IEICE TRANS. INF. & SYST., VOL.E101–D, NO.5 MAY 2018

PAPER

Graph-Based Video Search Reranking with Local and Global Consistency Analysis

Soh YOSHIDA[†a)], **Takahiro OGAWA**[††b)], **Miki HASEYAMA**[††c)], ***Members***, ***and*** **Mitsuji MUNEYASU**[†d)], ***Senior Member***

SUMMARY Video reranking is an effective way for improving the retrieval performance of text-based video search engines. This paper proposes a graph-based Web video search reranking method with local and global consistency analysis. Generally, the graph-based reranking approach constructs a graph whose nodes and edges respectively correspond to videos and their pairwise similarities. A lot of reranking methods are built based on a scheme which regularizes the smoothness of pairwise relevance scores between adjacent nodes with regard to a user's query. However, since the overall consistency is measured by aggregating only the local consistency over each pair, errors in score estimation increase when noisy samples are included within query-relevant videos' neighbors. To deal with the noisy samples, the proposed method leverages the global consistency of the graph structure, which is different from the conventional methods. Specifically, in order to detect this consistency, the propose method introduces a spectral clustering algorithm which can detect video groups, in which videos have strong semantic correlation, on the graph. Furthermore, a new regularization term, which smooths ranking scores within the same group, is introduced to the reranking framework. Since the score regularization is performed by both local and global aspects simultaneously, the accurate score estimation becomes feasible. Experimental results obtained by applying the proposed method to a real-world video collection show its effectiveness.

key words: *video search reranking, graph learning, graph consistency analysis, spectral clustering*

1. Introduction

With the explosive growth of social media, a great number of videos are being generated and shared on the Internet. For example, YouTube has over a billion users and people watch hundreds of millions of hours every day*. Thus, many techniques have been developed for multimedia searches. Owing to the success of information retrieval businesses, such as Google, Bing, and Yahoo!, most search engines employ text-based techniques by using nonvisual information such as surrounding text and user-provided tags, associated with visual content. However, since textual information is sometimes noisy or unavailable, the inconsistency between textual features and visual contents can cause poor image/video search results [1], [2].

To improve the text-based search performance and overcome the semantic gap between text information and video contents, visual search reranking has been the focus of attention in recent years [3]–[9]. This technique adjusts the initial ranking orders by mining visual content or leveraging some auxiliary knowledge. Most reranking methods have been developed on the basis of the following three assumptions: (1) visual contents with dominant patterns are expected to be ranked higher than others, (2) visual contents with similar visual appearance are to be ranked closely, and (3) top-ranked contents in initial search results are expected to be ranked relatively higher than the others. Under these assumptions, visual information is introduced to refine the initial search result.

A lot of reranking methods are formulated as finding the optimal ranked list from the perspective of Bayesian theory [10], [11] and manifold discovery [12], [13]. These reranking approach assumes that relevant multimedia documents such as images and videos lie on a manifold in visual feature space. Then the reranking is accomplished by graph-based learning methods. Therefore, we call it graph-based reranking. Generally, the approach constructs a graph, where the nodes are multimedia documents and the edges reflect their pairwise similarities. The initial relevance of each document can be viewed as the stationary probability of each node and can be transitioned to other similar nodes until some convergence conditions are satisfied. This graph representation of search results can be integrated into a regularization framework by considering the following two terms: a graph regularizer that keeps the ranking positions of visually similar documents close and a loss term insuring that the reranked results do not change too much from the initial ranking list.

Although many different methods have been proposed, the visual consistency between similar video contents is not always guaranteed due to the complexity of real-world video contents. Then, in several cases, search performance may be even degraded after the reranking. This is because that most graph-based methods measure visual consistency pairwisely. The overall consistency is measured by aggregating the local consistency over each pair. Thus, errors in score estimation increase when noisy samples are included in each pair. To solve this problem, we introduce the idea

Manuscript received September 1, 2017.
Manuscript revised December 20, 2017.
Manuscript publicized January 30, 2018.

†The authors are with Kansai University, Suita-shi, 564–8680 Japan.
††The authors are with Hokkaido University, Sapporo-shi, 060–0814 Japan.
a) E-mail: sohy@kansai-u.ac.jp
b) E-mail: ogawa@lmd.ist.hokudai.ac.jp
c) E-mail: miki@ist.hokudai.ac.jp
d) E-mail: muneyasu@kansai-u.ac.jp

DOI: 10.1587/transinf.2017EDP7277

*https://www.youtube.com/yt/press/statistics.html

of social network analysis. Specifically, community detection methods have attracted great research interests in the past years [14]. A community consists of a group of nodes that are densely connected to each other but sparsely connected to other dense groups. Since a community structure in networks usually reveals the common topic or interest, the consistency over an area among a same community means a video group whose videos have strong correlation with its neighbor. We call it global consistency. However, these consistency analysis is not considered for improving performance of video search reranking. Therefore, it is desirable to develop a novel algorithm that regularizes graph consistency based on both local and global aspects, simultaneously.

In this paper, we propose a novel graph-based reranking with local and global consistency analysis. We adopt the following two procedures: (A) detection of the global consistency over the graph and (B) modeling of the graph-based reranking considering both local and global consistency.

First, in (A), we detect the global consistency by adopting a spectral clustering algorithm [15] to the constructed graph. Given a similarity graph, a spectral clustering algorithm finds a partition of the set of its nodes into clusters. This algorithm satisfies the following hold: nodes in different clusters are dissimilar to each other, which aims to minimize the between-cluster similarities; and nodes in the same cluster are similar to each other, which aims to maximize the within-cluster similarities. From the clustering result, we extract center nodes corresponding to representative nodes of each cluster. Then we define a new affinity matrix representing the similarity between center nodes and the similarity between nodes among the same video group.

In (B), we model reranking using the graph and the affinity matrix which reflects global consistency over the graph. Our reranking model is built based on a Bayesian formulation [10] and its multimodal expansion [9]. In this paper, we introduce a new graph regularizer that smooths the ranking scores among the same video group obtained by the procedure in (A). For a video, instead of calculating the consistency with each of its neighbors individually, the proposed regularizer considers the consistency with all of videos among the same group simultaneously. By using this term with the previous regularization framework, the proposed method can suppress the influence of noisy videos. Furthermore, it is difficult to assign the appropriate parameters to the two types of affinity matrices. In order to integrate two aspects, we introduce the graph-learning approach and tune these parameters automatically. Finally, by minimizing the objective function including three terms, i.e., graph local and global regularizer terms and a loss term, the desired consistency over the graph is guaranteed. Therefore, performance improvement by the graph-based reranking becomes feasible.

The contribution of this work is summarized as follows:

1) We propose a graph global consistency detection approach for video search reranking. This enables integration of global consistency analysis into a graph-based regularization framework.
2) The proposed method simultaneously regularizes the smoothness of the ranking scores between not only adjacent nodes but also nodes among the same video group. This approach enables suppression of the influence of noisy videos' score propagation.

This paper is an extended version of [16]. In this paper, the following three aspects are enhanced. 1) In order to improve the robustness of the algorithm for obtaining the affinity matrix of each aspect, we introduce the graph-based learning approach in our method. By using this approach, tuning parameters for determining the scale of the affinity matrix are automatically learned. 2) We complement discussions of parameters to set manually. 3) We collect a Web video search dataset using 15 queries and study the effectiveness of the proposed method by comparing it with various conventional graph-based reranking methods.

