A Multimodal and Multilevel Ranking Scheme for Large-Scale Video Retrieval

Steven C. H. Hoi, *Member, IEEE*, and Michael R. Lyu, *Fellow, IEEE*

***Abstract*—A critical issue of large-scale multimedia retrieval is how to develop an effective framework for ranking the search results. This problem is particularly challenging for content-based video retrieval due to some issues such as short text queries, insufficient sample learning, fusion of multimodal contents, and large-scale learning with huge media data. In this paper, we propose a novel multimodal and multilevel (MMML) ranking framework to attack the challenging ranking problem of content-based video retrieval. We represent the video retrieval task by graphs and suggest a graph based semi-supervised ranking (SSR) scheme, which can learn with small samples effectively and integrate multimodal resources for ranking smoothly. To make the semi-supervised ranking solution practical for large-scale retrieval tasks, we propose a multilevel ranking framework that unifies several different ranking approaches in a cascade fashion. We have conducted empirical evaluations of our proposed solution for automatic search tasks on the benchmark testbed of TRECVID2005. The promising empirical results show that our ranking solutions are effective and very competitive with the state-of-the-art solutions in the TRECVID evaluations.**

***Index Terms*—Content-based video retrieval, graph representation, multilevel ranking, multimodal fusion, multimedia retrieval, semi-supervised ranking, support vector machines.**

I. Introduction

With the rapid growth of digital devices, internet infrastructures, and web technologies, video data nowadays can be easily captured, stored, uploaded, and shared over the Web. Although general search engines have been well developed, searching video content over the Web is still a challenging task. Typically, most Web search engines index only the metadata of videos and search through a text-based approach. However, without the understanding of media content, general search engines have limited capacity of retrieving relevant video information effectively. Thus, there is much scope to improve the retrieval performance of traditional meta-data based search engines through exploiting media content. Content-based video retrieval (CBVR) is becoming a promising direction for developing better video search engines in future [1], [2].

In fact, content-based image/video retrieval is not new for researchers in the multimedia community. In the past decade, content-based image retrieval (CBIR) has been actively studied in signal processing, pattern recognition, and multimedia communities [3]. In recent years, there has been a rapid growth of research attention to content-based video retrieval. Since 2001, the TREC Video Retrieval (TRECVID) evaluation testbed has been set up for conducting benchmark evaluations of video search tasks [4]. Typically, a content-based video search engine can be built upon a traditional text-based search engine using the extracted video texts, such as speech recognition transcripts, closed captions, and video Optical Character Recognition (OCR) text.

Although such video search engines inherit the mature techniques of traditional search engines, some natures of video data make the video search tasks much more difficult than traditional search tasks of text documents. For example, text documents usually contain little noise, while text transcripts of video data are often pretty noisy, as they are usually obtained from automatic speech recognition and text OCR processing. Therefore, it is often inadequate to apply text-based search engine solutions straightforwardly for video retrieval tasks.

In contrast to text data with traditional text search tasks, video data contain more other resources, such as low-level visual content, audio, and high-level visual concepts. A lot of recent research efforts in multimedia retrieval areas have shown that the fusion of information from multiple modalities, including texts and low-level visual content as well as high-level semantic concepts, helps to improve the retrieval performance of traditional text-based approaches in video search tasks [5]–[7]. These approaches often improve the text-only method by leveraging multiple visual query examples for exploiting both textual and visual information effectively.

Despite recent promising improvements, content-based retrieval on large-scale video data is still a very challenging task. There are still a lot of open challenging problems. Among them, one of the most challenging issues is how to develop an efficient ranking scheme of combining resources from multiple modalities to rank searching results effectively toward large-scale retrieval tasks. To this end, we propose a multimodal and multilevel (MMML) ranking framework intended to maximize the effectiveness of retrieval performance and reduce the computational cost of ranking for large-scale video retrieval tasks. We have implemented our solution and evaluated it on the TRECVID benchmark dataset. Our approach, which does

Manuscript received September 25, 2007; revised November 28, 2007. This work was supported in part by University Startup Grant RG67/07 from the School of Computer Engineering, Nanyang Technological University, and in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4150/07E). A short version of this draft was invited to the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2007), in the special session of "Web Image/Video Search Technologies." The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael G. Christel.

S. C. H. Hoi is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: chhoi@ntu.edu.sg).

M. R. Lyu is with the Department of Computer Science and Engineering of the Chinese University of Hong Kong, Hong Kong (e-mail: lyu@cse.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2008.921735

1520-9210/$25.00 © 2008 IEEE

not engage any high-level concept detectors, achieved very promising results, comparing with the best results reported in the TRECVID evaluations. This shows our framework could be a promising solution for building future content-based video search engines on large-scale video data.

The rest of this paper is organized as follows. Section II discusses some state-of-the-art work related to content-based video retrieval in recent years. Section III reviews the problems and challenging issues in content-based video retrieval and gives the motivation of our proposed solution. Section IV presents our multimodal and multilevel ranking framework and the methodologies employed in our solution. Section V discusses the testbed for experimental evaluations and the techniques we employ for feature representation. Section VI shows empirical evaluations of our solutions and implementations on video search tasks in the TRECVID testbed. Section VII discusses some future directions to improve the performance of our current solution. Section VIII sets out our conclusions.

II. Related Work

In this section, we review some existing work related to video search techniques in content-based video retrieval. From the the early 1990s, some institutes have already launched projects linked to digital video libraries for accessing video content intelligently. Some earlier pioneering works include CMU Informedia Digital Video Library projects [1], [8], Columbia VideoQ project of object-oriented search engine [9], and the iVIEW project of Multilingual Digital Video Library at CUHK [2], etc. At that time, research efforts focused more on video processing, such as automatic speech recognition, shot boundary detection, video story segmentation, video object detection, and video summarization, etc. Recently, video retrieval becomes more and more popular along with the surge of Web search engines and the rapid growth of huge volume video data. A variety of video retrieval techniques have been studied in multimedia community recently. We briefly describe some popular ranking approaches for video retrieval, especially for those work related to our proposed learning techniques.

Some early work of applying machine learning techniques for video retrieval may be the (negative) pseudo-relevance feedback methods, which studied SVM techniques to learn ranking functions on visual information to improve traditional text-based retrieval approaches [10], [11]. IBM researchers [12] also studied some variants of SVM techniques together with NN for ranking video shots in TRECVID video search tasks [12]. Both of them reported better results than traditional nearest neighbor search methods thanks to the state-of-the-art performance of discriminative learning techniques. Some other variants of large margin methods have also been studied such as the ranking logistic regression [13]. Most of these work did not explicitly explore the unlabeled data during the learning tasks. Therefore, we categorize these approaches as "supervised ranking," which usually adopt supervised learning techniques.

Recently, researchers have been aware of the importance of unlabeled data for video retrieval tasks. Some recent methods suggested solutions of combining supervised learning and unsupervised learning methods [14], [15]. One limitation of these unsupervised ranking approaches is the high computation cost of the clustering steps, which is critical to large-scale video retrieval tasks. Different from previous work, we suggest the graph based semi-supervised ranking method, and solve its scalability problem through a multilevel ranking scheme. In contrast to other previous approaches, our multimodal and multilevel solution is able to learn the ranking functions on both labeled and unlabeled data more naturally, and integrate multimodal contents for the ranking tasks more effectively. Finally, we are aware that there was some graph-based solutions for multimedia annotation, such as the mexed media graph approach with a linear graph propagation technique [16]. But we argue that a linear solution may be too limited in learning with challenging video retrieval tasks.

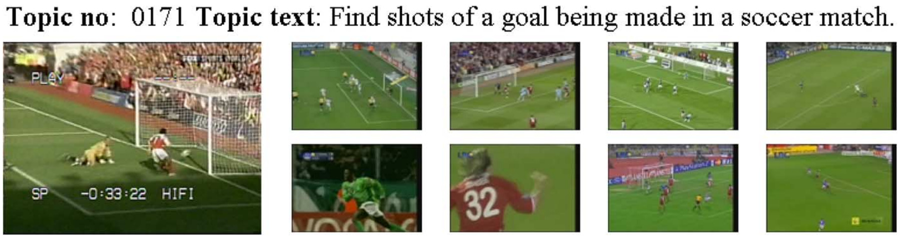
III. Problems and Motivations

In this section, we formally discuss the problem of content-based video retrieval and address several open challenging issues. We then indicate the motivations and philosophy of our approach for solving these problems.

