4-2013

Circular reranking for visual search

Ting YAO

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Citation

1

Circular Reranking for Visual Search

Ting Yao, Chong-Wah Ngo, *Member, IEEE*, and Tao Mei, *Senior Member, IEEE*

***Abstract*— Search reranking is regarded as a common way to boost retrieval precision. The problem nevertheless is not trivial especially when there are multiple features or modalities to be considered for search, which often happens in image and video retrieval. This paper proposes a new reranking algorithm, named circular reranking, that reinforces the mutual exchange of information across multiple modalities for improving search performance, following the philosophy that strong performing modality could learn from weaker ones, while weak modality does benefit from interacting with stronger ones. Technically, circular reranking conducts multiple runs of random walks through exchanging the ranking scores among different features in a cyclic manner. Unlike the existing techniques, the reranking procedure encourages interaction among modalities to seek a consensus that are useful for reranking. In this paper, we study several properties of circular reranking, including how and which order of information propagation should be configured to fully exploit the potential of modalities for reranking. Encouraging results are reported for both image and video retrieval on Microsoft Research Asia Multimedia image dataset and TREC Video Retrieval Evaluation 2007-2008 datasets, respectively.**

***Index Terms*— Circular reranking, multimodality fusion, visual search.**

I. Introduction

THE rapid development of Web 2.0 technologies has led to the surge of research activities in visual search. While visual documents are rich in audio-visual content and user-supplied texts, commercial visual search engines to date mostly perform retrieval by keyword matching. A common practice to improve search performance is to rerank the visual documents returned from a search engine using a larger and richer set of features. The ultimate goal is to seek consensus from various features for reordering the documents and boosting the retrieval precision. There are two general approaches along this direction: visual pattern mining [8] and multi-modality fusion [1], [2]. The former mines the recurrent patterns, either explicitly or implicitly, from initial search results and then moves up the ranks of visually similar documents. Random walk [9], for instance, performs self-reranking through identifying documents with similar patterns based on inter-image similarity and initial rank scores. This category of approaches, nevertheless, seldom explores the joint utilization of multiple modalities. Instead, different modalities are treated independently. Furthermore, the utilization of a modality is often application dependent, making it difficult to generalize the mining for general-purpose search. Multi-modality fusion, in contrast, predicts the importance of modalities, for instance, through fusion weight learning, and linearly combines them for reordering documents. The fusion, however, is done at the decision stage. More specifically, the estimation of fusion weights is mainly derived from the ranking scores in different ranked lists. There is no mechanism, however, where the interaction among multiple modalities could be exploited for reranking in a principle way.

This paper proposes a novel algorithm, named circular reranking, that takes advantages of both pattern mining and multi-modality fusion for visual search. More importantly, modality interaction is taken into account, on one hand to implicitly mine recurrent patterns, and on the other, to leverage the modalities of different strength for maximizing search performance. Figure 1 shows an overview of our proposed work compared with the existing methods. Given a ranked list of visual documents returned from a search engine, conventional methods use to perform random walk to rerank the results as shown in Figure 1(a). There are variants of approaches arisen from this methodology, for instance, conducting random walk on the original text space [31] or a new space built upon visual features [8], [10]. Typically each space is viewed as a graph that specifies the document proximity. More sophisticated approaches include lately fusing the reranked results from random walks in different feature spaces, or conversely, performing random walk on a unified graph that is fused from multiple features [9], [21]. Regardless of these different versions, a common issue not fully explored and studied is how the modalities should interact in view that their abilities in answering different queries could vary largely. We address this issue, as shown in Figure 1(b), from the viewpoint of mutual reinforcement. Specifically, different modalities interact by exchanging their feature spaces while preserving the original document scores for random walk. The exchange results in the following outcome: the ranks for documents which remain sharing similar local view of proximity in a different space tend to be upgraded. Take the text and bag-of-visual-words (BoW) feature spaces in Figure 1(b) as an example, the second and fourth images in the initial ranked list are similar in both text and BoW feature spaces. After reinforcement as in Figure 1(b), these group of images remains close in proximity and thus their ranks are likely to be moved up after random walk. Meanwhile, the second and third images

Manuscript received January 12, 2012; revised October 14, 2012; accepted December 4, 2012. Date of publication December 24, 2012; date of current version February 12, 2013. This work was supported in part by the National Natural Science Foundation of China under Grant 61272290 and Grant 61228205 and a grant from Microsoft Research Asia Windows Phone Academic Program FY12-RESOPP-107. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ton Kalker.

T. Yao and C.-W. Ngo are with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: tingyao2@student.cityu.edu.hk; cscwngo@cityu.edu.hk).