The remainder of this paper is organized as follows. In Sect. 2, we review the related work on the visual search reranking for image and video retrieval. Section 3 presents the proposed method, which retrieves videos using a graph-based reranking framework with local and global consistency analysis. Section 4 provides experimental results that verify the performance of the proposed method. Finally, Sect. 5 presents concluding remarks.

2. Related Work

2.1 Visual Search Reranking

Visual search reranking has been widely investigated for improving the search performance of images, videos and other multimedia documents. The existing visual search reranking efforts can be mainly classified into two categories according to whether there are query examples available, which are called example-based reranking and self-reranking.

For the first category, these methods need several examples in addition to a text-query. Yan et al. [3] regard the query examples as relevant samples and several bottom-ranked results in a ranking list as irrelevant ones. A Support Vector Machine (SVM) model is then learned based on these samples to rerank the search results. Natsev et al. [4] improve the robustness of this example-based approach by a bagging strategy. They collect multiple irrelevant sample sets and then generate different ranking lists accordingly. These ranking lists are aggregated to generate the final reranked result. Liu et al. [5] use the query examples to discover the relevant and irrelevant concepts for a given query and identify an optimal set of document pairs by using an information theory. A ranking list is then directly recovered from this pair set. These methods can improve search performance if good visual examples are provided. However, these methods cannot be used in the cases when there is no visual example available.

For the second category, the self-reranking approach does not rely on query examples. It aims to improve text-based search by mining the visual information of images or videos. In many cases, we can assume that the top-ranked documents are the few "relevant" (called pseudo relevant) documents that can be viewed as "positive". This is in contrast to relevance feedback where users explicitly provide feedback by labeling the results as positive or negative. Kennedy et al. [6] regard top and bottom-ranked results in a ranking list as pseudo relevant and irrelevant samples respectively to discover the related concepts. The detection results of the related concepts are then used as high-level features in SVM to build classifiers for reranking. Hsu et al. [17] formulate the reranking process as a random walk over a context graph, where videos are nodes and the edges between them are weighted by multimodal similarities. Jing et al. [8] apply the PageRank [18] to product image search and design the VisualRank algorithm for reranking. After a similarity-based image link graph is generated, an iterative computation similar to PageRank is utilized to rerank the images. Yang et al. [19] extract multiple features from each image and collect a training set that contains several queries and labeled search results. Reranking is then regarded as a supervised learning task. Tian et al. [10] model the textual and visual information from the probabilistic perspective and formulate visual reranking as an optimization problem in the Bayesian framework, named Bayesian visual reranking. This method encodes the assumptions that the reranked results do not change much from the initial ranking list and the ranking positions of visually similar images are close.

However, its fundamental deficiency lies in the noise, i.e., it is not guaranteed that the irrelevant instances are always apart from the top returns, which would push away true positive after reranking in many cases. In this work, to perform robust visual reranking in this kind of situation, we investigate video search reranking with local and global consistency analysis based on community detection approach. By learning the adaptive similarity weights of each aspect, we will show that our approach can effectively integrate two aspects to boost ranking performance.

2.2 Graph-Based Learning

Graph-based learning has been introduced into visual reranking in the past year. One major advantage of graph-based learning is to encode the data structure into the data similarity measurement to refine inference and modeling. In these methods, a graph is constructed based on the given data, where nodes and edges respectively correspond to samples and their pairwise similarities. They are usually formulated in a regularization scheme with two terms. One term is used to enforce the function to be smooth on the graph, and the other term is used to keep the function consistent with prior information such as the labeling information of several samples. The algorithms can be accomplished by a random walk process.

He et al. [12] adopt a graph-based method named manifold-ranking in image retrieval. Wang et al. [9] developed a multi-graph learning approach to fuse multiple feature channels based on semi-supervised learning. In [20], multiple graphs from different retrieval methods are fused by summing up the edge weights, and then a graph alignment is conducted to build an overall similarity graph. In [8], [10], [21], the initial ranking list is refined on the graph by propagating the ranking scores through the edges.

Unfortunately, the regularization term used in these methods measures the graph consistency pairwisely. Specifically, the overall consistency is measured by aggregating the local consistency over each pair. The consistency on the graph is multiplewise instead of pairwise since it is a term defined over the whole neighboring samples. Therefore, the consistency approximated through pairwise regularizers is not satisfactory enough. Our method is inspired by [9], [14]. Our approach first detects the global consistency of the overall graph. By using the multimodal graph learning method, we then fuse the two types of graphs and then estimate an optimal relevance score with regard to the user's query.

3. Graph-Based Video Search Reranking with Consistency Analysis

In this section, we describe our proposed reranking approach. We first introduce the existing graph-based reranking methods with a general regularization scheme. We then present our approach including consistency analysis and new graph regularization. For clarity, the notations and definitions throughout this paper are summarized in Table 1.

3.1 Graph-Based Reranking with Local Regularizer

We first follow [10] to define several terms in reranking. Let $\bar{\mathbf{r}} = [\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_N]^T$ and $\mathbf{r} = [r_1, r_2, \ldots, r_N]^T$ denote vectors of the initial ranking scores and the relevance scores, which correspond to the video set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. $\bar{r}_i$ and r_i are the initial ranking scores, which are calculated from the ranking position by keyword search, and the relevance scores with regard to the user's query. We also use

Table 1 Notation table.

Notation	Definition
$\mathcal{X}$, $\mathbf{x}_i$	The Video set and ith video in a ranking list.
$\bar{\mathbf{r}}$, $\bar{r}_i$	The vector of the initial ranking scores and the score of $\mathbf{x}_i$.
$\mathbf{r}$, r_i	The vector of the relevance scores and the score of $\mathbf{x}_i$.
L, G	Indicators for local and global aspects.
$\mathbf{W}_\bullet$	The affinity matrix of videos.
$\mathbf{A}_\bullet$	The transformation matrix including the affinity matrix.
$\mathbf{L}_\bullet$, $\tilde{\mathbf{L}}_\bullet$	The graph Laplacian and the normalized graph Laplacian derived from $\mathbf{W}_\bullet$.
$\mathbf{D}_\bullet$	The degree matrix derived from $\mathbf{W}_\bullet$.
$\mathbf{O}$	The centroids of spectral clustering.
$\mathbf{C}$	The node set which corresponds to each centroid.
$\alpha_\bullet$, ρ	Tuning parameters.
N	The number of videos.
K	The number of clusters for spectral clustering.
T, T_1	The iteration time in the alternating optimization.

$\mathbf{x}_i$ to denote its feature vector. In this paper, three kinds of visual features and one kind of audio feature are adopted (described in 4.1).

Generally, graph-based reranking can be formulated as a regularization framework. The objective function is then defined as:

$$\arg\min_{\mathbf{r}} Q(\mathbf{r}) = R(\mathbf{r}, \mathbf{W}) + \rho L(\mathbf{r}, \bar{\mathbf{r}}), \quad (1)$$

where the first part is a regularization term that makes the ranking scores of visually similar videos close, the second part is a loss term that estimates the difference between $\mathbf{r}$ and $\bar{\mathbf{r}}$, and ρ is a trade-off parameter. As the term $R(\mathbf{r}, \mathbf{W})$, a graph $\mathcal{G}$ is constructed with nodes being the videos and similar videos are linked by edges. Then graph Laplacian [22] and normalized graph Laplacian [23] can be widely utilized. When constructing the graph $\mathcal{G}$, each video is connected with its k-nearest neighbors [10]. $\mathbf{W}$ is an affinity matrix in which W_{ij} indicates the visual similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. In this paper, we use $\mathbf{W}_L$ and $\mathbf{W}_G$ as the affinity matrices for local and global aspects, respectively. For the local aspect, if two videos $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected as the edge, the similarity W^L_{ij} is calculated based on the Gaussian kernel with the scaling parameter σ_L. Otherwise, two videos are not connected $W^L_{ij} = 0$. We define the affinity matrix $\mathbf{W}_L \in \mathbb{R}^{N\times N}$ by taking W^L_{ij} as its (i, j)th element. Through minimizing the objective function $Q(\mathbf{r})$, the optimum ranking score list $\mathbf{r}^*$ can be derived as $\mathbf{r}^* = \arg\min_{\mathbf{r}} Q(\mathbf{r})$ using the local regularizer $R(\mathbf{r}, \mathbf{W}_L)$.