In general, a content-based video retrieval problem can be defined as an information retrieval task of searching relevant video shots from a collection of videos with respect to a query topic, which is formed by some text description and/or a set of visual query samples. For instance, Fig. 1 shows a query example in the TRECVID 2005 benchmark evaluation, which contains a short text sentence and a set of nine image samples. Typically, the collection of video data considered in a video retrieval task has only raw video clips without explicit text information. The implicit text information of the video clips usually can be captured through a preprocessing step. This often involves automatic speech recognition and video OCR processing. However, the quality of the extracted text data is often rather poor due to the long-standing difficulty of pattern recognition on natural images and videos. Even the texts extracted from good quality videos, such as well-structured new videos, can be pretty poor in practice [4]. The nature of content-based video retrieval makes the video retrieval task much more challenging than a traditional information retrieval task. Some of these challenges include the following aspects.

1) Text description of a query topic may be quite short. This poses a challenge for searching video shots by text over the noisy text transcripts extracted from video corpora.
2) Only a small number of positive visual examples will be provided. Collecting many labeled visual examples from users would be expensive in practice.
3) There are a variety of resources from multiple modalities, including text transcripts from video corpora, low-level visual content, and audio content, etc. The combination of various resources is still an open issue.
4) The volume of video collections can be very huge. It is critical to developing a ranking scheme with both excellent retrieval performance and scalability performance toward large-scale applications.

To attack the above challenges, in this paper, we propose a multimodal and multilevel ranking framework for tackling these issues in a unified solution, which can significantly boost the effectiveness of retrieval tasks whilst importantly reducing the



Topic no: 0171 **Topic text**: Find shots of a goal being made in a soccer match.

Fig. 1. Example of a query topic in the TRECVID search task.

computational cost. The main ideas of our solution for tackling these challenges can be summarized as follows.

1) To handle the short query, we alleviate this problem in the text based retrieval stage by engaging pseudo-relevance feedback (PRF) techniques.
2) To solve the small sample learning problem, we suggest the semi-supervised ranking (SSR) method by applying semi-supervised learning techniques, in which we engage the "pseudo" negative examples for the learning task.
3) To combine resources from multiple modalities, we represent a video retrieval task by graphs and learn multimodal functions by fusing multimodal resources smoothly over the graphs.
4) To make the SSR scheme practical for large-scale video retrieval, we propose a multilevel ranking framework of integrating several learning methods in a cascade fashion, which significantly reduces the computational cost meanwhile keeps excellent retrieval performance.

IV. Multimodal and Multilevel Ranking Framework

A. *Overview of Our Framework*

In this section, we present our novel multimodal and multilevel ranking framework and the methodology of our solution. First of all, we show how to represent video structures by graphs, to which a variety of graph based learning techniques can be applied to solve the video ranking problem. Based on the graph representation, we propose a high-level ranking architecture, which implements our multimodal and multilevel ranking framework. We then describe how to learn an effective function for ranking search results by investigating a graph based semi-supervised learning technique. In order to fuse multimodal resources properly, we extend the semi-supervised ranking scheme for learning multimodal ranking function, which can combine other resources over graphs smoothly under a probabilistic view of graph based learning. Finally, we suggest a multilevel ranking scheme of comprising multiple ranking methods of different learning capability and computational cost, which intends to maximize the effectiveness of the learning scheme and improve the computational efficiency of the ranking procedure.

B. *Graph Based Representation for Video Retrieval*

Video data contain rich resources from multiple modalities, including text transcripts from speech recognition and low-level visual content. In general, a video clip consists of an audio channel and a 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 a series of image *frames* presented in a time sequence order. Typically, such a video sequence can be represented by a hierarchical structure: *video*, *video stories*, and *video shots*. A video shot is often represented by one (or more) representative frame(s), termed *key frames*, which are selected from the most important frames in a video 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 retrieval task, a video shot corresponding to some particular representative frame(s) is usually regarded as the basic unit to be ranked and retrieved.

Given the above video structure, we can represent a video retrieval task by graphs, which can be interpreted in probability from a random walk viewpoint [17]. Fig. 2 gives an example to illustrate the idea. Fig. 2(a) shows a set of video stories containing both textual and visual contents for retrieval; Fig. 2(b) describes the corresponding graph with respect to a given query topic "Q". In the figure, the "T" node represents the text content of the video story, while the "S" node represents the visual content of the video shot. In this figure, we assume that the shots within a same video story share the same textual contents. This assumption can be properly extended to other situations. Note that links between "S" nodes are not plotted in the figure and only one visual query node "$\mathrm{Q_V}$" is given for simplicity.

Based on the graph representation, given a query topic, the retrieval task can be regarded as the problem of finding the target shots, i.e., "S" nodes in the figure, with maximal probabilities of being hit by the starting query node in a random walk viewpoint. Specifically, a starting query node "Q," considering two modalities, has two routes for hitting the targets of "S" nodes. One is to go through the path of the "$\mathrm{Q_V}$" node; the other is to go through the path of the "$\mathrm{Q_T}$" node. We will show that this graph based representation is beneficial to multimodal fusion of video ranking in a subsequent section.

C. *Semi-Supervised Ranking Over Graphs*

In this part, we formally present a semi-supervised ranking (SSR) solution based on the above graph representation. Let us first describe the basic idea of our approach. First of all, for a given query topic, we can build a graph G with respect to the query topic. As a result, the video ranking task can be formulated into a graph based learning problem of looking for a smooth ranking function g over the graph. The value of the function g on each node can be regarded as the relevance score of the node with respect to the query. Further, a smooth ranking function enjoys a probabilistic interpretation from a random walk perspective [17]. Let us consider a particle starting from a query node in a random walk behavior. The value of function g on a

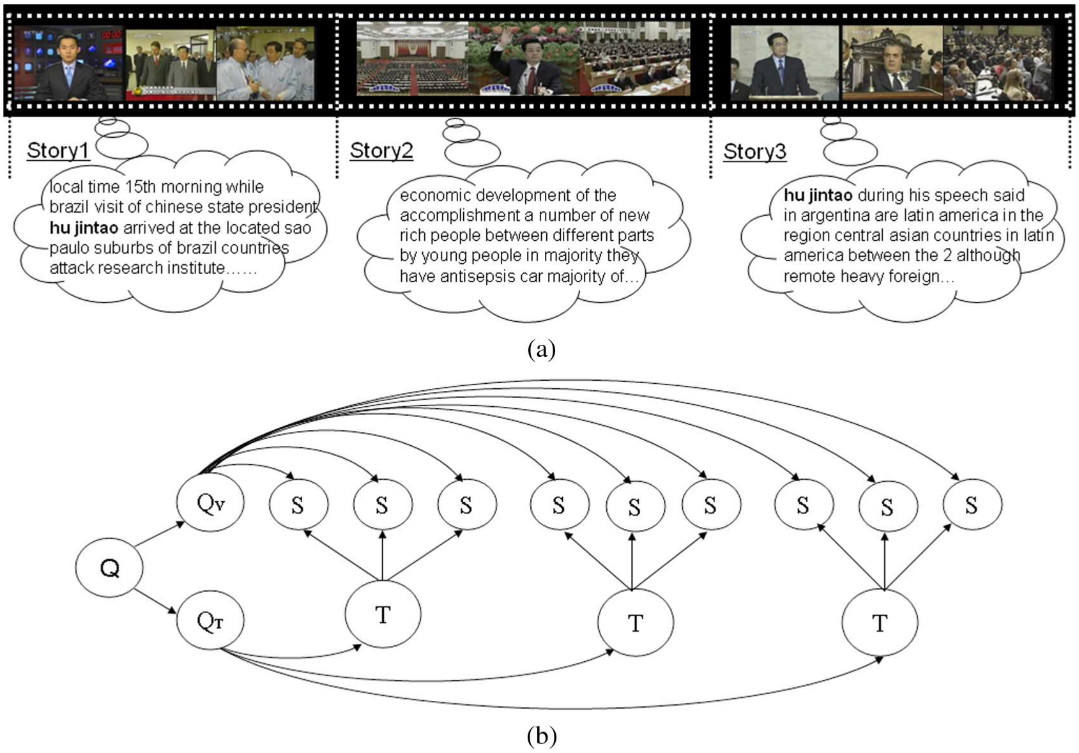

Story1
local time 15th morning while brazil visit of chinese state president **hu jintao** arrived at the located sao paulo suburbs of brazil countries attack research institute……
Story2
economic development of the accomplishment a number of new rich people between different parts by young people in majority they have antisepsis car majority of…
Story3
hu jintao during his speech said in argentina are latin america in the region central asian countries in latin america between the 2 although remote heavy foreign…
(a)
Q
QV
QT
S
T
(b)

Fig. 2. Example illustrating graph representation of video retrieval. (a) List of three video stories; (b) Graph representation of a video retrieval task.

searching node can be regarded as the probability of the particle hits this node from the query. Now the video ranking problem becomes how to learn the smooth ranking function over graphs. Since we can treat the query nodes as the labeled nodes and the searching nodes as the unlabeled nodes, the graph based ranking problem can be turned into an equivalent semi-supervised learning problem over graphs. In this paper, we investigate a graph based semi-supervised learning method in [18] to solve our video ranking problem, which has been shown effective for image retrieval applications [19]. Note that other emerging semi-supervised learning techniques [20], [21] can also be investigated to solve the ranking problem in our proposed framework.