T. Mei is with Microsoft Research Asia, Beijing 100190, China (e-mail: tmei@microsoft.com).

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/TIP.2012.2236341

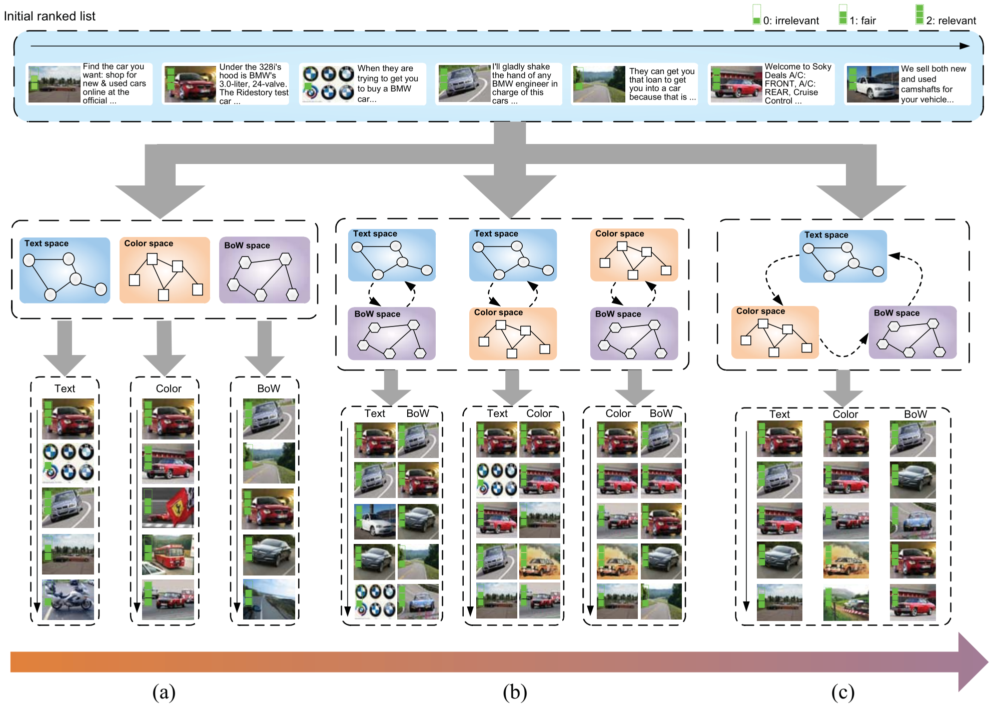
Initial ranked list
0: irrelevant
1: fair
2: relevant
Find the car you want: shop for new & used cars online at the official ...
Under the 328i's hood is BMW's 3.0-liter, 24-valve. The Ridestory test car ...
When they are trying to get you to buy a BMW car...
I'll gladly shake the hand of any BMW engineer in charge of this cars ...
They can get you that loan to get you into a car because that is ...
Welcome to Soky Deals A/C: FRONT, A/C: REAR, Cruise Control ...
We sell both new and used camshafts for your vehicle...
Text space
Color space
BoW space
Text
Color
BoW
(a)
(b)
(c)

Fig. 1. Reranking the initial search result (top) returned from Bing for the query "Find the images with car." The retrieved images are modeled as graphs respectively in different feature spaces, with nodes (images) attributed by the ranking scores and edges representing feature proximity. (a) Random walks: perform reranking by treating each feature space independently [9]. (b) Mutual reinforcement: exchanges modality spaces in a pairwise manner for random walks [30]. (c) Circular reranking: iteratively updates the image ranks by circular mutual reinforcement (this paper). Note that the final reranked list can be picked from the best performed modality or by linearly combining lists from different modalities (details in Section IV-C).

in the initial ranked list have similar textual descriptions, but dissimilar visual appearances in the BoW feature space. With mutual reinforcement, the rank of these two images becomes far apart from each other in the reranked lists produced by text and BoW spaces. In brief, similar to visual mining, the approach in Figure 1(b) implicitly mines the recurrent patterns of documents through random walk; but different from existing methods, modality interaction is considered by reinforcing the mutual exchange and propagation of information relevancy across different spaces.

By consolidating the idea of mutual reinforcement between two modalities, circular reranking arranges the reinforcement in a circular manner, as illustrated in Figure 1(c). The reinforcement is posted as a multi-random walk optimization problem, where the updated scores of documents as a result of mutual reinforcement is continuously propagated from one modality to another. The optimization converges when the propagation does not lead to further change of document ordering, and ideally, results in better ranking as shown in Figure 1(c). The preliminary version of this work, which performs co-reranking or mutual reinforcement in pairwise manner, is published in [30]. In this paper, we generalize this work to multiple modalities, where interaction is explored holistically among all the modalities, as opposed to locally based on multiple pairwise reinforcements. Furthermore, we address several issues arisen from this extension. These issues involve classical problems such as the dynamic adaptation of modality weights for information fusion. We analyze these problems in the paper and present solution about the use of modality importance for circular ordering of features when more than two modalities are considered. This ordering fully leverages the modalities of different strength for maximizing performance of query-dependent search. In addition, extensive empirical studies are also conducted for both image and video search reranking.

The main contribution of this work is the proposal of circular reranking for addressing the issue of multi-modality interaction in visual search. This issue also leads to the elegant view of how modalities of different strength should be leveraged for fusion, which is a problem not yet fully understood in the literature. The remaining sections are organized as follows. Section II describes the related work. Section III presents the problem formulation of circular reranking and its solution, while Section IV further details the ordering of modalities in the circular layout according to their ability in query answering. Section V presents the experimental results for image and video reranking. Finally, Section VI concludes this paper.

II. Related Work

We briefly group the related works for visual search reranking into two categories: recurrent pattern mining and multi-modality fusion. The former assumes the existence of common patterns among relevant documents for reranking. The later predicts or learns the contribution of a modality in search reranking.

A. Recurrent Pattern Mining

The research in this direction has proceeded along three different dimensions: self-reranking [6], [9], [7], crowd-reranking by exploiting online crowdsourcing knowledge [14], and example-based reranking by leveraging user-provided queries [15], [28].