3.2 Global Consistency Detection

This subsection shows how to detect global consistency by using a spectral clustering algorithm [15]. In this paper, global consistency means that videos on the same video group structure, typically referred to as a cluster, are likely to have a high similarity. Since this structure in the graph usually reveals the common topic or interest, the consistency over a local area within the same graph means that each sample has strong correlation with its neighbor. Thus, if we can deduce a sample's score in its neighbors precisely, it is regarded that this sample is locally consistent.

Spectral clustering unveils the video group structure by exploiting the eigen-structure of the graph Laplacian matrix $\mathbf{L}_L$, where $\mathbf{L}_L = \mathbf{D}_L - \mathbf{W}_L$ and $\mathbf{D}_L$ is a diagonal matrix and its (i, i)th element is the sum of ith row of $\mathbf{W}_L$. Let $\mathbf{U}$ consist of the unit-length eigenvectors which are associated with the K smallest eigenvalues of $\mathbf{L}_L$, namely $\mathbf{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$, which is a K-dimensional embedding of the graph. The information of each node is therefore captured by a point in $\mathbb{R}^K$. In order to discover the video group structure, k-means clustering is applied to the rows of $\mathbf{U}$ and returns the video group labels $\mathbf{z} = \{z_1, \ldots, z_N\} \in \{1, \ldots, K\}$ and K centroids $\mathbf{O} = \{\mu_1, \ldots, \mu_K\}$. Then we detect nodes $\mathbf{C} = \{c_1, \ldots, c_K\}$, which correspond to each centroid $\mathbf{O}$ and are called center nodes. A spectral clustering algorithm is provided in Algorithm 1 with the input being the affinity matrix $\mathbf{W}_L$ and the pre-specified number of groups K. Its outputs are the estimated labels $\mathbf{z}$ and the center nodes $\mathbf{C}$.

Algorithm 1 Global consistency detection using a spectral clustering algorithm

Input: The affinity matrix $\mathbf{W}_L$ of the video graph $\mathcal{G}$ and K
Output: Label set $\mathbf{z}$ and center nodes $\mathbf{C}$

procedure GLOBALCONSISTENCYDETECTION($\mathbf{W}$, K)
  $d^L_{ii} \leftarrow \sum_{j=1}^{N} W^L_{ij}$
  $\mathbf{D}_L \leftarrow \mathrm{diag}\{d_{11}, \ldots, d_{NN}\}$
  $\mathbf{L}_L \leftarrow \mathbf{D}_L - \mathbf{W}_L$
  $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\} \leftarrow$ unit-length eigenvectors of $\mathbf{L}_L$ which are associated with the K smallest eigenvalues of $\mathbf{L}_L$
  $\mathbf{U} \leftarrow \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$
  Cluster labels for all nodes and centroids of K groups $(\mathbf{z}, \mathbf{O}) \leftarrow$ results of k-means clustering on the rows of $\mathbf{U}$ with K centres
  $\{c_1, \ldots, c_K\} \leftarrow$ nodes corresponding to each centroid $\mathbf{O} = \{\mu_1, \ldots, \mu_K\}$
  $\mathbf{C} \leftarrow \{c_1, \ldots, c_K\}$
  return $(\mathbf{z}, \mathbf{C})$
end procedure

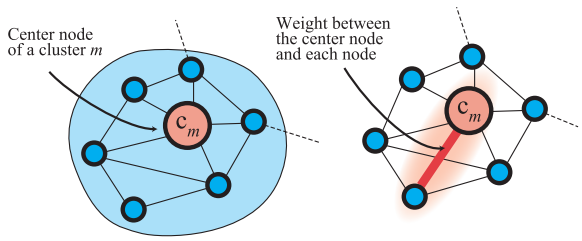

Fig. 1 Center node detection and similarity definition based on shortest path problem.

The goal of our reranking is to regularize smoothness of the ranking scores between not only adjacent nodes but nodes among the same video group simultaneously. Therefore, we define a new weight W^G_{ij}, which represents the similarity between each node and its center node among the same video group. As shown in Fig. 1, if two videos $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same label $\mathbf{z}$ and $\mathbf{x}_j \in \mathbf{C}$, we connect them by an edge and calculate its weight W^G_{ij}. We define the affinity matrix $\mathbf{W}_G \in \mathbb{R}^{N\times N}$ by taking W^G_{ij} as its (i, j)th element. By using the affinity matrix $\mathbf{W}_G$, we formulate the reranking problem.

3.3 Proposed Graph-Based Reranking Algorithm

We develop our approach based on normalized graph Laplacian and ranking distance. Typically, the similarity of kth aspect ($k \in \{L, G\}$) between ith and jth videos is firstly defined as $W^k_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2_k)$, where σ_k is the scaling parameter of the Gaussian function that converts distance to similarity. However, Euclidean distance may not be appropriate as the most suitable distance metric [24]. Therefore, we replace the Euclidean distance metric with the following Mahalanobis distance metric, which can be learned an optimization framework:

$$W^k_{ij} = \exp\left(-(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M}_k (\mathbf{x}_i - \mathbf{x}_j)\right), \quad (2)$$

where $\mathbf{M}_k$ is a symmetric positive semi-define real matrix. We decompose $\mathbf{M}_k$ as $\mathbf{M}_k = \mathbf{A}_k^T\mathbf{A}_k$, where $\mathbf{A}_k \in \mathbb{R}^{d\times d}$ and is substituted it into Eq. (2) as

$$W_{ij}^k = \exp\left(-\|\mathbf{A}_k(\mathbf{x}_i - \mathbf{x}_j)\|^2\right). \tag{3}$$

This is equivalent to transform each video $\mathbf{x}_i$ to $\mathbf{A}_k\mathbf{x}_i$. For the initialization, we set $\mathbf{A}_k$ to a diagonal matrix $\mathbf{I}/\sigma_k$, where σ_k is the median value of the pairwise Euclidean distance of the videos in the kth aspect.

The proposed method considers local and global aspects in the graph. Here, we linearly combine the normalized graph Laplacian regularizers. Mathematically, in order to smooth reranking scores based on both global and local consistencies, we model the reguralizer term so as to combine local and global terms as follows:

$$\begin{aligned} R(\mathbf{r}, \mathbf{A}_L, \mathbf{A}_G) &= \sum_{k\in\{L,G\}}\sum_{i,j}\alpha_k W_{ij}^k\left(\frac{r_i}{\sqrt{d_{ii}^k}} - \frac{r_j}{\sqrt{d_{jj}^k}}\right)^2 \\ &= \sum_{k\in\{L,G\}}\alpha_k\mathbf{r}^T\tilde{\mathbf{L}}_k\mathbf{r}, \end{aligned} \tag{4}$$

where α_k is the weight for local and global regularizers. The weights satisfy $0 \le \alpha_k \le 1$ and $\alpha_L + \alpha_G = 1$. d_{ii}^k is the sum of the ith row of $\mathbf{W}_k$, $\tilde{\mathbf{L}}_k = \mathbf{I} - \mathbf{D}_k^{-1/2}\mathbf{W}_k\mathbf{D}_k^{-1/2}$ is the normalized graph Laplacian, and $\mathbf{D}_k$ is the diagonal matrix whose (i,i)th element is d_{ii}^k.