Let us first consider a graph with single modality, i.e., the visual modality. For a video retrieval task, assume that there are a set of l query examples $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, and a set of u video shots $U = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$ to be ranked, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the visual features of a video shot. The relevance value, $y_i \in [0,1]$, is equal to 1 for a positive query example, and 0 for a negative one. Let us construct a graph $G = (V, E)$, where the vertex set $V = L \cup U$ includes the set of query examples L and the set of unlabeled video shots U to be ranked, and the edge set E is the set of pairwise links between any two examples in the vertex set. We then construct a symmetric weight matrix W to characterize the data manifold structure. Specifically, the weight w_{ij} between any two examples $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{R}^d$ is computed below:

$$w_{ij} = \exp\left(-\sum_{k=1}^{d} \frac{(x_{ik} - x_{jk})^2}{\sigma_k^2}\right)$$

where x_{ik} is the k-th dimensional value of the visual example $\mathbf{x}_i$ and σ_k is the length scale parameter for each dimension. In practice, we simply set all of them to a same constant value.

Now the ranking task is equivalent to assigning a relevance value to each example in the set of unlabeled video shots U. Specifically, the goal is to learn some real-valued function $g : V \mapsto [0,1]$ on the graph G according to some criteria. To this purpose, first of all, we constrain g to take values $g(\mathbf{x}_i) = g_l(\mathbf{x}_i) = y_i$ for the examples in the query set. Then, we look for a function g to ensure that is smooth with respect to the constructed graph. Specifically, we consider a quadratic energy function and find the smooth ranking function g by minimizing the quadratic energy as follows:

$$\arg\min_{g} \frac{1}{2}\sum_{i,j} w_{ij} \left(g(\mathbf{x}_i) - g(\mathbf{x}_j)\right)^2 \quad s.t. \quad g(\mathbf{x}_i) = y_i,\ i = 1, \ldots, l. \tag{1}$$

Let $\mathcal{E}(g)$ denote the objective energy function in the above minimization. In order to assign a probability distribution on the function g, we employ the Gaussian random field approach by formulating the Gaussian field as $p_t(f) = (e^{-t\mathcal{E}(f)}/Z_t)$, where t is a well-known "inverse temperature" parameter [17], and Z_t is a normalization factor. According to the theory of random field and harmonic functions [17], it can be shown that the minimum energy function g enjoys the *harmonic* property, i.e., the function g satisfies $\Delta g = 0$ on unlabeled data set U and is equal to g_l on the labeled set L, where Δ is the Laplace operator, i.e., a second order differential operator. The harmonic property means that the value of g at each unlabeled example is the average of g at the neighboring examples, i.e., $g(\mathbf{x}_i) = (1/\sum_j w_{ij}) \sum_{j=l+1}^{l+u} w_{ij} g(\mathbf{x}_j)$. According to the property of harmonic functions, one can show that the optimal function g should be unique for the above optimization.

In order to solve the harmonic function g using matrix operations, we calculate the diagonal matrix $D = \mathrm{diag}(d_i)$, where

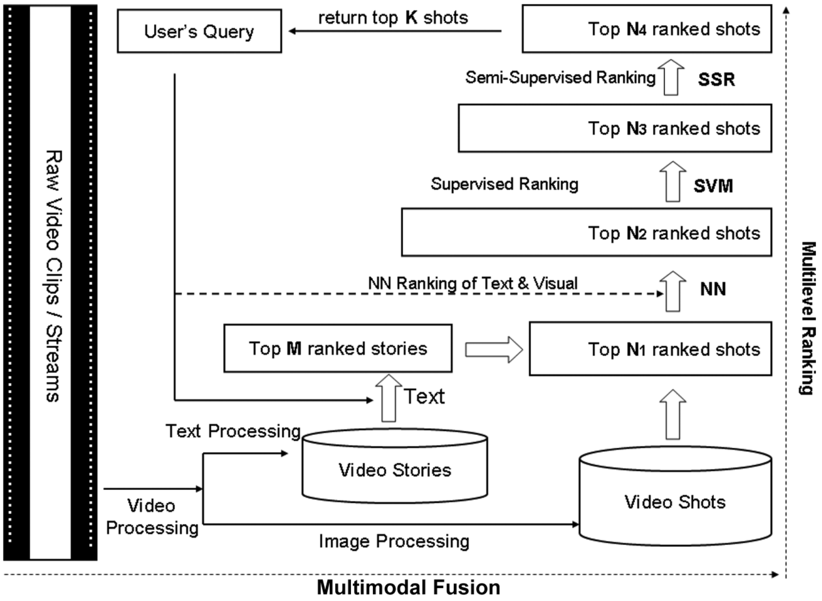


Fig. 3. Multimodal and multilevel ranking architecture.

$d_i = \sum_j w_{ij}$, and W is the weight matrix. Then let $P = D^{-1}W$ and split the matrices W, D, and P into four blocks similar to the following structure:

$$W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}. \tag{2}$$

Let us denote $g = \left[\begin{smallmatrix} g_l \\ g_u \end{smallmatrix}\right]$, where g_u consists of the values of function g on the set of unlabeled video shots U, 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:

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

D. Multimodal Fusion Through Graphs

The previous discussion of ranking data over graphs considers only the visual modality situation. We now investigate how to fuse information from other modalities to learn the ranking functions effectively. Since we have formulated the ranking task as a graph based learning problem above, the multimodal fusion problem can then be naturally handled through the graph based learning approach, which can be properly interpreted in probability from a random walk viewpoint. Specifically, in this section, we describe the method of fusing information from two modalities, i.e., textual and visual, to learn a multimodal ranking function based on the graph based learning principle.

Let us consider the example given in Fig. 2 with two modalities: textual and visual representations. If we first ignore the text modality, we can directly apply the previous graph based ranking method to learn a harmonic ranking function on the visual modality. If text information is included, each "S" node ("*video shot*") of the graph has two possible channels through which it can be hit from the starting query nodes. Let us assume the hitting probability from the text channel is given by $\rho (0 \leq \rho \leq 1)$, the probability of the other channel will be $1 - \rho$. Here, ρ is also regarded as a fusion coefficient of multimodal combination. Thus, we can solve the multimodal fusion problem by seeking the harmonic ranking function g_u on this enhanced graph $G' = (V', E')$, where $V' = V \cup L_T \cup U_T$ is a union of both visual and textual nodes, in which L_T denotes the textual nodes of the query and U_T denotes the textual nodes of the unlabeled video shots. The extended edge set E' only involves the additional links from every textual node to its corresponding shots in the same video story. According to the random walk theory, we can prove that the harmonic ranking function of this multimodal fusion problem given the enhanced graph G' should be:

$$g_u = (I - (1-\rho) P_{uu})^{-1} \left((1-\rho) P_{ul} g_l + \rho f_u^{\mathtt{txt}} \right) \tag{4}$$

where $f_u^{\mathtt{txt}}$ is the ranking function based on the textual modality, which is a known function of probabilistic ranking results. Note that the graph-based fusion is not restricted to two modalities, which can be easily extended for fusing information from other modalities. In the subsequent part, we will discuss its extension of including additional ranking information from a large margin supervised learning stage.

E. Efficient Multilevel Ranking Scheme

We have outlined the semi-supervised ranking framework of learning harmonic ranking functions with multimodal resources through graphs. However, for a large-scale video retrieval problem, directly applying the previous solution on the whole data will be computationally prohibited. To develop an efficient solution, we propose a multilevel ranking framework to learn the ranking functions through multiple learning stages with different computational cost. Fig. 3 shows the architecture of our proposed multimodal and multilevel ranking framework. The basic idea of our solution is to arrive a balance between retrieval performance and computational efficiency by adopting a learning strategy of lower computational cost in a low-level ranking stage and employ more effective learning strategies in a high-level ranking stage. Let us elaborate our framework in detail and discuss the methodology used in each ranking stage as follows.