Self-reranking seeks consensus from the initial ranked list as visual patterns for reranking. Fergues *et al.* employed probabilistic Latent Semantic Analysis (pLSA) for mining visual categories through clustering of images in the initial ranked list [6]. Candidate images are then reranked based on the distance to the mined categories. Similar in spirit, Hsu *et al.* employed information bottleneck (IB) reranking to find the clustering of images that preserves the maximal mutual information between the search relevance and visual features [7]. Later in [9], [17], Hsu *et al.* and Richter *et al.* further formulated the problem as random walk over a context graph built through linearly fusing multi-modalities for visual search. The idea of crowd-reranking is similar to self-reranking except that consensus is sought simultaneously from multiple ranked lists obtained from Internet resources. In [14], Liu *et al.* proposed a reranking paradigm by issuing query to multiple online search engines. Based on visual word representation, both salient and concurrent patterns are respectively mined to initialize a graph model for random walk based reranking.

Different from self- and crowd-reranking, example-based reranking relies on a few query examples provided by users for model learning. In [28], classifiers are learnt by treating query examples as positive training samples while randomly picking pseudo-negative samples from the bottom of initial ranked list. The classifiers which capture the visual distribution of positive and negative samples are then exploited for reranking. In another work by Liu *et al.* [15], query examples are utilized to identify relevant and irrelevant visual concepts, which are in turn employed to discover the rank relationship between any two documents using mutual information for correcting ranking of document pairs.

In short, while these approaches focus on the mining of recurrent patterns from different means, such as by random walk [9], external knowledge [15], and classifier learning [28], the interaction among modalities is not exploited for reranking. Our work in this paper contributes by studying not only mining patterns (or consensus) through random walk, but also how the consensus can be more reliably estimated by exploring modality interaction.

B. Multimodality Fusion

Multi-modality fusion based on weighted linear fusion is widely adopted in the literature [1]. Various combination operators such as ComSUM and Borda count have been proposed [27]. In this section, we only review the techniques based on how fusion weights are derived, where some of these techniques will be used together with circular reranking. Broadly, we can categorize the existing research into heuristic [20], adaptive [21], and query-class-dependent fusion [4], [13], [29].

Heuristic fusion rule-based defines the weights of modalities depending on types of user queries. For example, a set of weights are hand-crafted respectively for text-, concept- and visual-oriented queries [20]. Despite its simplicity, this strategy often yields satisfactory search performance as demonstrated in TRECVID evaluation [19]. Adaptive fusion, which determines the fusion weights dynamically on a per query basis, is seldom explored nevertheless. Typical approaches include exploiting the initial ranked lists of modalities based on score distribution [26] and rank agreement [21]. In [26], Wilkins *et al.* modeled the change of scores in a list to predict the importance of a modality. Specifically, the gradual (drastic) change of scores indicates the difficulty (capability) of a modality in distinguishing relevant from irrelevant items, and fusion weights are thus determined accordingly. In [21], Tan *et al.* leveraged rank agreement mined from multiple lists iteratively to update the weights of modalities until reaching an equilibrium stage.

In between the heuristic and adaptive fusion strategies, query-class-dependent fusion learns the weights based on training examples [4], [13], [29]. This scheme starts by predefining query classes, and then learning of weights is offline conducted on the query class level. During search, a given query is routed into one of the predefined classes, and the learnt weights are directly applied for fusion. This scheme is effective in general when the underlying query classes can be clearly defined and there are enough samples for weight learning. An alternative scheme based on concept-driven fusion was recently proposed by Wei *et al.* when these conditions could not be fulfilled [24]. Instead of performing query-to-class mapping, a query is mapped to a large number of visual concepts, in which the concept-to-modality relationship is learnt offline to determine the fusion weights.

The aforementioned approaches focus on the derivation of fusion weights through ranked lists and supervised learning. Our work is different in the way that we do not estimate fusion weights, but instead leverage the weights to indicate the importance of modalities for circular reranking. We exploit MAD [26] and query-class dependent [12] for this purpose, which will be elaborated in Section IV.

III. Circular Reranking

The basic idea of circular reranking is to facilitate interaction among different modalities through mutual reinforcement. In this way, the performance of strong modality is enhanced through communication with weaker ones, while the weak modality is also benefited by learning from strong modalities. We begin this Section by presenting the problem formulation, and followed by the analysis of solution convergence.

A. Problem Formulation

Suppose there are m single modalities, each associated with its initial ranked list consisting of N documents. We can have m views of relationship among these documents by constructing m affinity graphs each on a modality. In each graph, edges represent the document similarity and nodes carry the ranked scores. In this way, m runs of random walks can be conducted separately on each graph to rerank the initial results of m modalities. The spirit of circular reranking is to encourage modality interaction in a way that a modality starts random work by using the affinity graph of another modality, while preserving its original document scores. The exchange of affinity graphs explains the intuition of how consensus is reached among different modalities: the high scored documents can always propagate their scores to the similar set of documents and vice versa, despite the change of feature spaces. By arranging the m modalities linearly, such that $(i+1)^{th}$ order modality uses graph of the i^{th} order modality, this forms a circular ring. Each modality, once completing the random walk, will propagate the new result to influence its neighbor. We call this mutual reinforcement-based reranking as "circular reranking." Note that the framework has no limitation in terms of the type of modalities and the correlation between modalities. In other words, the used modalities could be of any types and with any correlations.