Accordingly, our algorithm can be formulated as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{r},\mathbf{A}_L,\mathbf{A}_G} Q(\mathbf{r}, \mathbf{A}_L, \mathbf{A}_G) &= \sum_{k\in\{L,G\}}\sum_{i,j}\alpha_k W_{ij}^k\left(\frac{r_i}{\sqrt{d_{ii}^k}} - \frac{r_j}{\sqrt{d_{jj}^k}}\right)^2 \\ &\quad + \rho\sum_{i,j\in S_{\bar{\mathbf{r}}}}\left(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j}\right)^2, \end{aligned} \tag{5}$$

where the loss term indicates the preference strength ranking distance [10] and $S_{\bar{\mathbf{r}}}$ is the set of pairs (i,j) whose relevance scores of all the sample-pairs $(\mathbf{x}_i, \mathbf{x}_j)$ satisfy $\bar{r}_i > \bar{r}_j$. Note that an appropriate scale of $\mathbf{A}_k$ for estimating $\mathbf{W}_k$ will also be automatically determined. The scaling parameter is usually very sensitive for graph-based learning, and it needs to be carefully tuned. The elimination of the parameter by automatically determining the scale of $\mathbf{A}_k$ is also an important element of our approach.

3.4 Alternating Optimization

The formulation shown in Eq. (5) is a minimization problem involving two variables to optimize. Since this objective is not convex, it is difficult to simultaneously recover both unknowns. However, if we hold one unknown constant and solve the objective for the other, we have two convex problems that can be optimally solved. In the rest of this section, we introduce an alternating optimization for our reranking framework, which iterates between the updates of $\mathbf{r}$ and $\mathbf{A}_k$.

3.4.1 Update for $\mathbf{r}$

By using the form of normalized graph Laplacian, we can rewrite Eq. (5) as follows:

$$\begin{aligned} &Q(\mathbf{r}, \mathbf{A}_L, \mathbf{A}_G) = \\ &\sum_{k\in\{L,G\}}\alpha_k\mathbf{r}^T\tilde{\mathbf{L}}_k\mathbf{r} + \rho\sum_{i,j\in S_{\bar{\mathbf{r}}}}\left(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j}\right)^2, \end{aligned} \tag{6}$$

If the transformation matrices $\mathbf{A}_L$ and $\mathbf{A}_G$ are constant, then denote $\beta_{ij} = 1/(\bar{r}_i - \bar{r}_j)$ and the relevance score list $\mathbf{r}$ can be updated by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{r}} Q(\mathbf{r}) \\ &= \min_{\mathbf{r}}\sum_{k\in\{L,G\}}\alpha_k\mathbf{r}^T\tilde{\mathbf{L}}_k\mathbf{r} + \rho\sum_{i,j\in S_{\bar{\mathbf{r}}}}\left\{1 - \beta_{ij}(\bar{r}_i - \bar{r}_j)\right\}^2 \\ &= \min_{\mathbf{r}}\sum_{k\in\{L,G\}}\alpha_k\mathbf{r}^T\tilde{\mathbf{L}}_k\mathbf{r} + \rho(\mathbf{r}^T\mathbf{L}^{(B)} - 2\mathbf{B}\mathbf{e})\mathbf{r}, \end{aligned} \tag{7}$$

where $\mathbf{L}^{(B)}$ is a graph Laplacian matrix defined over the graph $\mathcal{G}_B$ which has the same structure of $\mathcal{G}$ regarding the weight between nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ as $|\beta_{ij}|$. $\mathbf{B} = [\beta_{ij}]_{N\times N}$ is an anti-symmetric matrix, and $\mathbf{e}$ is a vector with all elements equal to 1.

Finally, the relevance score list $\mathbf{r}$ is derived by differentiating w.r.t $\mathbf{r}$ and equating it to zero as follows:

$$\mathbf{r} = \left(\sum_k \alpha_k\tilde{\mathbf{L}}_k + \rho\mathbf{L}^{(B)}\right)^{-1}\tilde{\rho}, \tag{8}$$

where $\tilde{\boldsymbol{\rho}} = 2\rho(\mathbf{B}\mathbf{e})$. It can be seen that different from the normalized graph Laplacian based learning, the two types of normalized graph Laplacian matrices have been linearly combined with weights α_k.

3.4.2 Update for $\mathbf{A}_k$

Now, we consider the optimization of $\mathbf{A}_k$ $(k = L, G)$. Since the optimization of both $\mathbf{A}_L$ and $\mathbf{A}_G$ is the same process, we describe that of $\mathbf{A}_L$ as an example. Considering $\mathbf{r}$ and $\mathbf{A}_G$ are fixed, we then derive the derivative of Q with respect to $\mathbf{A}_L$ as follows:

$$\begin{aligned} &\frac{\partial}{\partial\mathbf{A}_L}Q(\mathbf{A}_L, \mathbf{A}_G) \\ &= \alpha_L\frac{\partial}{\partial\mathbf{A}_L}\sum_{i,j}W_{ij}^L\left(\frac{r_i}{\sqrt{d_{ii}^L}} - \frac{r_j}{\sqrt{d_{jj}^L}}\right)^2 \\ &= \alpha_L\sum_{i,j}(h_{ij}^L)^2\frac{\partial W_{ij}^L}{\partial\mathbf{A}_L} \\ &\quad - W_{ij}^L h_{ij}^L\left(\frac{r_i}{\sqrt{(d_{ii}^L)^3}}\frac{\partial d_{ii}^L}{\partial\mathbf{A}_L} - \frac{r_j}{\sqrt{(d_{jj}^L)^3}}\frac{\partial d_{jj}^L}{\partial\mathbf{A}_L}\right), \end{aligned} \tag{9}$$

Algorithm 2 Gradient descent process for solving $\mathbf{A}_k$.

Input: Step-size parameter $\eta_t = 1$.
Output: The transformation matrix $\mathbf{A}_k$.
1: Set $\mathbf{A}_k^{(0)}$ to a diagonal matrix $\mathbf{I}/\sigma_k$, where σ_k is the median value of the pairwise Euclidean distances of the videos in the kth aspect.
2: **for** $t = 1$ to T_1 **do**
3: Let $\mathbf{A}_k^{(t+1)} = \mathbf{A}_k^{(t)} - \eta_t \frac{\partial Q}{\partial \mathbf{A}_k}|_{\mathbf{A}_k = \mathbf{A}_k^{(t)}}$.
4: **if** $Q(\mathbf{A}_k^{(t+1)}) < Q(\mathbf{A}_k^{(t)})$ **then**
5: $\eta_{t+1} = 2\eta_t$;
6: **else**
7: $\mathbf{A}_k^{(t+1)} = \mathbf{A}_k^{(t)}$, $\eta_{t+1} = \eta_t/2$.
8: **end if**
9: **end for**