In general, our multilevel ranking framework consists of four ranking stages: 1) Text-based Ranking, 2) Nearest Neighbour Reranking, 3) Large Margin Supervised Reranking, and 4) Multimodal Semi-Supervised Reranking. We discuss each of them as follows.

1) **Text-based Ranking**. Let us consider a CBVR search task given with both text and visual query samples. For video retrieval tasks, particularly for TRECVID, text based retrieval is usually more effective than a purely visual based approach. Moreover, text based retrieval enjoys the advantage of computational efficiency since efficient indexing techniques have been extensively studied in traditional communities of information retrieval and database. Therefore, we consider text based ranking methods in the first ranking stage. The retrieval performance of text based approaches may suffer significantly for some factors. One is the ill-defined text transcripts of video data, which are usually obtained from Automatic Speech Recognition (ASR) or video OCR techniques. Another is the short query problem, in which a query topic is usually formed by a few keywords or a short sentence. To alleviate these challenges to some extent, we employ pseudo-relevance feedback (PRF) (or query expansion) techniques for overcoming the short query [22]. The main idea is to assume the top k retrieved documents are relevant and then expand the original query using words selected from these top documents. By PRF techniques, we can improve the overall recall rate of text-based retrieval methods. For example, some relevant shots without words from the text content of original query can be retrieved through the PRF approach.
2) **Nearest Neighbour Reranking**. In the second ranking stage, we consider the nearest neighbour (NN) reranking method of combining visual and textual information. This may be one of the most efficient ways. For the textual modality, we employ the normalized ranking scores from the text based ranking stage for computing the ranking scores. For the visual modality, in which data are often represented in vector space, we calculate distances between data examples and query targets for dissimilarity measure using some distance metric. Typically we simply employ the Euclidean distance as the distance measure on the normalized data and use $k = 1$ for the nearest neighbor approach. Consequently, we formulate the ranking function of combining both textual and visual ranking scores as $f_u^{\mathrm{NN}} = \rho f_u^{\mathrm{txt}} + (1 - \rho) f_u^{\mathrm{EU}}$, where f_u^{EU} is the Euclidean distance on visual features and the parameter ρ is the factor to balance the tradeoff of textual and visual features (which is simply fixed to 0.5 in the experiments). The performance may be improved by engaging other advanced distance metrics [23], [24].
3) **Large Margin Supervised Reranking**. In the third ranking stage, we consider a supervised reranking method based on large margin learning techniques [25], which enjoy excellent discriminative performances. In theory, large margin learning balances between the structure risk and the empirical risk (fitting error) via a regularized learning framework:

$$f = \arg \min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_K^2 \tag{5}$$

where $\mathcal{L}$ is some loss function measuring the empirical fitting error, $\|f\|_K^2$ is a norm in a Reproducing Kernel Hilbert Space (RKHS) [25] $\mathcal{H}$ defined over the positive definite kernel function K, and λ is the regularization parameter controlling the tradeoff. The supervised large margin learning methods can be trained efficiently for medium-scale learning tasks. In our solution, we adopt Support Vector Machine (SVM), the most representative large margin method, which has achieved many empirical successes [25]. For a binary classification task, soft margin SVM is often formulated via kernel tricks as follows:

$$\min_{\mathbf{w},\xi,b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{l} \xi_i$$
$$\text{s.t.} \quad y_i (\mathbf{w} \cdot \Phi(\mathbf{x}_i) - b) \geq 1 - \xi_i, \ \xi_i \geq 0, \ i = 1, 2, \cdots, l \tag{6}$$

where C is a regularization parameter, $\Phi(\cdot)$ is a kernel mapping function, such that $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$, and y_i are labels of either $+1$ (positive) or -1 (negative). The primal SVM problem usually can be formulated into a dual form of quadratic program using Lagrange multipliers [25]. By solving the dual, the decision function can be obtained:

$$d_{\mathrm{SVM}}(\mathbf{x}_i) = \sum_{j=1}^{l} y_j \alpha_j k(\mathbf{x}_i, \mathbf{x}_j). \tag{7}$$

For a ranking task, one can simply rank the testing examples based on the SVM distance output values. To further combining the textual information, we can express the ranking function f_u^{LM} in the large margin ranking stage as follows:

$$f_u^{\mathrm{LM}} = \rho f_u^{\mathrm{txt}} + (1 - \rho) f_u^{\mathrm{SVM}} \tag{8}$$

where f_u^{SVM} is a normalized ranking function. To normalize the SVM output to the range of [0,1] effectively, we adopt the following approximation method of estimating the probabilistic output of SVMs:

$$f_u^{\mathtt{SVM}}(\mathbf{x}_i) = P(+|d_{\mathtt{SVM}}(\mathbf{x}_i)) = \frac{1}{1 + \exp(-A \times d_{\mathtt{SVM}}(\mathbf{x}_i) + B)} \tag{9}$$

where $d_{\mathtt{SVM}}(\mathbf{x}_i)$ is the SVM distance output, $A \in \mathbb{R}^+$ and $B \in \mathbb{R}$ are constant parameters. A more comprehensive study for probabilistic output of SVMs can be found in [26].

Finally we would like to remind a practical issue by applying SVMs (and similar machine learning techniques) for solving the video ranking tasks. Typically, we will encounter a barrier, i.e., there is often no negative example provided in a video retrieval task, which is a difficulty by using two-class SVMs. One way to solve this problem is to engage the pseudo negative examples [10], [11]. Specifically, to overcome the difficulty, we can form a set of pseudo negative examples by sampling the examples from the list of the least relevant results obtained from the previous ranking stage.

4) **Multimodal Semi-Supervised Reranking**. In the last ranking stage, we propose the semi-supervised reranking method to learn the ranking functions on both labeled and unlabeled data. For a video retrieval task, there are usually only a limited number of labeled examples. The semi-supervised ranking method is able to exploit the unlabeled data to effectively improve the ranking performance. However, it often incurs much computational cost using semi-supervised learning methods. To avoid much computational cost, we only engage a small port of the the most important unlabeled data by the multilevel ranking principle. As a result, the semi-supervised ranking function in (4) can be computed efficiently. Hence, we can significantly reduce the computational cost of the semi-supervised ranking step whilst without sacrificing the overall retrieval performance much.

While the harmonic ranking function by the semi-supervised learning has been shown with excellent performance for artificial and some UCI datasets [18], in practice, it may suffer from noisy data in real-world applications. To avoid overfitting to the unlabeled data, we suggest to integrate with the prior knowledge learned from the large margin ranking stage. Specifically, we fuse the SVM output into the semi-supervised ranking function based on the graph based fusion principle:

$$\begin{aligned} g_u = &(I - (1-\rho-\lambda)P_{uu})^{-1} \\ &\times \left((1-\rho-\lambda)P_{ul}g_l + \rho f_u^{\texttt{txt}} + \lambda f_u^{\texttt{SVM}}\right) \end{aligned} \quad (10)$$

where $\rho, \lambda \in [0,1]$ are fusion parameters that satisfy the constraint $0 \leq \rho + \lambda \leq 1$. Note that the SVM decision function $f_u^{\texttt{SVM}}$ is a probabilistic function of SVM output.

Remark: It is clear that the multimodal ranking function in (10) reduces to the semi-supervised ranking approach on single modality in (3) when we set $\rho + \lambda = 0$. On the other hand, if we set $\rho + \lambda = 1$, it becomes a traditional supervised ranking solution without engaging any information from unlabeled data.

In summary, we propose a novel multilevel ranking framework to learn multimodal ranking functions efficiently through four ranking stages using different learning strategies. In the first stage, the text-based ranking method yields a set of the top M ranked video stories, which are associated with a set of N_1 top video shots. In the second stage, the NN ranking method reranks the N_1 shots and outputs the top N_2 most relevant video shots. In the third stage, the SVM ranking method reranks the N_2 shots and outputs the top N_3 most relevant video shots. In the last stage, the SSR ranking method reranks the top N_4 shots of N_3 SVM output results. Finally, the multilevel ranking framework returns the top k shots for performance evaluation. It is clear that $N_1 > N_2 > N_3 > N_4$.

F. *Justification and Interpretation*

This multilevel ranking framework is significant to making the proposed SSR solution practical for large-scale applications. In fact, the proposed multimodal ranking solution together with the multilevel ranking scheme, not only significantly reduces the computational cost, but also learns better ranking functions than either the large margin supervised ranking method or the semi-supervised ranking method individually. We justify the effectiveness of our ranking solution properly by a "global filtering/local fitting" learning viewpoint.