Denote p_{nij} as the similarity between any two documents i and j characterized by the n^{th} modality, and v_{nj} and r_{nj} as the initial and updated scores of a document j, respectively. The circular reranking is formulated as following:

$$\begin{cases} r_{1j}^{(t)} = \omega_1 \sum_i r_{mi}^{(t-1)} p_{mij} + (1-\omega_1)v_{1j} \\ r_{2j}^{(t)} = \omega_2 \sum_i r_{1i}^{(t)} p_{1ij} + (1-\omega_2)v_{2j} \\ \qquad \cdots \\ r_{nj}^{(t)} = \omega_n \sum_i r_{(n-1)i}^{(t)} p_{(n-1)ij} + (1-\omega_n)v_{nj} \\ \qquad \cdots \\ r_{(m-1)j}^{(t)} = \omega_{(m-1)} \sum_i r_{(m-2)i}^{(t)} p_{(m-2)ij} + \\ \qquad (1-\omega_{(m-1)})v_{(m-1)j} \\ r_{mj}^{(t)} = \omega_m \sum_i r_{(m-1)i}^{(t)} p_{(m-1)ij} + (1-\omega_m)v_{mj} \end{cases} \tag{1}$$

where the superscript (t) denotes iteration, and the tradeoff parameter ω_n $(0 \le \omega_n \le 1)$ weights the importance of the propagated and initial scores. Note that the first term in the equation represents information exchange from neighboring modality, while the second term is the original document score. The circular reranking starts by running random walk on the $1st$ modality using the affinity graph from the m^{th} modality. Once the random walk of the $1st$ modality completes, the updated scores are passed to the next modality. The first iteration ends when all modalities complete their random walks. This procedure is repeated until there is no change of document scores. To this end, there are m reranked lists being produced, with each from a single modality. The final reranked score $\mathbf{R}^* \equiv [r_j^*]_{1\times N}$ can be obtained by the utilization of these m reranked lists which will be discussed in Section IV-C.

The spirit of random walk is to rank a document higher if it is in close proximity with other documents that also have high scores. Circular reranking enforces additional constraint that a document is ranked higher if the high scored documents which are close to it are consistently observed across affinity graphs of different modalities. In reverse, if an item shares different set of similar documents in different modality views, the item will receive lower updated score at the end of an iteration. Different from the algorithms such as [17], [20] which explicitly find consensus from modalities for fusion, circular reranking inherently imposes the consensus in the graph model as a constraint for adjusting document scores.

B. Convergence Property

Eq. (1) can be expressed in a matrix form as following

$$\begin{cases} \mathbf{R}_1^{(t)} = \omega_1 \mathbf{R}_m^{(t-1)} \mathbf{P}_m + (1-\omega_1)\mathbf{V}_1 \\ \mathbf{R}_2^{(t)} = \omega_2 \mathbf{R}_1^{(t)} \mathbf{P}_1 + (1-\omega_2)\mathbf{V}_2 \\ \qquad \cdots \\ \mathbf{R}_n^{(t)} = \omega_n \mathbf{R}_{n-1}^{(t)} \mathbf{P}_{n-1} + (1-\omega_n)\mathbf{V}_n \\ \qquad \cdots \\ \mathbf{R}_{m-1}^{(t)} = \omega_{m-1} \mathbf{R}_{m-2}^{(t)} \mathbf{P}_{m-2} + (1-\omega_{m-1})\mathbf{V}_{m-1} \\ \mathbf{R}_m^{(t)} = \omega_m \mathbf{R}_{m-1}^{(t)} \mathbf{P}_{m-1} + (1-\omega_m)\mathbf{V}_m \end{cases} \tag{2}$$

here $\mathbf{R}_n \equiv [r_{nj}]_{1\times N}$ and $\mathbf{V}_n \equiv [v_{nj}]_{1\times N}$ are $1 \times N$ matrix composed of the propagated and initial scores of N documents respectively. $\mathbf{P}_n$ is the affinity graph characterizing document similarities in n modality. Circular reranking is guaranteed to converge for having the following property:

$$\lim_{t\to\infty} (\mathbf{R}_n^{(t+1)} - \mathbf{R}_n^{(t)}) = 0 \tag{3}$$

Proof:

$$\begin{aligned} \mathbf{R}_n^{(t+1)} - \mathbf{R}_n^{(t)} &= \omega_n \mathbf{R}_{n-1}^{(t+1)} \mathbf{P}_{n-1} + (1-\omega_n)\mathbf{V}_n \\ &\quad - (\omega_n \mathbf{R}_{n-1}^{(t)} \mathbf{P}_{n-1} + (1-\omega_n)\mathbf{V}_n) \\ &= \omega_n (\mathbf{R}_{n-1}^{(t+1)} - \mathbf{R}_{n-1}^{(t)})\mathbf{P}_{n-1} \\ &= \omega_n (\omega_{n-1} \mathbf{R}_{n-2}^{(t+1)} \mathbf{P}_{n-2} + (1-\omega_{n-1})\mathbf{V}_{n-1})\mathbf{P}_{n-1} \\ &\quad - \omega_n (\omega_{n-1} \mathbf{R}_{n-2}^{(t)} \mathbf{P}_{n-2} + (1-\omega_{n-1})\mathbf{V}_{n-1})\mathbf{P}_{n-1} \\ &= \omega_n \omega_{n-1} (\mathbf{R}_{n-2}^{(t+1)} - \mathbf{R}_{n-2}^{(t)})\mathbf{P}_{n-2}\mathbf{P}_{n-1} \\ &= \ldots \\ &= \omega_n \omega_{n-1} \ldots \omega_1 \omega_m \omega_{m-1} \ldots \omega_{n+1} \\ &\quad (\mathbf{R}_n^{(t)} - \mathbf{R}_n^{(t-1)})\mathbf{P}_n \mathbf{P}_{n+1} \ldots \mathbf{P}_m \mathbf{P}_1 \ldots \mathbf{P}_{n-1} \\ &= (\omega_n \omega_{n-1} \ldots \omega_1 \omega_m \omega_{m-1} \ldots \omega_{n+1})^t \\ &\quad (\mathbf{R}_n^{(1)} - \mathbf{R}_n^{(0)})(\mathbf{P}_n \mathbf{P}_{n+1} \ldots \mathbf{P}_m \mathbf{P}_1 \ldots \mathbf{P}_{n-1})^t. \end{aligned} \tag{4}$$