Algorithm 3 Optimization process of the reranking algorithm

Input: Tuning parameters α_k and trade-off parameter ρ. The affinity matrices $\mathbf{W}_L^{(0)}, \mathbf{W}_G^{(0)}$ for initialization.
Output: The relevance score list $\mathbf{r}$.
1: Set $A_L^{(0)}, A_G^{(0)}$ to diagonal matrices $\frac{\mathbf{I}}{\sigma_L}, \frac{\mathbf{I}}{\sigma_G}$, respectively, where $\sigma_\bullet$ is the median value of the pairwise Euclidean distances of the videos in each aspect.
2: **for** $t = 1$ to T **do**
3: Compute the tth optimal relevance score list $\mathbf{r}^{(t)}$ according to Eq. (8).
4: Update tth transformation matrices $\mathbf{A}_L^{(t+1)}$ and $\mathbf{A}_G^{(t+1)}$ sequentially according to Algorithm 2.
5: Update the similarity matrices $\mathbf{W}_k^{(t+1)}$ as Eq. (3).
6: **end for**

where

$$h_{ij}^L = \frac{r_i}{\sqrt{d_{ii}^L}} - \frac{r_j}{\sqrt{d_{jj}^L}},$$
$$\frac{\partial W_{ij}^L}{\partial \mathbf{A}_L} = -2W_{ij}^L \mathbf{A}_L (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j), \quad (10)$$
$$\frac{\partial d_{ii}^L}{\partial \mathbf{A}_L} = \sum_{j=1}^{N} \frac{\partial W_{ij}^L}{\partial \mathbf{A}_L}.$$

In order to solve the optimization of $\mathbf{A}_L$ using Eq. (9), we adopt a gradient descent process. In the gradient descent process, we dynamically adapt the step-size in order to accelerate the process while guaranteeing its convergence. Denote $\mathbf{A}_L^{(t)}$ as a result of $\mathbf{A}_L$ in tth turn of the iterative process. If $Q(\mathbf{A}_L^{(t+1)}, \mathbf{A}_G) < Q(\mathbf{A}_L^{(t)}, \mathbf{A}_G)$, i.e., the cost function obtained after the gradient descent is reduced, we double the step-size. Otherwise, we decrease the step-size and do not update $\mathbf{A}_L$. The process is shown in Algorithm 2. In this process, we denote $Q(\mathbf{A}_L)$ as the value of the object function when entering $\mathbf{A}_L$. After the iteration of $\mathbf{A}_L$, $\mathbf{r}$ and $\mathbf{A}_L$ are fixed, and $\mathbf{A}_G$ is calculated by the same way as $\mathbf{A}_L$.

The whole alternating optimization process is illustrated in Algorithm 3. After the alternating optimization, the proposed method returns videos in accordance with the optimal relevance score $\mathbf{r}$ as the video searching result.

Table 2 15 Event Queries.

Queries
UEFA EURO 2016 highlights
Sochi 2014 Winter Olympics opening ceremony
World Figure Skating Championships 2016
Rio 2016 Summer Olympics games
NBA Finals 2014 highlights
CCTV new years gala 2016
Speech at Apec China 2014 speech
2014 Hong Kong protests
2014 Israel Gaza Conflict air strikes
Malaysia Airlines Flight 17 crash moment
New York Fashion Week 2014 runway
November 2015 Paris attacks
Flood in Indonesia 2014
Calbuco Volcano Eruption in Chile
Italy earthquake 2016

4. Experimental Results

In this section, we verify the effectiveness of our proposed method. We first describe the datasets collected from YouTube[†] and the measurements in the experiments. We then analyze the performance of our method of video search reranking.

4.1 Datasets and Features

Datasets: While the research on video search has recently received intensive attention, the public datasets do not reflect current social event topics. To substantially evaluate our approach, we collected a new dataset with rank information from YouTube for video search reranking. Specifically, the used videos were crawled from YouTube by using 15 event queries as shown in Table 2. There is a MSRA-MM (Microsoft Research Asia Multimedia) dataset [25] as a well-known dataset for video search. In this task, 9 categories of videos are searched. Therefore, we use 15 queries in the experiments. These queries cover current topics of news from events, which were selected by reference to the categories "Categories:2014, 2015, and 2016" from Wikipedia[††]. For each query, we obtained max top-500 videos, and analyzed the related videos of each video by using YouTube API[†††]. Furthermore, the associated contextual information such as tags, titles and descriptions were also crawled together with videos. This dataset is a real-world Web video dataset containing the original ranking information. By using these videos, we construct the video graph $\mathcal{G}$. When constructing the graph $\mathcal{G}$, each sample is connected with its k-nearest neighbors. The neighborhood size is set to 5. For the iteration times T and T_1, we set them to 5 and 10, respectively. Our method assigns the initial score $\bar{r}_i = 1 - \hat{r}_i/N$, where $\hat{r}_i$ is the rank of video $\mathbf{x}_i$ returned by the search engine.

[†]http://www.youtube.com
[††]http://en.wikipedia.org/wiki/Category:2014, 2015, and 2016
[†††]https://developers.google.com/youtube/v3/

