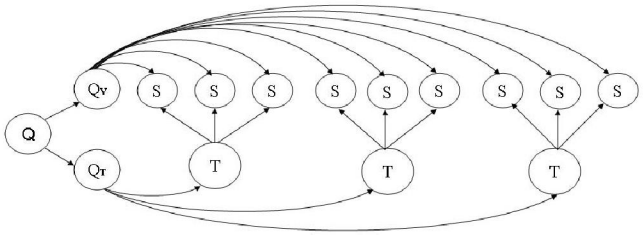
Videos contain rich resources from multiple modalities, including text transcripts from speech recognition and low-level visual contents. In general, a video clip consists of audio channel and visual channel. From the audio channel, text information can be extracted through speech recognition processing. High-level semantic events may also be detected from the audio channel. A video sequence in the visual channel can be regarded as a series of image *frames* presented in a time sequence. Typically, such a video sequence can be represented by a hierarchical structure: video, video *stories*, and video *shots*. A video shot is usually represented by a representative frame, or termed *key frame*, which is selected from frames presented in the shot. A video story is a video scene describing a complete semantic story, which is formed by a series of continuous video shots. In a video search task, a video shot is typically regarded as the basic search unit.

Given the above video structure, for a video search task, we can represent the problem by a graphical model, which can be interpreted as a random walk problem from a probabilistic view [6]. Fig. 1 gives an example to illustrate the idea. Fig. 1(a) shows a set of video stories for retrieval; Fig. 1(b) describes the corresponding graph with respect to a given query topic. The "T" node represents the text content of the video story, while the "S" node represents a video shot. Note that links between "S" nodes are not plotted in the figure for simplicity. Hence, given a query topic formed by texts (Q_T) and visual contents (Q_V), the retrieval task can be regarded as the problem of finding the shots ("S" nodes in the figure) with maximal probabilities on the graph.

2.3. Learning Harmonic Ranking Functions

Based on the graph representation, given a retrieval task, a graph G can be constructed. Then, the video ranking task can be formulated into a learning problem of looking for a smooth function g over the graph. The value of function g on each node is regarded as the relevance score of the node with respect to the query target. From a random walk viewpoint [6], considering a particle starting from a query node, then the value of function g on a searching node can be regarded as the probability that the particle from the starting query node hits the current node. Next, we show how to use the principles of Gaussian field and harmonic function to learn a harmonic ranking function over the constructed graphs [7, 8].

(a) Set of video stories



(b) Graph representation of video structure

Fig. 1. Example of showing graph representation of video retrieval. Note that links between "S" nodes in (b) are not plotted for simplicity.

Let us first consider a graph of single modality. For a search task, assume there are l labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and u unlabeled examples $\{\mathbf{x}_i\}_{i=l+1}^{l+u}$ to be ranked. The label value, y_i is either equal to 1 for a positive example or 0 for a negative one. Let us construct a graph $G = (V, E)$, where vertex set $V = L \cup U$, and L and U are the sets of labeled and unlabeled examples, respectively. We then construct a weight matrix W, which characterizes the data manifold structure. The weight w_{ij} between any two examples $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{R}^d$ can be computed as $w_{ij} = \exp\left(-\sum_{k=1}^{d} \frac{(x_{ik}-x_{jk})^2}{\sigma_k^2}\right)$, where x_{ik} is the k-th component of the example $\mathbf{x}_i$ and σ_k is the length scale parameter of each dimension.

Now the ranking task is equivalent to the problem of assigning a real-valued label to each example in the unlabled set U. Namely, the goal is to learn some real-valued function $g : V \mapsto \mathcal{R}$ on the graph G according to some proper criteria. First, we constrain g to take values $g(\mathbf{x}_i) = g_l(\mathbf{x}_i) = y_i$ on the labeled examples. Then, we look for a function g which is smooth with respect to the constructed graph. To this purpose, we try to find the function g that minimizes the quadratic energy function as follows:

$$g = \arg \min_{g|_L = g_l} \frac{1}{2} \sum_{i,j} w_{ij} \left(g(\mathbf{x}_i) - g(\mathbf{x}_j)\right)^2 \qquad (1)$$