It is easy to see that Eq.(3) can be derived when each row of $\mathbf{P}_n$ is normalized to 1, and $0 \le \omega_1, \omega_2, \ldots, \omega_m \le 1$. □

C. Complexity Analysis

The complexity is governed by the number of modalities m, and the time required to complete m rounds of random walks in an iteration. Since the typical complexity of running a random walk is $O(N^2)$ where N is the number of documents, each iteration of circular reranking takes $O(m \times N^2)$. Thus, the running time complexity is $O(T \times m \times N^2)$, where T is the number of iterations before convergence. In our experiments, circular reranking usually converges for $T < 20$. Take 1,000 documents to be reranked by three modalities for example, our algorithm takes less than 0.002 seconds on a regular PC (Intel dual-core 3.33GHz CPU and 4GB RAM) to complete the whole reranking process. Since the algorithm is for reranking rather than ranking, where in practice only top 500 or 1,000 of documents in the initial ranked list will be considered, the speed is highly efficient and provides almost instant response.

IV. Reranking Property

The arrangement of modality in circular manner restricts how information should be exchanged from one to another. On the other hand, allowing the exchange to be happened between any two arbitrary modalities is also infeasible due to excessive number of possible pairs. Thus, a fundamental problem of circular reranking is the impact of modality ordering in affecting the effectiveness of information propagation. The order apparently is also characterized by the role that a modality plays in query answering. Furthermore, since the importance of a modality is expected to change in response to different types of queries, ideally one should weight the modality significance and then order modalities adaptively to maximize performance. In this section, we discuss the issues about the ordering, prediction, and utility of modalities in the context of circular reranking.

A. Order Analysis

Without loss of generality, we consider the case when there are three modalities available. The close form solution of the modality in the $3rd$ order is

$$\mathbf{R}_3^{\infty} = \mathbf{S}(\mathbf{I} - \omega_3\omega_2\omega_1\mathbf{P}_3\mathbf{P}_2\mathbf{P}_1)^{-1} \tag{5}$$

where

$$\mathbf{S} = (1-\omega_3)\mathbf{V}_3 + \omega_3(1-\omega_2)\mathbf{V}_2\mathbf{P}_2 + \omega_3\omega_2(1-\omega_1)\mathbf{V}_1\mathbf{P}_1\mathbf{P}_2. \tag{6}$$

Obviously, Eq.(5) is governed by the matrix $\mathbf{S}$ since the values of other parts remain unchanged over iteration. $\mathbf{S}$ is a linear combination of three terms, where each term is contribution from different modalities. The first term $(1-\omega_3)\mathbf{V}_3$ is its initial document score; the second term $\omega_3(1-\omega_2)\mathbf{V}_2\mathbf{P}_2$ represents the influence propagated from $2nd$ modality; and the third term $\omega_3\omega_2(1-\omega_1)\mathbf{V}_1\mathbf{P}_1\mathbf{P}_2$ is from $1st$ modality which has been refined by $2nd$ modality. Thus the inference of $2nd$ modality is indeed higher than that of $1st$ modality since it affects both the second and third terms of $\mathbf{S}$. This simple observation leads to the important conclusion that, to maximize the performance of $3rd$ modality, one should place a modality with stronger performance next to it and followed by weaker ones.

Generalizing this conclusion to m modalities, referring to Eq.(2), it means to order the modalities in ascending order of their significance. In other words, the strongest performed modality should be placed as $\mathbf{R}_m$, followed by $\mathbf{R}_{m-1}$ which is the second strongest and so on. The weakest performed modality should be ordered in the first position as $\mathbf{R}_1$. Circular reranking will start by using the affinity graph of the strongest modality (m^{th} modality) to influence the weakest one ($1st$ modality) to ensure the largest possible improvement for $1st$ modality. The refined $\mathbf{R}_1$ is then in turn used to influence the $2nd$ weakest modality. In the end, the m^{th} modality is affected and refined by $m-1$ weaker modalities before the start of next iteration.