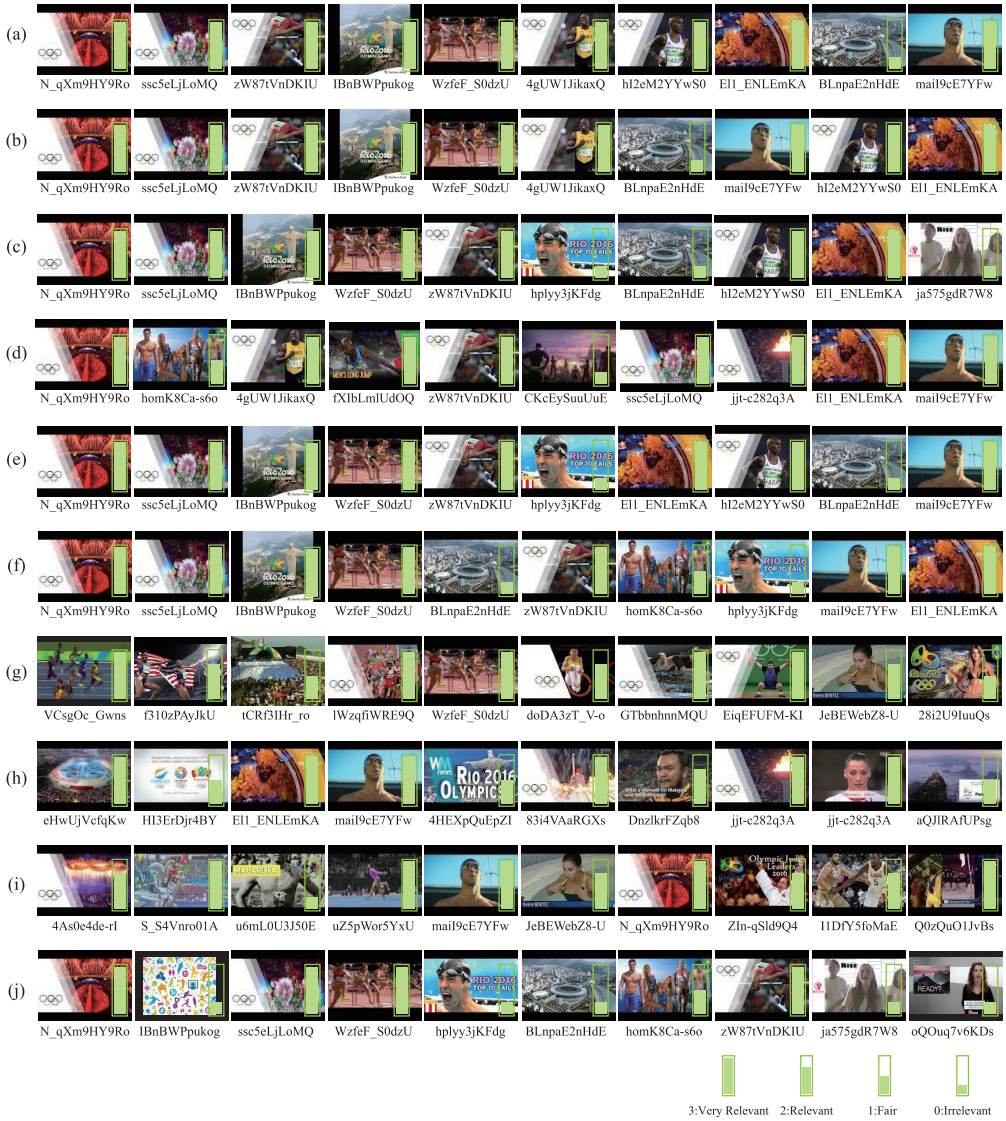

Fig. 2 Visual results of the video reranking from different approaches of the specific query (Rio 2016 Summer Olympics Games): (a) Ours, (b) Ours (without $\mathbf{A}_k$ optimization), (c) Ours ($\alpha_G = 0$), (d) SocialRank, (e) MGL, (f) Bayesian, (g) VisualRank, (h) RandomWalk, (i) BM25, (j) Initial (No method). Note that the corresponding YouTube IDs are shown below the images.

Features: For query videos, we extract the following the sequential features from the whole videos and the frame-level visual and audio features from keyframes. Note that we denote I-frames of the MPEG-4 video as the keyframes.

C3D: We apply the C3D model [26] pre-trained on the *Sports 1M* dataset to compute representations with 512 dimensions.

Inception-v3: We apply the Inception-V3 model [27] pre-trained on the *ImageNet 1K* classification task to compute representations with 2048 dimensions.

HSV Color histogram: We use the HSV color histogram to exploit the color information. To contain spatial information, keyframes are divided into 25 blocks of the same size. A 1600-dimensional HSV normalized color histogram of each region with 4 bins in each color space is extracted.

MFCC: Mel-frequency cepstral coefficients (MFCC), which describe the short-time spectral shape of audio frames, are extracted to capture audio information. MFCC are widely used not only for speech recognition but also for generic audio classification. ΔMFCC, ΔΔMFCC, log-power, Δlog-power and ΔΔlog-power are extracted in addition to the MFCC. The dimension of the audio feature is 39 including 12-dimensional MFCC.

These sequential and frame-level visual and audio features are combined by early fusion followed by PCA to reduce the dimension to 256. The video-level feature $\mathbf{x}_i$ of the ith video is mean-pooled from frame-level features.

4.2 Evaluation Metrics

The performance evaluation of our method is voted by eight volunteers who are invited to assign the relevance scores for top N videos of each query. The averaged relevance score is used to measure the retrieval results.

The performance is measured by the widely used average precision (**AP**), which averages the precision obtained when each relevant video occurs. We average the APs over all the 15 queries to obtain the mean AP (**MAP**) as an overall performance measurement. Then, to measure the video search performance, the normalized discounted cumulative gain (**NDCG**) [28], which is commonly used measure in information retrieval when there are more than two relevance levels, is adopted. For a given query, the NDCG score at position d in the ranking list is calculated as follows:

$$\mathrm{NDCG@}d = Z_d \sum_{j=1}^{d} \frac{2^{t^j} - 1}{\log(1+j)}, \tag{11}$$

where t^j is the degree of the jth video in the ranking list and Z_d is a normalization constant chosen to guarantee that NDCG@d is 1 for a perfect ranking. For each video, in the experiments, the relevance degree t^j was judged manually on four scales: "0:Irrelevant", "1:Fair", "2:Relevant", and "3:Very Relevant". To evaluate the overall performance, we average the NDCGs over all queries to obtain the mean NDCG (**MNDCG**).

4.3 Reranking Results

To evaluate the performance of the proposed reranking algorithm, we first compare the proposed method with the following eight reranking methods:

1) No method, i.e., the initial search results without reranking. This method is denoted as "Initial".
2) The text-based search results based on the Okapi BM-25 formula [29] using the associated contextual information of each video. The method is denoted as "BM25".
3) The random walk method proposed in [17]. The method is denoted as "RandomWalk".
4) Graph-based reranking proposed in [8]. The method is denoted as "VisualRank".
5) Bayesian reranking proposed in [10]. The method is denoted as "Bayesian".
6) Multimodal graph-based reranking proposed in [9], which is the state-of-the-art for graph-based reranking. The method is denoted as "MGL".
7) Social ranking proposed in [30]. User information is utilized to boost the retrieval performance. A regularization framework which fuses the visual and views information is introduced. The method is denoted as "SocialRank".
8) The proposed reranking method without the global regularizer. That means we fix $\alpha_G = 0$. The method is denoted as "Ours ($\alpha_G = 0$)"
9) The proposed reranking method with assigning equivalent scaling parameters to two aspects. That means we

Table 3 MAP comparison of video reranking performance.

Methods	MAP
Ours	**0.735**
Ours (without $\mathbf{A}_k$ optimization)	0.729
Ours ($\alpha_G = 0$)	0.684
SocialRank	0.731
MGL	0.727
Bayesian	0.717
VisualRank	0.635
RandomWalk	0.529
BM25	0.583
Initial (No method)	0.597

Table 5 p values of the significance test comparison. The performance measure is MAP.

Methods	p values
versus Ours (without $\mathbf{A}_k$ optimization)	3.16×10^{-3}
versus Ours ($\alpha_G = 0$)	5.75×10^{-4}
versus SocialRank	2.65×10^{-2}
versus MGL	1.05×10^{-3}
versus Bayesian	6.25×10^{-4}
versus VisualRank	1.94×10^{-7}
versus RandomWalk	1.78×10^{-7}
versus BM25	1.15×10^{-5}
versus Initial (No method)	1.45×10^{-6}

Table 4 MNDCG@d comparison of the video reranking performance.

Methods	@5	@10	@20	@30	@40	@50	@60	@70	@80	@90	@100
Ours	**0.901**	**0.895**	**0.837**	**0.825**	0.811	**0.792**	**0.762**	**0.751**	**0.749**	**0.751**	**0.746**
Ours (without $\mathbf{A}_k$ optimization)	0.893	0.881	0.832	0.824	0.804	0.791	0.751	0.735	0.733	0.731	0.733
Ours ($\alpha_G = 0$)	0.806	0.781	0.773	0.754	0.744	0.733	0.721	0.701	0.699	0.685	0.679
SocialRank	0.899	0.894	0.831	0.823	**0.813**	0.789	0.742	0.743	0.741	0.739	0.729
MGL	0.871	0.850	0.845	0.796	0.785	0.778	0.761	0.749	0.738	0.751	0.733
Bayesian	0.850	0.811	0.786	0.759	0.753	0.742	0.738	0.735	0.730	0.723	0.722
VisualRank	0.671	0.645	0.647	0.661	0.659	0.651	0.656	0.652	0.649	0.644	0.651
RandomWalk	0.581	0.578	0.575	0.581	0.582	0.573	0.571	0.586	0.587	0.579	0.590
BM25	0.682	0.679	0.659	0.641	0.638	0.642	0.633	0.628	0.625	0.632	0.631
Initial (No method)	0.677	0.663	0.668	0.668	0.661	0.665	0.665	0.654	0.654	0.652	0.653

Table 6 MNDCG@d comparison of the video reranking performance when the initial search results obtained by each query equally contain 80% of noisy samples.