According to the graph theory, the minimum energy function g enjoys the *harmonic* property, which means that the value of g at each unlabeled example is the average of g at the neighboring examples. In order to solve the harmonic function g by matrix operations, we calculate the diagonal matrix $D = \mathrm{diag}(d_i)$, where $d_i = \sum_j w_{ij}$ and W is the weight matrix. Then we let $P = D^{-1}W$ and split the matrices W, D,

and P into four blocks similar to the following structure:

$$W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix} \quad (2)$$

Let us denote $g = \begin{bmatrix} g_l \\ g_u \end{bmatrix}$, where g_u consists of the values of function g on the unlabeled data, which is regarded as the final desirable ranking function. Consequently, the harmonic solution to this final ranking function g_u can be represented by the matrix operations as follows [7, 8]:

$$g_u = (D_{uu} - W_{uu})^{-1} W_{ul} g_l = (I - P_{uu})^{-1} P_{ul} g_l \ . \quad (3)$$

2.4. Multimodal Fusion Through Graphs

Now let us discuss the issue of fusing multimodal resources to learn the ranking functions. The previous graph-based representation enables us to coherently fuse multimodal resources through graphs with proper probabilistic interpretation.

Let us consider the example given in Fig. 1. There are two modalities: text modality and visual modality. If we ignore the text modality, we can directly apply the above harmonic ranking solution to the visual modality. If the text information is included, each "S" node in the graph has two possible channels of hitting the starting query nodes. If we assume the hitting probability from the text channel is given by ρ, then the probability of the other channel will be $1 - \rho$. Here, ρ is also regarded as a combination coefficient of multimodal fusion. Thus, we tackle the multimodal fusion problem by finding the harmonic ranking function h_u on the enhanced graph derived as follows:

$$h_u = (I - (1-\rho)P_{uu})^{-1}((1-\rho)P_{ul} g_l + \rho f_u) \ . \quad (4)$$

2.5. Multilevel Ranking Framework

We have outlined a semi-supervised ranking framework of learning harmonic ranking functions with multimodal resources through graphs. However, for a large-scale problem, directly applying the previous solution on the whole data may not be computationally efficient. To attack this problem, we propose a multilevel ranking framework to achieve a good balance between retrieval performance and computational efficiency. The main idea is to combine multiple ranking strategies of different learning costs in multilevel learning stages. Fig. 2 shows the architecture of our proposed multimodal and multilevel ranking framework. In the first ranking stage, we employ a text-based approach to retrieve the top M ranked video stories associated with a collection of N_1 video shots. In the second stage, combining with visual information, a nearest neighbor (NN) ranking strategy is engaged to retrieve the top N_2 video shots among N_1 shots. In the third stage, a supervised large margin learning method, Support Vector Machine (SVM), is employed to rank on the N_2 shots and output the top N_3 shots. Finally, we apply the semi-supervised ranking method to rerank the top N_4 shots of SVM results and return users top K shots. It is clear that $N_1 \geq N_2 \geq N_3 \geq N_4$.

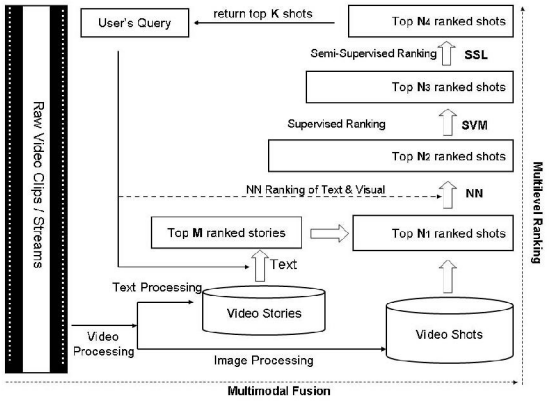


Fig. 2. Architecture of multimodal and multilevel ranking framework.

3. EXPERIMENTAL RESULTS

3.1. Experimental Testbed

We perform experiments on the TRECVID 2005 testbed [9]. The dataset contains 277 broadcast news videos totalling 171 hours from 6 channels in 3 languages (English, Chinese, and Arabic). The automatic speech recognition (ASR) and machine translation (MT) transcripts are provided by NIST [2]. We consider the automatic search task consisting of 24 query topics. Each query contains a text sentence and several image examples [9]. For performance evaluation metrics, we employed non-interpolated average precision (AP) of a single query, and mean average precision (MAP) of multiple queries.

3.2. Textual Processing

Textual information comes from ASR and MT transcripts. The ASR transcripts are all time-stamped at the word level, while the MT transcripts are time-stamped at the sentence level. The text transcripts are segmented in video story level according to some story boundary detection method [10]. Shots within a video story share the same text block. All text stories and queries are parsed by a text parser with a standard list of stop words. The Okapi BM-25 formula is used as the retrieval model together with pseudo-relevance feedback (PRF) for text search.

3.3. Visual Feature Representation

Three types of visual features are used: color, shape and texture. For color, we use Grid Color Moment. Each image is partitioned into 3×3 grids and three types of color moments

are extracted for each grid. Thus, an 81-dimensional color moment is adopted for color.