B. Modality Importance

The significance of modalities with respect to a query could be utilized in two ways. First, as presented in the previous subsection, the modalities are prioritized to determine the circular ordering. Second, the tradeoff parameter ω_n in Eq.(1), which weights the importance of initial document scores provided by a modality, could also be set according to modality importance. In this way, stronger modality will have higher influence than weaker one during information propagation. In practice, however, the knowledge of modality significance is not always available. Furthermore, the significance of a modality could be varied depending on the nature of a given query. This implies the need for dynamically deciding the importance of modality on-the-fly during query time. In the literature, nevertheless, the problem of modality ranking remains an open question, though there have been several approaches being proposed. Here, we discuss two popular approaches: MAD (Mean Average Distance) [25], [26] and query-class-dependent fusion [12], and their use for circular reranking.

MAD directly derives the weights of modalities from their ranked lists. The weights, which are sought independently from one modality to another, are utilized to reflect the relative importance of modalities. The derivation is query-dependent for each modality and judged based on the distributions of scores in a ranked list. The hypothesis is that the shape of a curve, which depicts how scores are distributed in the descending order of their query relevance, reveals the degree of randomness in ranking. The basic idea is that the rapid change of document scores in the top part of a ranked list indicates the ability of a modality in distinguishing the relevant items from irrelevant ones. Conversely, the gradual change of scores gives clue that a modality is incapable of making clear decision and the ranking of documents is likely to be random. In MAD, the document initial scores $\mathbf{V}_n \equiv [v_{nj}]_{1\times N}$ of a ranked list given by the n^{th} modality are first normalized using max-min normalization as following

$$v'_{nj} = \frac{v_{nj} - \min\{v_{nj}\}}{\max\{v_{nj}\} - \min\{v_{nj}\}} \tag{7}$$

The mean average distance in scores of any two adjacent documents is then computed as

$$\text{MAD} = \frac{\sum_{j=1}^{N-1}(v'_{nj} - v'_{n(j+1)})}{N-1} \tag{8}$$

The rapid change in scores is then quantitatively determined by computing the ratio of MAD within the top subset of a ranked list ($topset$), versus that of a larger subset of the same ranked list ($largerset$). The ratio is named as Similarity Cluster (SC) in [26] and defined as

$$\text{SC} = \frac{\text{MAD}\{\text{topset}\}}{\text{MAD}\{\text{largerset}\}}. \tag{9}$$

In our implementation, the two subsets are selected as following: the top 10% ranked documents form the $topset$, while the top 90% of documents form the $largerset$. Basically the higher the value of SC is, the more important is a modality to a given query.

Query-class-dependent fusion requires supervised learning for inferring modality weights. The basic assumption is that user queries could be classified into a limited number of classes, where queries in each class have more or less similar view of modality importance. For instance, a query such as

Fig. 2. MSRA-MM dataset [23]. The 68 queries representing different visual concepts for web image search: (1) *angel*, (2) *animals*, (3) *baby*, (4) *backgrounds*, (5) *baseball*, (6) *batman*, (7) *beach*, (8) *bees*, (9) *birds*, (10) *boy*, (11) *cake*, (12) *car*, (13) *cartoon*, (14) *cat*, (15) *children*, (16) *chocolates*, (17) *cow*, (18) *cowboys*, (19) *disney*, (20) *dogs*, (21) *dragons*, (22) *earth*, (23) *email*, (24) *fish*, (25) *flags*, (26) *flowers*, (27) *food*, (28) *football*, (29) *frogs*, (30) *fruit*, (31) *games*, (32) *ghosts*, (33) *golf*, (34) *hairstyles*, (35) *hawaii*, (36) *heart*, (37) *horses*, (38) *hotels*, (39) *indians*, (40) *jesus*, (41) *lion*, (42) love, (43) *maps*, (44) *medical*, (45) *military*, (46) *nokia*, (47) *panda*, (48) *party*, (49) *people*, (50) *pigs*, (51) *plants*, (52) *police*, (53) *ronaldinho*, (54) *rose*, (55) *school*, (56) *snakes*, (57) *spider*, (58) *sports*, (59) *stars*, (60) *tiger*, (61) *trees*, (62) *turtles*, (63) *war*, (64) *waterparks*, (65) *weather*, (66) *wolves*, (67) *women*, and (68) *youtube*.

"find me videos about Lady Gaga" is likely to be answered by text than visual, while a query such as "find me images with stars of night sky" is likely to be answered by visual than text. With this assumption, the optimal fusion weights for each class are offline learnt from training examples using various search strategies, aiming to optimize the average search performance for queries within the same class. In other words, different from MAD, the fusion weights are optimized on a query-class basis. During retrieval, a query is routed to one of the classes, and the learnt optimal weights are applied directly for modality fusion. Readers can refer to the survey [12] for technical details. In our case, we employ query-class-dependent fusion for learning the tradeoff parameters, i.e., ω_n, for circular reranking. The learning basically characterizes the importance of initial ranked scores produced by different modalities w.r.t the nature of a query class.

C. Result Utilization

The result of interacting m modalities by circular reranking will produce m reranked lists each from a modality. Basically, there are three major ways of utilizing this result. The simplest way is by prioritizing the importances of modalities as described in Section IV-A, and then directly adopting the reranked score $\mathbf{R}_m$ produced by m^{th} modality as the final score $\mathbf{R}^*$. An alternative way is by introducing another fusion stage to linearly combine the m ranked lists from different modalities. There are various algorithms such as ComSUM [27] that can be employed for this purpose. Finally, for target-oriented search task such as color-based retrieval where color is the dominant feature while other modalities play the complementary roles, the circular reranking can be ordered in the manner that color is placed as the m^{th} modality followed by other modalities according to their importances. In this way, the final score $\mathbf{R}^*$ is simply taken from the reranked score produced by color.