Global Filtering: From a geometric viewpoint, a large margin learning method, e.g., SVM, intuitively is to learn the "separating" hyperplane maximizing the margin between the boundaries of positive and negative examples. The margin maximization principle is motivated by the "structure risk minimization" (SRM) theory, which minimizes a bound on the risk over the structure on some set of functions [25]. The SRM theory enables the large margin learning methods to choose the simplest model for fitting the training data. For a retrieval task, the large margin learning method can "globally" filter most irrelevant examples from the relevant set effectively, consequently, we can obtain a smaller set of top ranked examples with a high recall rate.

Local Fitting: However, large margin learning methods may not improve the average precision performance effectively when the number of labeled examples is limited. This problem can be overcome through combining semi-supervised learning methods. In our solution, the principle of our semi-supervised ranking method is to learn harmonic ranking functions for minimizing the fitting error of the approximated function on the training data (including labeled and unlabeled data) through fitting the "local" manifold structure smoothly. Therefore, it can improve the average precision of the ranking results by taking advantage of the local manifold information of both labeled and unlabeled data. Since the semi-supervised ranking may overfit the unlabeled data, we propose to engage a proper regularization term into the multimodal semi-supervised ranking function through fusing the SVM result in (10). The similar idea of regularization has also been studied for semi-supervised learning in the recent machine learning community [21].

In sum, the proposed MMML ranking framework, without incurring much computational cost, can improve the overall retrieval performance by combining the "global filtering" stage with large margin supervised ranking methods and the "local fitting" stage with graph based semi-supervised ranking methods.

V. Experimental Testbed and Feature Representation

A. *Overview of Testbed*

The dataset in our experimental testbed is based on the news video dataset of TRECVID 2005. The total amount of news video data is about 169 hours of videos: 43 in Arabic, 52 in Chinese, and 74 in English. These video data were collected by the Linguistic Data Consortium during November of 2004, and were digitized and transcoded to MPEG-1 format. The search test collection contains 140 video files and 45 765 reference shots. A suite of commercial software was used for automatic

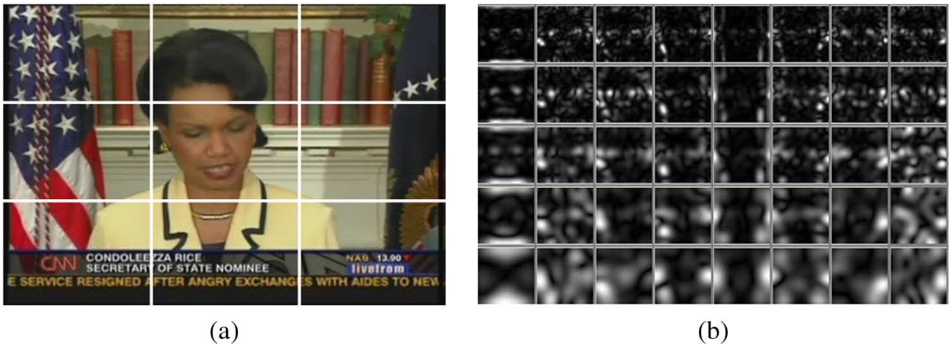

CNN
CONDOLEEZZA RICE
SECRETARY OF STATE NOMINEE
NAS 13.90
livefrom
E SERVICE RESIGNED AFTER ANGRY EXCHANGES WITH AIDES TO NEW
(a)
(b)

Fig. 4. Example of visual feature representation: (a) 3∗3 Grid color moments; (b) 5∗8 Gabor subimages.

speech recognition (ASR) and machine translation (MT) for textual extraction. We consider the automatic search task developed by TRECVID, which consists of 24 query topics [27].

The success of a video retrieval task critically relies on the effectiveness of feature representation of video data. In general, video data contain various resources from multiple modalities, including text, speech, audio, and visual images. In our approach, we mainly focus on two types of information. One is the textual resources extracted from automatic speech recognition. The other is visual information extracted from key frames of video sequences. We discuss how to extract effective features representing these two types of information.

B. Textual Processing

Textual information was extracted using the ASR and MT procedures. The ASR transcripts are all time-stamped at the word level, while the MT transcripts are time-stamped at the sentence level. Since the retrieval unit of a video retrieval task is the video shot, the problem arises of how to relate the text information to the video shots. A possible intuitive solution may be to consider relating text blocks to each video shot by partitioning text blocks at the shot level. This may be somewhat difficult for the transcripts time-stamped at the sentence level. Moreover, we are aware the fact that shots in the same video story are usually more likely to be relevant. Thus, a more reasonable method is to relate the text information to video shots at the story level, i.e., video shots in a single story share the same text information. To this purpose, we adopt a common automatic video story segmentation method to detect the boundaries of video stories.

Once the text stories are obtained, all text stories and query texts are then 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 [22]. In our implementation, the Lemur toolkit was adopt for textual processing and indexing[1].

C. Visual Feature Representation

We extract three kinds of visual features to represent the key frames of video shots, including color, shape, and texture, which have been extensively studied in CBIR [28].

For color, we use color moment. In our approach, we implement a modified color moment, called grid color moment (GCM). Specifically, for each given key frame, we split it into 3×3 equal grids and extract color moments for each of the nine grids. Fig. 4(a) shows a grid based image example. Three types of color moments are computed: mean, variance and skewness in each color channel (H, S, and V), respectively. Thus, an 81-dimensional grid color moment is adopted as the color features for each image.

For shape, we employ an edge direction histogram [29]. To acquire the edge direction histogram, an input color image is first converted into a gray image, and a Canny edge detector is then applied to obtain its edge image. Based on the edge images, the edge direction histogram can then be computed. Each edge direction histogram is quantized into 36 bins of 10 degrees each. In addition, we use a bin to count the number of pixels which do not provide edge information. Hence, a 37-dimensional edge direction histogram is employed to represent the shape features.

For texture, we employ Gabor feature representation [30]. In our approach, each image is first scaled to the size of 64×64. Then, the Gabor wavelet transformation is applied to the scaled image at five scale levels and eight orientations, which results in a total of 40 subimages for each input image. Fig. 4(b) shows an example of Gabor subimages via Gabor transformation. For each subimage, we calculate three types of statistical moments to represent the texture features, including mean, variance, and skewness. Therefore, we use a 120-dimensional feature vector to represent the texture features of each image.

In total, a 238-dimensional feature vector is used to represent the key frame of a video shot.

VI. Experimental Results

A. Overview of Experimental Evaluations

To examine the effectiveness of our solutions, we performed a set of extensive evaluations on video retrieval tasks. In particular, our empirical studies were conducted to address the following questions.

1) How effective is our text retrieval solution? Which retrieval models perform more effective for video retrieval?
2) Can pseudo-relevance feedback (PRF) improve the retrieval performance in solving the short-query issue?
3) Can visual information improve the text based retrieval method? How effective is a regular visual based reranking method based on our extracted features?

[1] http://www.lemurproject.org/

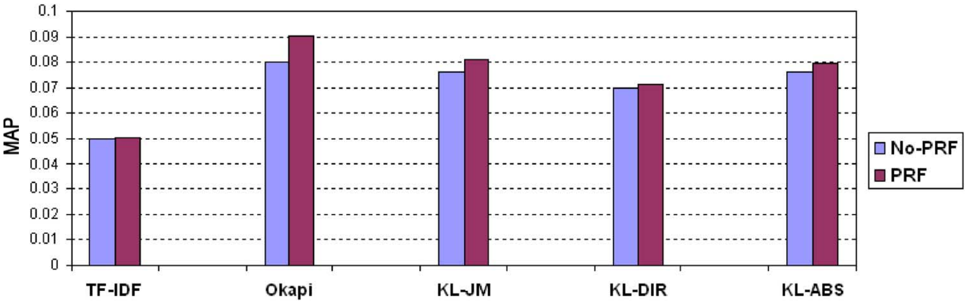


Fig. 5. Evaluation of text-based retrieval methods on TRECVID2005.

4) How effective is the MMML ranking solution? Can it compete with existing methods?
5) How efficient is the MMML ranking solution? Can it achieve a good tradeoff between retrieval performance and computational efficiency?

For performance evaluation, we followed the benchmark evaluation of an automatic search task in TRECVID. The performance metric in our experiments is non-interpolated average precision (AP) for a single query, which corresponds to the area under an ideal (non-interpolated) recall/precision curve. As the AP is only for a single query, we employ the mean average precision (MAP), the mean of average precisions for multiple queries, to evaluate the overall average performance across the set of different queries in our testbed. More details can be found in [4].