Methods	@5	@10	@20	@30	@40	@50	@60	@70	@80	@90	@100
Ours	0.766	**0.741**	**0.715**	0.664	0.605	**0.607**	**0.587**	**0.609**	**0.635**	**0.638**	**0.656**
Ours (without $\mathbf{A}_k$ optimization)	0.741	0.729	0.690	0.663	0.602	0.577	0.575	0.586	0.623	0.627	0.644
Ours ($\alpha_G = 0$)	0.702	0.664	0.698	0.629	0.600	0.592	0.556	0.529	0.524	0.534	0.522
SocialRank	0.748	0.714	0.696	0.641	0.589	0.586	0.584	0.595	0.632	0.635	0.653
MGL	**0.767**	0.731	0.698	**0.677**	**0.621**	0.587	0.585	0.596	0.633	0.636	0.654
Bayesian	0.739	0.689	0.698	0.643	0.591	0.577	0.575	0.586	0.623	0.627	0.644
VisualRank	0.448	0.538	0.536	0.563	0.596	0.552	0.572	0.591	0.586	0.585	0.579
RandomWalk	0.258	0.457	0.451	0.482	0.518	0.476	0.460	0.480	0.476	0.475	0.465
BM25	0.364	0.317	0.395	0.417	0.377	0.384	0.360	0.375	0.393	0.402	0.417
Initial	0.652	0.634	0.611	0.629	0.579	0.592	0.556	0.529	0.525	0.534	0.523

fix $\mathbf{A}_k = 1/\sigma_k$. The method is the same as [16] and denoted as "Ours (without $\mathbf{A}_k$ optimization)"

For fair comparison, the comparisons 3) - 7) were implemented by using the same video-level features as shown in Sect. 4.1.

Figure 2 shows the top results with comparisons between the proposed method and other methods for an example query "Rio 2016 Summer Olympics games". It is obvious that our approach is superior to all compared methods owing to our capability to rank the relevance videos by using multiple types of objects and multiple types of relationships. The results of the MAP comparison are shown in Table 3. It can be seen that the proposed reranking algorithm has a better performance than the other methods. This demonstrates the robustness of our algorithm.

Next, we show the video retrieval results obtained by using the proposed method and the other retrieval methods. Table 4 demonstrates the MNDCG@5,10,20,30,40,50,60,70,80,90,100 of different methods. Overall, our proposed graph-based reranking outperforms the other methods, and the improvements are consistent and stable at different depths of NDCG. Especially, using the proposed method, the value of the MNDCG@100 shows an improvement of 0.017 and 0.013 over SocialRank and MGL, which are the state-of-the-art methods in reranking, respectively.

To verify whether the improvement of the proposed method is statistically significant, we further perform a statistical significance test. Here, we conduct paired T-test at the 5% significance level between ours and all other methods. The p values are shown in Table 5. The T-test is conducted over 15 queries. From this result, we can see that the improvement of the proposed method is statistically significant.

Table 6 shows the simulation results to verify the robustness to noisy videos. It is observed that the average of noise ratio, which means the ratio of relevant and irrelevant videos, is originally 72% in our dataset. Thus, in this experiment, we randomly insert noisy videos from other queries' ranking lists in the target initial ranking list so that the ratio of noise videos is 80%. Table 6 also demonstrates the MNDCG@5,10,20,30,40,50,60,70,80,90,100 of different methods. As shown in Table 6, our proposed graph-based reranking outperforms the other methods, and the improvements are consistent and stable at most of different depths of NDCG. Thus, it can be seen that our method including the global consistency analysis can improve the robustness to nosy videos.

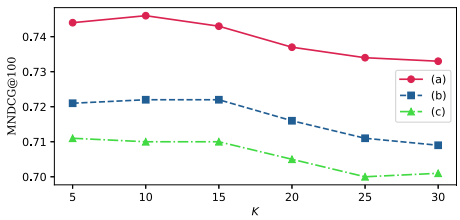

Fig. 3 Performance comparisons between different center nodes detection approaches for different parameter K in terms of MNDCG@100: (a) Ours, (b) PageRank, (c) HITS.

Next, in order to confirm the effectiveness of the proposed center node detection using a spectral clustering algorithm, we compare the proposed method with two popular representative node detection schemes including:

1) The PageRank algorithm [18] which was used in Google and was designed as a method for link analysis. The method is defined as "PageRank".
2) The HITS (Hypertext Induced Topic Selection) algorithm [31]. HITS makes the distinction between *hubs* and *authorities* and computes them in a mutually reinforcing way. The method is defined as "HITS".

Note that for implementation of PageRank and HITS, we also used the same video graph and its affinity matrix as those used in the proposed method. To further analyze the results, we compare the results of the different parameter K, which is the number of center nodes. Figure 3 depicts the performance of three types of methods, the proposed method, HITS and PageRank with different K ranging from 5 to 30 in terms of NDCG@100. From the results, we can see that the proposed method always gives better performance, and the best number for K is 10.

Finally, we also test the sensitivity of the two parameters ρ and α_L, which are used in the proposed method. We first set $\alpha_L = 0.5$ and vary ρ from 0.001 to 1. Figure 4 demonstrates the performance curve with respect to the variation of ρ. We then set $\rho = 0.1$ and vary α_L ($\alpha_G = 1 - \alpha_L$) from 0.1 to 0.9. Figure 5 demonstrates the performance curve with respect to the variation of α_L. Here, we also illus-

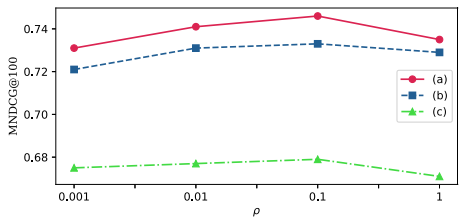

Fig. 4 Illustration of the effects of the parameter ρ in terms of MNDCG@100: (a) Ours, (b) Ours (without $\mathbf{A}_k$ optimization), (c) Ours ($\alpha_G = 0$).

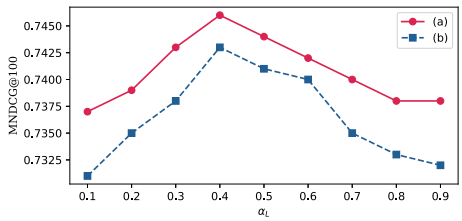

Fig. 5 Illustration of the effects of the parameter α_L in terms of MNDCG@100: (a) Ours, (b) Ours (without $\mathbf{A}_k$ optimization).

trate the performance of the methods based on the proposed method. From the results we can see that the performance of our approach will not be significantly degraded when the two parameters vary in a fairly wide range, and it can keep outperforming the other methods.

From the above experimental results, we can verify the effectiveness of the proposed method using the local and global consistency analysis. Therefore, the proposed method improves the performance of graph-based reranking in video searches.

4.4 Complexity Analysis

From the above solution process, we can see that its computational cost mainly contains three parts, which are for detecting global consistency, updating $\mathbf{r}$, and updating $\mathbf{A}_{\{L,G\}}$, respectively. First, the computational cost of the global consistency detection is $O(K^3 + KNt)$, where K is the number of clusters, N is the number of videos, and t is the number of k-means iterations. In the graph-based reranking method, we sparsify $\mathbf{W}_{\{L,G\}}$ by only keeping the l largest components in each row, where l is the number of neighbors for each video. From Eq. (8), we can see that the cost for updating $\mathbf{r}$ is $O(Nl)$. For updating $\mathbf{A}_L$ and $\mathbf{A}_G$, from the process in Algorithm 2, we can see that the cost is $O(T_1 Nld^2)$. Overall, the total time complexity for reranking is $O(K^3 + KNt + T(Nl + T_1 Nld^2))$, where d is the dimensionality of video feature vectors, and T and T_1 are the iteration times of optimization, respectively.