V. Experiments

We empirically verify the merit of circular reranking from three aspects: 1) when strong and weak modalities are interacted, 2) when the modality significance is predicted by MAD, and 3) when the tradeoff parameters are learnt by query-class-dependent fusion. Three experiments were conducted respectively on the simulated, image and video datasets. The first experiment examines the degree on which search performance could be boosted when interacting modalities of different strength by circular reranking (V-A). The second experiment compares the performance of proposed approach with late fusion under oracle setting on MSRA-MM web image dataset [23]. The optimal results are then compared against a more practical scenario when MAD is used to estimate the modality importance (V-B). Finally, the adaptive setting of parameters for circular reranking is demonstrated by employing query-class-dependent fusion for video search conducted on TRECVID 2007 and 2008 datasets [22] (V-C).

TABLE I

Search Result Improvement for 2-D Point Ranking by Using Circular Reranking. Result Indicates the Improvement (Percentage of Reserved Pairs Being Corrected) Introduced for a Modality (First Row) After Interacting With Another Modality (Second Row)

Initial	x		$x+y$	
	y	$x+y$	x	xy
20% Noise	19.1%	64.1%	42.7%	48.5%
30% Noise	17.6%	62.5%	22.7%	32.4%
40% Noise	16.8%	60.9%	19.4%	31.9%

A. Simulation

We first examined how search performance is affected when modalities of strong and weak interact through circular reranking. The experiment was conducted on a toy dataset of 500 points randomly generated on the x-y coordinate. These points were ordered according to their Euclidean distance from the origin. We generated three lists from here by randomly exchanging the positions of 20%, 30% and 40% of points respectively, representing the initial ranked lists of different noise degree. Our task is to rerank the points in the lists by four modalities: x, y, xy and $x + y$. Apparently the former two modalities are considered weak in point ordering, while the later two are regarded as strong modalities. Furthermore, the modalities carry different levels of correlation: x and y are independent, x and y are moderately correlated with $x + y$, while xy and $x+y$ are highly correlated. Note that the quality of an initial rank list will impact the performance of reranking.

We designed four runs to compare the performances of circular reranking as a result of combining two different modalities. In the simulation, the generated noisy lists were directly treated as different initial ranked lists for reranking. Table I shows the search performances for modalities x and $x + y$ when circular reranking is employed. By using x-only modality for reranking, there are 50% of adjacent point pairs being incorrectly ordered in the top-10 result list. While for $x + y$ which being a relatively strong modality, the percentage of reverse point pairs is 6.7%. Table I basically shows different levels of improvement when modalities of different strength are integrated for circular reranking. For instance, when weak modality x is interacted with a strong modality $x + y$, more than 60% of improvement is attained. When combining with weak modality y, however, less than 20% of improvement is observed. Similar results are also obtained for strong modality $x + y$, where the larger improvement is achieved when interacting with modality xy than x. When further taking into account the quality of an initial ranked list, similar improvement trend is also observed, except that the degree of improvement is inversely proportional to the increase of noise level. This simulation verifies the basic features of circular reranking. The best possible improvement in reranking is achieved when interacting with a strong modality. In addition, the result from strong modality could also be improved when interacting with weaker one. In general, improvement is always expected even when interacting two weak modalities such as x and y and starting from a noisy initial list. It is worth noticing that the degree of improvement is directly related to the strength of modalities in interaction, and we observe no clear relationship of how modality correlation will impact the performance.

B. Web Image Reranking

The second experiment was conducted on the MSRA-MM dataset that consists of 68 queries collected from the query log of Microsoft Bing Search [3]. The queries covers various visual concepts as listed in Figure 2. Each query is associated with a ranked list of 900 images, and the surrounding text of each image is also provided. In total, the dataset contains 60,275 images. In the experiment, the top 500 images are reranked since in practice very few relevant images could be found when going deeper into the list [9], [14]. Three modalities are used for reranking: text, color moment (CM) and bag-of-visual-words (BoW). The text modality represents each image as a vector of keywords extracted from its surrounding text. To consider only salient terms for retrieval, only top 2000 frequent keywords are considered and each word is weighted by its term frequency. CM divides each image into 5-by-5 partitions and extracts a feature vector of 225 dimensions. BoW generates a visual vocabulary of 2000 words to describe image. Given an image, scale-invariant feature transform (SIFT) descriptors are computed for local regions detected by Difference of Gaussian (DoG) [16]. Each descriptor is then encoded as a word, and similar to text modality, BoW represents an image as a vector of visual words weighted by term frequency.

In the MSRA-MM dataset, the ground-truth of queries is given. The relevancy of each image is labeled in the scale of 0 to 2, indicating relevant (2), fair (1) or irrelevant (0). We adopted Normalized Discounted Cumulative Gain (NDCG) which takes into account the measure of multi-level relevancy as the performance metric. Given a query q, the NDCG score at the depth of d in the ranked list is defined by

$$NDCG@d = Z_d \sum_{j=1}^{d} \frac{2^{r^j} - 1}{\log(1 + j)} \tag{10}$$

where r^j represents the rating of an image in the ground-truth, Z_d is a normalization constant and is chosen so that $NDCG@d = 1$ for perfect ranking.