B. Performance Evaluation of Text-Based Retrieval

In this part, experiments were conducted to evaluate the performance of text-only retrieval methods. This set of experiments aimed to answer two of the questions set out above: one is to examine which text retrieval model is more effective on the testbed; the other is to evaluate whether the PRF approaches exceed the retrieval performance of typical retrieval methods without PRF. To find an effective text retrieval method, five representative text retrieval approaches were compared in our performance evaluation.

1) *TF-IDF*: the well-know term frequency and inverse document frequency retrieval method [31].
2) *Okapi*: the Okapi BM25 retrieval algorithm [32].
3) *KL-JM*: the Kullback–Leibler (KL) divergence measure using the Jelinek-Mercer smoothing approach [33].
4) *KL-DIR*: the KL-divergence measure with the Bayesian smoothing using Dirichlet priors [22].
5) *KL-ABS*: the KL-divergence measure with the Absolute discounting smoothing [34].

In the experiments, for each of the above method, we compare two variants: one is without PRF and the other is with PRF. Let us now examine their performance on automatic video search tasks in TRECVID 2005. To ensure an objective evaluation of the compared methods, we use a set of default parameters in the smoothing models, which are empirically tuned from traditional document retrieval tasks. Fig. 5 shows the MAP results of our empirical evaluation. Let us first compare the results without PRF. Among the five retrieval methods, the Okapi BM25 and KL divergence based language modeling methods are significantly better than the regular TF-IDF method. This is because the TF-IDF method often suffers from some bias problem of long documents. But the Okapi BM25 method can adjust the weights of terms to balance the bias problem. Among the other retrieval methods, the Okapi BM25 algorithm achieved the best results in our evaluation, though some KL divergence based retrieval methods were reported with better results than the Okapi BM25 method in some traditional text document retrieval tasks [22].

TABLE I
COMPARISON OF FOUR DIFFERENT RANKING APPROACHES IN OUR IMPLEMENTATION

Methods	MAP TOP1000	Prec. TOP10	Prec. TOP15	Prec. TOP20	Prec. TOP30
Text	0.0902 (+0.00%)	0.1333	0.2000	0.1917	0.1847
Text+NN	0.1046 (+15.96%)	0.2583	0.2444	0.2375	0.2097
Text+SVM	0.1171 (+29.82%)	0.3250	0.3000	0.2806	0.2562
MMML	**0.1267** **(+40.47%)**	**0.3708**	**0.3333**	**0.3042**	**0.2681**

Next, let us compare the performance of the retrieval methods with and without PRF. From the average results shown in Fig. 5, we can observe that the retrieval approaches with PRF always outperform the methods without PRF in this evaluation. The most evident case is the Okapi BM25 retrieval method, in which the MAP result was boosted from 8.02% without PRF to 9.02% after using PRF, which is the best retrieval result using text-only retrieval approaches.

C. Performance Evaluation of the MMML Ranking Scheme

The previous experiments have shown the effectiveness of our text retrieval solution for video retrieval tasks. In this part, we will examine the effectiveness of visual reranking scheme and the proposed multimodal and multilevel (MMML) learning framework for ranking in video search tasks. In our implementation, we first employ the previous text based retrieval method to retrieve a set of the top $N_1 = 20000$ ranked video shots. Based on the N_1 video shots, the basic NN ranking method is engaged to rerank the video shots through a linear combination of text ranking scores and visual similarity scores. The text ranking scores are normalized ranking values from the text

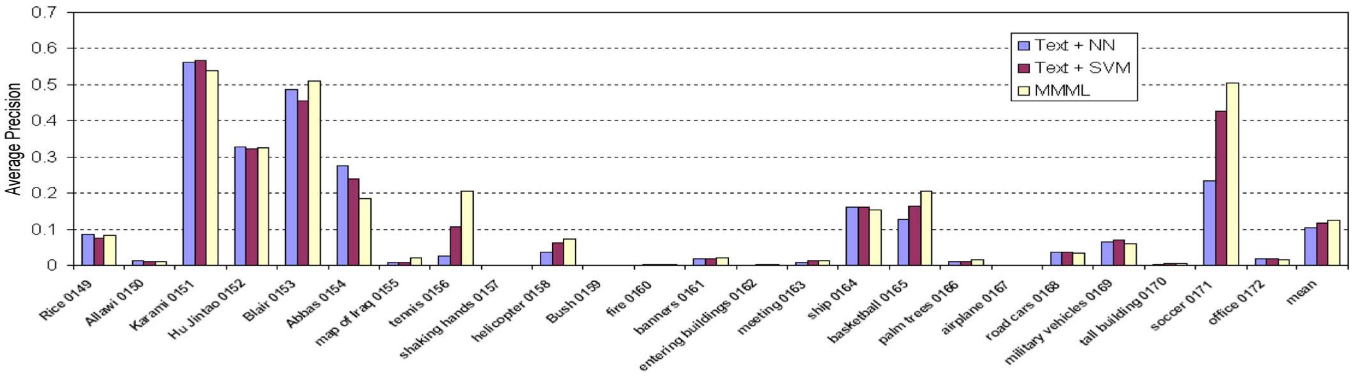


Fig. 6. Evaluation of the proposed multimodal and multilevel ranking scheme on the 24 queries in TRECVID 2005.

based retrieval stage, while the visual similarity scores are normalized scores by measuring Euclidean distances on visual features. The basic NN reranking stage outputs a set of the top $N_2 = 2000$ ranked shots among the N_1 video shots. Next, the SVM learning stage is conducted for reranking the $N_2 = 2000$ video shots, from which the top $N_3 = 1000$ ranked shots are returned. Finally, the semi-supervised ranking stage reranks the top $N_4 = 100$ video shots of the $N_3 = 1000$ SVM results. Finally, 1000 video shots of combined SVM and semi-supervised reranking results are returned for performance evaluation.[2] For comparison, we also implement another multimodal ranking scheme based on SVM learning, which has been regarded as a promising solution in some previous work.

Let us first compare the overall performance of our MMML ranking scheme to other two multimodal approaches, NN and SVM. Table I shows the comparison results. From the results, we can see that all three multimodal approaches are able to achieve significant improvements over the state-of-the-art text baseline retrieval approach. This shows that multimodal solution is successful and promising for video retrieval tasks. Comparing the three multimodal ranking approaches, the NN ranking approach was the worst solution, achieving a MAP improvement of 15.96% over the text baseline approach. The SVM ranking solution obtained a MAP improvement of 29.82% over the text baseline, which is better than the NN one. Finally, our MMML ranking achieved the best improvement of 40.47% over the text baseline.

In the TRECVID benchmark evaluation, top 1000 shots are usually returned to be evaluated in a search task. However, for real-world applications, users may be more interested in obtaining a small number of relevant shots, rather than being offered the top 1000 shots. To examine the performance with a small number of top shots, we evaluate the average precision with TOP ten, 15, 20, and 30 shots. From the results in Table I, we can see that for most of compared methods, the precision of top ranked results decreases when the number of ranked examples increases. The only exceptional case is for the text baseline approach where the precision of TOP 10 results is lower than the TOP 15 results. This shows that a text-based approach without accessing the visual content usually cannot return very precise video shots in the top ranked results. Further, we compare several visual ranking approaches, it is clear that our MMML ranking approach consistently achieved the best results among all the compared schemes. To examine details of the performance on specific queries, we also list the comparison results of all 24 query topics in Fig. 6. Similar observations can be drawn from the results, i.e. our MMML ranking scheme outperforms the other two approaches in most cases though there are a few exceptional cases. For example, for the 0154 query topic, finding shots of “Mahmoud Abbas,” the MMML ranking method is worse than the nearest neighbor and SVM approaches. To figure out the reason, we examine the testbed and find that there are a number of ambiguous examples in the query set, which are not relevant to the query topic. Hence, the MMML ranking approach may be overfitting to the noisy labeled examples. Nonetheless, the average improvement by the MMML solution is still rather significant, which shows the proposed MMML ranking solution is effective at combining resources from multiple modalities for the video retrieval tasks.