Besides theoretical analysis, we also test the time cost experimentally for the proposed method. It is implemented by using Python and run on a workstation with Intel Xeon E5-2620 v3, 2.4 GHz, 32GB memory in a single thread. By averaging the time cost of the all queries, our method can rank videos within 10s when $N = 500$ in a single thread. From the theoretical analysis and the experimental test discussed above, we can see that the efficiency of the proposed method is acceptable for real applications.

5. Conclusions

This paper has presented a method to improve performance of graph-based Web video search reranking. We first construct the video graph and detect global consistency over the graph by using a spectral clustering algorithm. From the clustering result, we extract center nodes, which are representative nodes of each cluster and then define the new affinity matrix and the global regularizer representing the similarity between center nodes and each node among the same video group. Secondly, by considering both local and global graph consistency, video search reranking is formulated as an optimization problem. The effectiveness of integrating local and global regularizers has been demonstrated. We have also compared our method with several existing reranking methods, and the results demonstrate the superiority of our method.

Acknowledgments

This research was financially supported by JSPS KAKENHI Grant Number 17K12687.

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 Trans. Pattern Anal. Mach. Intell., vol.22, no.12, pp.1349–1380, 2000.
[2] A. Hauptmann, R. Yan, W.H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news," IEEE Trans. Multimedia, vol.9, no.5, pp.958–966, 2007.
[3] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," Proceedings of International Conference on Content-Based Image and Video Retrieval, vol.2728, pp.238–247, 2003.
[4] A.P. Natsev, M.R. Naphade, and J. Tešić, "Learning the semantics of multimedia queries and concepts from a small number of examples," Proceedings of the ACM International Conference on Multimedia, pp.598–607, 2005.
[5] Y. Liu and T. Mei, "Optimizing visual search reranking via pairwise learning," IEEE Trans. Multimedia, vol.13, no.2, pp.280–291, 2011.
[6] L.S. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," Proceedings of the ACM International Conference on Image and Video Retrieval, pp.333–340, 2007.
[7] W.H. Hsu, L.S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," Proceedings of the ACM International Conference on Multimedia, pp.35–44, 2006.
[8] Y. Jing and S. Baluja, "VisualRank: Applying pagerank to large-scale image search," IEEE Transanctions on Pattern Analysis and Machine Intelligence, vol.30, no.11, pp.1877–1890, 2008.
[9] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," IEEE Trans. Image Process., vol.21, no.11, pp.4649–4661, 2012.
[10] X. Tian, Y. Yang, J. Wang, X. Wu, and X.-S. Hua, "Bayesian visual reranking," IEEE Trans. Multimedia, vol.13, no.4, pp.639–652,

2011.
[11] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," ACM Computing Surveys, vol.46, no.3, pp.38:1–38:38, 2014.
[12] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang, "Manifold-ranking based image retrieval," Proceedings of the ACM International Conference on Multimedia, pp.9–16, 2004.
[13] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol.290, no.5500, pp.2323–2326, 2000.
[14] M.A. Porter, J.P. Onnela, and P.J. Mucha, "Communities in networks," Notices of the AMS, vol.56, no.9, pp.1082–1097, 2009.
[15] U. von Luxburg, "A tutorial on spectral clustering," Statistics and Computing, vol.17, no.4, pp.395–416, 2007.
[16] S. Yoshida, T. Ogawa, and M. Haseyama, "Graph-based Web video search reranking through consistency analysis using spectral clustering," Proceedings of the IEEE International Conference on Multimedia and Expo, pp.1–6, 2016.
[17] W.H. Hsu, L.S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," Proceedings of the ACM International Conference on Multimedia, pp.971–980, 2007.
[18] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Computer Networks and ISDN Systems, vol.30, no.1-7, pp.107–117, 1998.
[19] L. Yang and A. Hanjalic, "Supervised reranking for Web image search," Proceedings of the ACM International Conference on Multimedia, pp.183–192, 2010.
[20] S. Zhang, M. Yang, T. Cour, K. Yu, and D.N. Metaxas, "Query specific rank fusion for image retrieval," vol.7573, pp.660–673, 2012.
[21] W. Liu, Y.-G. Jiang, J. Luo, and S.-F. Chang, "Noise resistant graph ranking for improved web image search," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.849–856, 2011.
[22] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," Proceedings of the International Conference on Machine Learning, pp.912–919, 2003.
[23] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," Proceedings of the International Conference on Neural Information Processing Systems, pp.321–328, 2004.
[24] B. Geng, D. Tao, and C. Xu, "DAML: Domain adaptation metric learning," IEEE Trans. Image Process., vol.20, no.10, pp.2980–2989, 2011.
[25] H. Li, M. Wang, and X.-S. Hua, "MSRA-MM 2.0: A large-scale web multimedia dataset," Proceedings of the IEEE International Conference on Data Mining Workshops, pp.164–169, 2009.
[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," Proceedings of the IEEE International Conference on Computer Vision, pp.4489–4497, 2015.
[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[28] C.D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.
[29] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," Foundations and Trends in Information Retrieval, vol.3, no.4, pp.333–389, 2009.
[30] D. Lu, X. Liu, and X. Qian, "Tag-based image search by social re-ranking," IEEE Trans. Multimedia, vol.18, no.8, pp.1628–1639, 2016.
[31] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, vol.46, no.5, pp.604–632, 1999.

Soh Yoshida received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2012, 2014, and 2016, respectively. He joined the Faculty of Engineering, Kansai University, in 2016, where he is currently an Assistant Professor. His research interests are Image/Video Semantic Analysis and Information Retrieval. He is a member of the ACM, the IEEE, the IEICE, and the ITE.

Takahiro Ogawa received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008, where he is currently an Associate Professor. His research interests are multimedia signal processing and its applications. He has been an Associate Editor of the ITE Transactions on Media Technology and Applications. He is a member of the ACM, the EURASIP, the IEICE, and the ITE.

Miki Haseyama received the B.S., M.S., and Ph.D. degrees in electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor in 1994. She was a Visiting Associate Professor with Washington University, USA, from 1995 to 1996. She is currently a Professor with the Graduate School of Information Science and Technology, Hokkaido University. Her current research interests include image and video processing and its development into semantic analysis. She has been the Vice President of the Institute of Image Information and Television Engineers (ITE), Japan, an Editor-in-Chief of the ITE Transactions on Media Technology and Applications, and the Director of the International Coordination and Publicity, Institute of Electronics, Information, and Communication Engineers (IEICE). She is a member of the IEICE, the ITE, and the ASJ.

Mitsuji Muneyasu received the B.E. and M.E. degrees in system engineering from Kobe University in 1982 and 1984, respectively, and Doctor of Engineering degree from Hiroshima University, Japan, in 1993. In 1984, he joined Oki Electric Industry Co., Ltd., in Tokyo, Japan. From 1990 to 1991, he was a Research Assistant at the Faculty of Engineering, Tottori University, Tottori, Japan. From 1991 to 2001, he was a Research Assistant and Associate Professor at the Faculty of Engineering, Hiroshima University, Higashi-Hiroshima, Japan. In 2001 he joined the Faculty of Engineering, Kansai University, Osaka, Japan, where he is currently a Professor. His research interests include image processing theory and nonlinear digital signal processing. He is a member of IEICE, IEEE, and IPSJ.