For shape, we employ an edge direction histogram. A Canny edge detector is used to get the edge image and then the edge direction histogram is computed. Each histogram is quantized into 36 bins of 10 degrees each. An additional bin is used to count the number of pixels without edge information. Hence, a 37-dimensional edge direction histogram is used for shape.

For texture, we adopt Gabor feature. Each image is scaled to 64×64. Gabor wavelet transformation is applied on the scaled image with 5 scale levels and 8 orientations, which results in 40 subimages. For each subimage, three moments are computed: mean, variance, and skewness. Thus, a 120-dimensional feature vector is adopted for texture.

In total, a 238-dimensional feature vector is employed to represent each key frame of video shots.

3.4. Performance Evaluation

We compare our multimodal and multilevel (MMML) ranking scheme with other traditional approaches. For comparison, the text-only approach is used as the baseline method. We also implemented two other ranking approaches: text search with visual reranking by Nearest Neighbor (Text+NN) and visual reranking by Support Vector Machines (Text+SVM). All ranking methods return top 1,000 ranked shots for evaluation.

In our MMML ranking, we first perform text-based ranking to retrieve the top 20,000 shots. Secondly, visual reranking by NN is conducted to retrieve the top 3,000 shots. Thirdly, SVM is used to rerank them and return the top 1,000 shots. Finally, semi-supervised ranking is performed to rerank the top 100 shots of the SVM results.

Table 1 summarizes the comparison of MAP results. First, our implementation of text-based method achieves the MAP result of 0.0902, which is competitive with the best results reported in the TRECVID 2005. For visual reranking methods, we can see that the NN reranking method is able to achieve an improvement of 15.96% over the baseline method. This shows that the set of visual features is quite effective. Further, we found that the reranking method by SVM achieves more significant results than the NN method.

Finally, our MMML solution achieves the best MAP result of 0.1244, which improves the baseline method by 37.92%. Fig. 3 also gives specific evaluation results on all 24 queries of TRECVID 2005. These encouraging results show that our framework is effective and promising for content-based video retrieval tasks.

4. CONCLUSION

In this paper we propose a novel multimodal and multilevel ranking framework for content-based video retrieval. We address several challenges of content-based video retrieval and solve them effectively in our framework. Empirical results have shown that our method is effective and promising for future large-scale content-based video search engines.

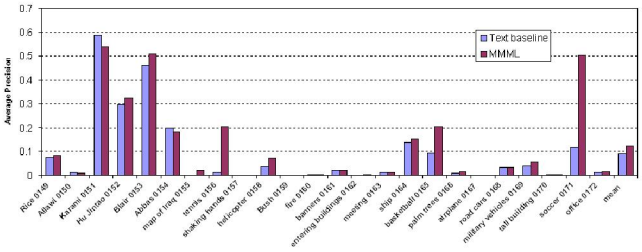


Fig. 3. Evaluation on all 24 queries of TRECVID 2005

Table 1. Comparison of MAP results of different ranking methods

Methods	MAP	Improvement
Text Baseline	0.0902	0%
Text+NN	0.1046	+15.96%
Text+SVM	0.1150	+28.38%
MMML	0.1244	+37.92%

5. ACKNOWLEDGMENT

The work described in this paper was fully supported by a grant from the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong.

6. REFERENCES

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE T. PAMI*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] TRECVID, "TREC video retrieval evaluation," in *http://www-nlpir.nist.gov/projects/trecvid/*.

[3] Rong Yan, Jun Yang, and Alexander G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *Proc. ACM International Conference on Multimedia*, New York, NY, USA, 2004, pp. 548–555.

[4] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. ACM International Conference on Multimedia*, Singapore, 2005, pp. 399–402.

[5] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM International Conference on Multimedia*, Santa Barbara, CA, USA, 2006, pp. 35–44.

[6] P. Doyle and J. Snell, "Random walks and electric networks," *Mathematical Assoc. of America*, 1984.

[7] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proc. ICML*, 2003.

[8] Steven C. H. Hoi and Michael R. Lyu, "A semi-supervised active learning framework for image retrieval," in *Proc. IEEE CVPR*, 2005.

[9] P. Over, W. Kraaij, and A. F. Smeaton, "TRECVID 2005 an overview," in *Proc. TRECVID Workshop*, 2005.

[10] Winston H. Hsu and Shih-Fu Chang, "Visual cue cluster construction via information bottleneck principle and kernel density estimation," in *Proc. CIVR*, Singapore, 2005.