Finally, in order to examine whether our proposed scheme is competitive with other approaches, we also compare the average performance of our solution to other existing approaches. Fig. 7 shows the comparison of our results to the results reported by TRECVID participators. From the comparison, we can observe that our MMML solution is among the best of all compared schemes, which represent the state-of-the-art approaches in TRECVID [12], [35]. While the improvements of our solution over the best TRECVID approaches were not very significant, we should address some differences between our solution and other approaches. In our solution, we only consider text and visual information, but some other schemes with best results have included additional high-level features with visual concept detectors. Nonetheless, these results are only presented to show that our MMML approach is, if not better than, competitive to the existing state-of-the-art solutions. We also want to note that all combination coefficients used in our multimodal fusion scheme are simply fixed to default values (0.5) for the two modalities of normalized data without tuning. We believe that better results could be achieved by our solution when including additional information and better parameter selection schemes. From these observations, we can conclude that our solution is effective and promising toward video retrieval tasks.

[2] This number of 1000 is the requirement for benchmark evaluation in TRECVID 2005.

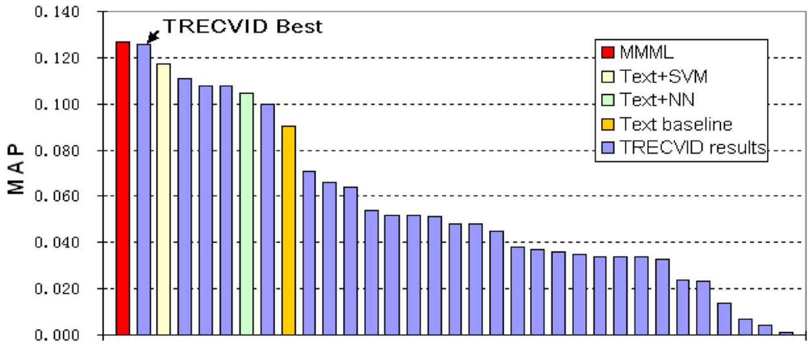


Fig. 7. Comparison of our solutions to other results in TRECVID 2005.

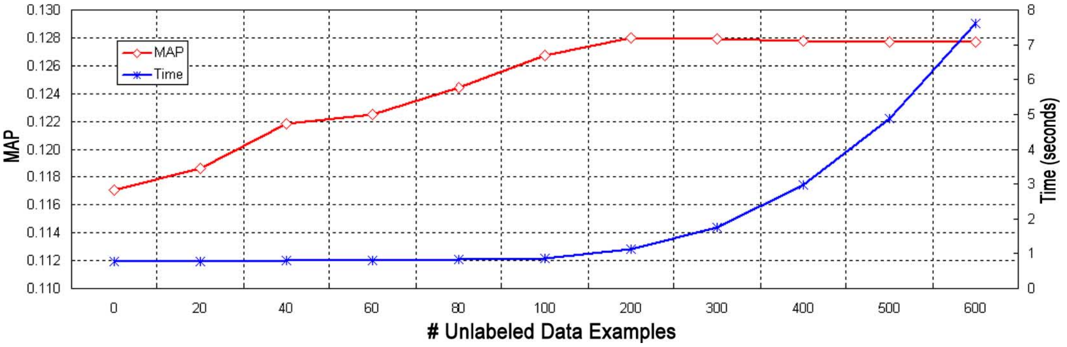


Fig. 8. Evaluation of computational time and retrieval performance with respect to the number of unlabeled examples.

D. Evaluation on Computational Efficiency

In this part, we empirically evaluate the time efficiency of our scheme and investigate its relationship to the retrieval effectiveness. In our multilevel ranking scheme, three ranking stages account for the main computational cost, i.e., the NN ranking, the SVM ranking, and the semi-supervised ranking. For the NN ranking, it is of linear computational complexity, which may be solved more efficiently by adopting some existing indexing techniques [36]. For SVM training, as it is essentially a quadratic program, there are efficient algorithms to solve it with global optima for a medium scale dataset, such as Sequential Minimal Optimization (SMO) techniques [26] whose empirical complexity is about $\mathcal{O}(n^{2.3})$. Since the number of training examples is usually small in a video retrieval task, the SVM training usually can be conducted efficiently. For the semi-supervised ranking, the semi-supervised learning scheme requires the matrix inversion, which results in an $\mathcal{O}(n^3)$ time complexity. Therefore, it is important to choose the number of unlabeled examples N_4 in the SSR stage carefully in order to reduce the overall computational performance of the MMML ranking scheme.

To examine how is the influence of the number of unlabeled examples engaged in the SSR stage with respect to the computational efficiency and the retrieval performance, we conducted experiments to evaluate the performance impact with different values of N_4 in the retrieval tasks. Fig. 8 shows the evaluation results. From the Figure, we can see that when the number of unlabeled examples N_4 was equal to 0, the MMML ranking scheme was reduced to the supervised SVM ranking solution. When N_4 was smaller than 300, the MAP retrieval performance improved when N_4 was increased further. This is because the more unlabeled data are included, the better performance can be achieved by the SSR approach. Specifically, we can see that when N_4 was equal to the default value 100 used in our previous experiments, the MAP result reached 12.67%, which is significantly better than the regular SVM based approach. The best retrieval performance was obtained when N_4 was selected between the range of 200 to 300, at which points the MMML ranking scheme achieved the maximum MAP result of 12.80%. When N_4 was greater than 300, we found some interesting and a bit surprising results: the retrieval performance tended to degrade slightly when the number of unlabeled data was increased beyond the threshold value of 300. This phenomenon can be explained in terms of overfitting and noisy data issues. When large amount of unlabeled data are engaged, much noisy data may be included. Consequently, the performance of the SSR scheme may degrade, caused by overfitting to the local data without proper regularization.

In terms of efficiency, we can see that the computational time consistently increased when the number of unlabeled examples increased. In particular, when the number of unlabeled examples was greater than 200, the computational time dramatically increased. This shows that large-scale semi-supervised ranking would be infeasible without a proper deployment. However, we were able to find that the SSR scheme can be conducted efficiently when the number of unlabeled examples is smaller than 200. Therefore, to ensure an efficient solution, we should consider a number of unlabeled examples smaller than 200. In our previous experiments, our default scheme engaging 100 unlabeled examples can be conducted efficiently, whilst its average retrieval performance is close to the best results we observed. Based on the observations, we can see that our proposed MMML ranking scheme is able to achieve a good balance between the retrieval performance and the computational efficiency.

VII. Future Directions

In this section, we address the limitations of our current approach and point out several future directions.

First of all, we mainly use text and visual information in the ranking tasks. In future, we will include additional information from other modalities. For example, we can study high-level concept detection techniques [12] and investigate some concept models to improve the ranking performance [37], [38]. It is likely that the proposed scheme could be improved by engaging additional information.

Second, in our current implementation, the fusion parameters are simply set to default parameters. In future work, we could study more intelligent solutions to determine the optimal parameters for multimodal fusion.

Third, in our current ranking solution, we consider only the query-independent approach for automatic search tasks. For future work, the query-class dependent weighting methods [37] can also be extended to our solution for further improving the retrieval performance. How to develop an effective query-class dependent algorithm using the graph based ranking framework will be an interesting research issue in future work.

Moreover, we will apply our solution to solving other problems of content-based video retrieval, such as interactive video retrieval [39] and image/video annotation [40]. To these problems, we will explore the proposed multimodal and multilevel framework together with active learning techniques [41], [42] to overcome these open challenges. We may also study more effective kernel learning methods, such as the nonparametric kernel learning for improving the retrieval performance [43].

Lastly, we may study more efficient indexing techniques in our current solution. For large-scale video retrieval applications, determining how to index the data is important, a topic which was excluded in the previous discussion. To enable efficient solution of queries, some emerging indexing techniques, such as Locality-Sensitive Hashing [36] and SVM indexing [44], can be investigated when building an efficient ranking and indexing scheme for large-scale content-based video search engines.

VIII. Conclusions

In this paper we proposed a novel multimodal and multilevel ranking scheme for large-scale video retrieval. The proposed framework not only achieves considerably better retrieval performance than traditional approaches, but also is practically efficient for large-scale applications. The main contributions of this work can be summarized as follows.

First of all, we modeled the ranking problem of video retrieval through graph representation and formulated the retrieval task as a graph based learning problem. The suggested graph based ranking scheme can smoothly fuse a variety of resources from multiple modalities, which enjoys a nice interpretation from the random walk view.

Second, we proposed the semi-supervised ranking (SSR) method to resolve the video retrieval tasks. Different from traditional supervised ranking approaches, the SSR solution is able to exploit both labeled and unlabeled data for the retrieval tasks effectively. The proposed method can be naturally extended to learn a multimodal ranking function by fusing multimodal contents smoothly.

Further, we suggested an efficient multilevel ranking solution to solve the scalability problem of the SSR method. The proposed multilevel ranking scheme not only significantly improves the computational efficiency of our ranking solution, but also avoids the overfitting problem by combining large margin ranking and semi-supervised ranking in a unified scheme. To the best of our knowledge, this may be the first effective semi-supervised ranking scheme applicable for large-scale applications.

Finally, we conducted an extensive set of experiments to empirically evaluate every aspect of the algorithms and techniques proposed in our solution. While our methodology was studied for video retrieval, we expect some ideas of our proposed techniques can be applicable to other similar problems in the field of multimedia, and may offer some valuable insights to other research areas.

References

[1] H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens, "Intelligent access to digital video: Informedia project," *IEEE Computer*, vol. 29, no. 5, pp. 46–53, 1996.

[2] M. R. Lyu, E. Yau, and K. S. Sze, "iview: An intelligent video over internet and wireless access system," in *Proc. 11th Int. World Wide Web Conf. (WWW2002), Practice and Experience Track*, Honolulu, HI, 2002.

[3] 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 Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[4] TRECVID, TREC Video Retrieval Evaluation [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[5] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. 12th ACM Int. Conf. Multimedia*, New York, 2004, pp. 572–579.

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

[7] K.-S. Goh, E. Y. Chang, and W.-C. Lai, "Multimodal concept-dependent active learning for image retrieval," in *Proc. 12th ACM Int. Conf. Multimedia*, New York, 2004, pp. 564–571.

[8] M. Christel, T. Kanade, M. Mauldin, R. Reddy, M. Sirbu, S. Stevens, and H. Wactlar, "Informedia digital video library," *Commun. ACM*, vol. 38, no. 4, pp. 57–58, 1994.

[9] S.-F. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong, "A fully automatic content-based video search engine supporting multi-object spatio-temporal queries," *IEEE Trans. Circuits Syst. Video Technology*, vol. 8, no. 3, pp. 602–615, Mar. 1998.

[10] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. Int. Conf. Image and Video Retrieval (CIVR 2003)*, Urbana-Champaign, IL, 2003, pp. 238–247.

[11] R. Yan, A. Hauptmann, and R. Jin, "Negative pseudo-relevance feedback in content-based video retrieval," in *Proc. ACM Multimedia Conf. (MM 2003)*, Berkeley, CA, 2003.

[12] A. Amir, G. Iyengar, J. Argillander, M. Campbell, A. Haubold, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev, J. R. Smith, J. Tesic, and T. Volkmer, "IBM research trecvid-2005 video retrieval system," in *Proc. TRECVID Workshop*, Washington, DC, 2005.

[13] R. Yan and A. G. Hauptmann, "Efficient margin-based rank learning algorithms for information retrieval," in *Proc. Int. Conf. Image and Video Retrieval (CIVR 2006)*, Tempe, AZ, 2006, ACM Press.

[14] J. Yuan, J. Li, and B. Zhang, "Learning concepts from large scale imbalanced data sets using support cluster machines," in *Proc. 14th ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 441–450.

[15] W. H. Hsu, L. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Multimedia Conf. 2006*, Santa Barbara, CA, 2006.

[16] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," in *Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'04)*, Seattle, WA, 2004, pp. 653–658.

[17] P. Doyle and J. Snell, "Random walks and electric networks," in *Math. Assoc. Amer.*, 1984.

[18] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proc. 20th Int. Conf. Machine Learning (ICML'03)*, Washington, DC, 2003.

[19] S. C. H. Hoi and M. R. Lyu, "A semi-supervised active learning framework for image retrieval," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, CA, 2005.

[20] J. Zhu, Semi-Supervised Learning Literature Survey Carnegie Mellon Univ., 2005, Tech. Rep..

[21] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[22] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inform Syst.*, vol. 22, no. 2, pp. 179–214, 2004.

[23] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'06)*, New York, Jun. 17–22, 2006.

[24] L. Si, R. Jin, S. C. H. Hoi, and M. R. Lyu, "Collaborative image retrieval via regularized metric learning," *ACM Multimedia Syst. J.*, vol. 12, no. 1, pp. 34–44, 2006.

[25] V. N. Vapnik*, Statistical Learning Theory*. Hoboken, NJ: Wiley, 1998.

[26] J. C. Platt, "Probabilities for support vector machines," *Adv. Large Margin Classifiers* 1999.

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

[28] S. C. H. Hoi, M. R. Lyu, and R. Jin, "A unified log-based relevance feedback scheme for image retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 4, pp. 509–524, 2006.

[29] A. K. Jain and A. Vailaya, "Shape-based retrieval: A case study with trademark image database," *Pattern Recognit.*, no. 9, pp. 1369–1390, 1998.

[30] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 8, pp. 837–842, 1996.

[31] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 1998, pp. 275–281.

[32] S. E. Robertson and S. Walker, "On relevance weights with little relevance information," in *Proc. 20th Int. ACM SIGIR Conf. (SIGIR'97)*, 1997, pp. 16–24.

[33] F. Jelinek and E. L. Mercer, "Interpolated estimation of markov source parameters from sparse data," *Pattern Recognit. Practice*, 1980.

[34] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modeling," *Comput. Speech Lang.*, pp. 1–38, 1994.

[35] A. G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang, "CMU informedia's trecvid 2005 skirmishes," in *Proc. TRECVID Workshop*, Washington, DC, 2005.

[36] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Computational Geometry*, New York, 2004, pp. 253–262.

[37] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights for automatic video retrieval," in *Proc. ACM Multimedia Conf. (MM 2004)*, 2004.

[38] S. Dagtas, W. Al-Khatib, A. Ghafoor, and R. Kashyap, "Models for motion-based video indexing and retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 88–101, Jan. 2000.

[39] M. Christel and R. Yan, "Merging storyboard strategies and automatic retrieval for improving interactive video search," in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, Amsterdam, The Netherlands, 2007.

[40] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Image annotation refinement using random walk with restarts," in *Proc. 14th ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 647–650.

[41] S. Tong and E. Y. Chang, "Support vector machine active learning for image retrieval," in *Proc. Ninth ACM Int. Conf. Multimedia (MM'01)*, 2001, pp. 107–118.

[42] S. C.H. Hoi, R. Jin, and M. R. Lyu, "Large-scale text categorization by batch mode active learning," in *Proc. 15th Int. World Wide Web conference (WWW'06)*, CITY?, U.K., May 23–26, 2006.

[43] S. C. Hoi, R. Jin, and M. R. Lyu, "Learning non-parametric kernel matrices from pairwise constraints," in *Proc. 24th Int. Conf. Machine Learning (ICML'07)*, OR, June 20–24, 2007.

[44] N. Panda and E. Y. Chang, "Kdx: An indexer for support vector machines," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 6, pp. 748–763, 2006.

Steven C. H. Hoi (M'06) received the B.S. degree in computer science from Tsinghua University, Beijing, China, and the M.S. and Ph.D. degrees in computer science and engineering from the Chinese University of Hong Kong.

He is currently an Assistant Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include multimedia information retrieval, statistical machine learning, Web search and data mining.

Michael R. Lyu (F'04) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., the M.S. degree in electrical and computer engineering from the University of California at Santa Barbara, and the Ph.D. in computer science from the University of California, Los Angeles, in 1981, 1982 and 1988, respectively.

He is currently a Professor in the Department of Computer Science and Engineering, the Chinese University of Hong Kong. He is also Director of the Video over InternEt and Wireless (VIEW) Technologies Laboratory. His research interests include software reliability engineering, distributed systems, fault-tolerant computing, mobile networks,Web technologies, multimedia information processing, and E-commerce systems. He has published over 250 refereed journal and conference papers in these areas. He was the editor of two book volumes: *Software Fault Tolerance* (New York: Wiley, 1995), and *The Handbook of Software Reliability Engineering* (Piscataway, NJ: IEEE and New York: McGraw-Hill, 1996).

Dr. Lyu received Best Paper Awards in ISSRE'98, and ISSRE2003. He initiated the First International Symposium on Software Reliability Engineering (ISSRE) in 1990. He was the Program Chair for ISSRE'96, and General Chair for ISSRE2001. He was also PRDC'99 Program Co-Chair, WWW10 Program Co-Chair, SRDS2005 Program Co-Chair, and PRDC2005 General Co-Chair, and served in program committees for many other conferences. He served on the Editorial Board of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and has been an Associate Editor of IEEE TRANSACTIONS ON RELIABILITY, *Journal of Information Science and Engineering*, and Wiley's *Software Testing, Verification & Reliability Journal*. Dr. Lyu is an IEEE and AAAS Fellow for his contributions to software reliability engineering, and software fault tolerance.