1) Initial Ranked Score: The initial ranked list for each modality is generated from the original list provided by the dataset. Similar to [5], [7], the ranking of images in an initial list is based on the intuition that relevant items are of similar to each other, while irrelevant items are likely to be different from one another. We employed k-means to partition the set of images in the original list as clusters. The images are then ranked according to the size of the clusters they belong to. Denote the initial ranked list of N images for n^{th} modality as $\mathbf{V}_n \equiv [v_{ni}]_{1\times N}$, the score of an image ranked at i^{th} position is computed as

$$v_{ni} = \lambda \times v_{ni}^c + (1 - \lambda) \times v_{Bi} \tag{11}$$

where v_{ni}^c is the score assigned as a result of clustering, and v_{Bi} is the ranking in the original list normalized by N. The tradeoff parameter between the two terms is set as $\lambda = 0.9$ as in [7], with the assumption that the original list is noisy and should be given much less weighting.

2) Performance Comparison: We compared the following approaches for performance evaluation.

- Single modality random walk [9]. Each modality reranks the initial search list by performing random walk on its affinity graph. The nodes in the graph are initialized by the scores computed using Eq.(11), and the edges between nodes are weighted by the cosine similarity between them. The three runs based on Text, CM and BoW are named as RWT, RWC and RWB respectively.
- Co-reranking [30]. Similar to circular reranking, except that only two modalities text and BoW are considered for mutual reinforcement. We name this run as CO in short.
- All-pairs Co-reranking. This run performs co-reranking for all the pairs among the modalities Text, CM and BoW. The result is reported based on the pair that exhibits the best performance. We name this run as CO^*.
- Late fusion [27]. There are variants of algorithms for linear fusion of multiple ranked lists. We employed CombSUM which exhibits superior performance than others as demonstrated in [27]. ComSUM first normalizes the scores of images in a list in the range of [0,1]. The final ranking of an image is decided by weighted summation of its normalized scores from multiple lists. The optimal weight of each modality is estimated by exhaustive grid search at per query level. The result of

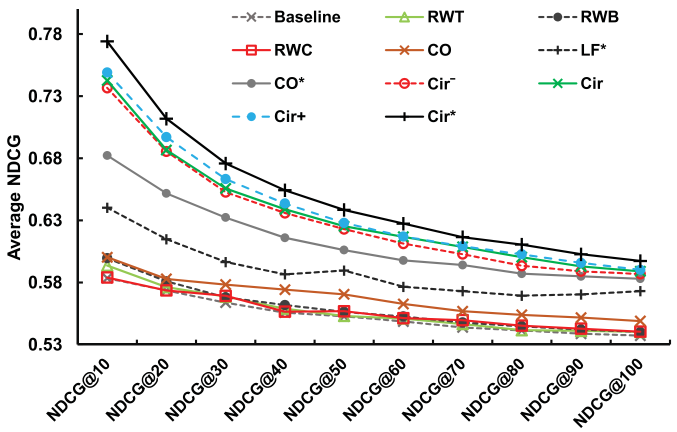

Fig. 3. NDCG of different approaches for web image reranking.

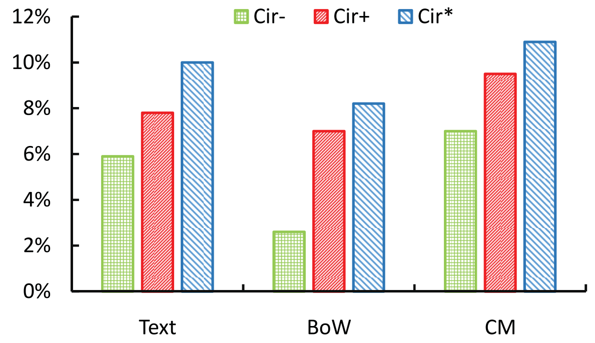

Fig. 4. Performance improvement of different modalities by circular reranking. The performances are compared against the results w.r.t their ranked lists at NDCG@50.

fusion basically indicates the best possible performance of late fusion in this dataset. We name this run as LF^*.

- Circular reranking. We designed four runs for our proposed approach: Cir^+, Cir^-, Cir and Cir^*. The first run Cir^+ ranks the importance of modalities by MAD on their initial ranked lists, and orders them accordingly for reranking. The result is reported for the modality that achieves the highest NDCG. The second run Cir^- ranks the modalities in reverse order of their importance, and the result of the modality that gets the lowest NDCG is shown. Basically, Cir^+ and Cir^- show the best and worst performances respectively when predicting modality significance by MAD. Cir is the actual run, where the result is reported based on the most important modality as selected by MAD. Finally, Cir^* is an oracle run that exhaustively evaluates all the possible orders of modalities, and the result corresponds to the order that exhibits the highest NDCG is reported.

Figure 3 shows the NDCG performances of eleven runs averaged over 68 queries in MSRA-MM dataset. The NDCG of baseline is evaluated based on the original list of a query provided by MSRA-MM. By random walk, the three single modalities runs slightly improve the baseline. RWT appears to be a weaker modality since the surrounding texts of images are not always content related. Overall, the results across different depths of NDCG consistently indicate that reranking using multi-modalities leads to a larger performance boost against baseline compared to single modality. Furthermore, circular reranking utilizing all three modalities also exhibits significantly better performance than CO which uses only text and BoW for mutual reinforcement. CO^* significantly improves CO, but the overall performance in the best case is still lower than that of circular reranking in the worst case (Cir^-). The result basically indicates the advantage of exploring interaction holistically among all modalities than in the pairwise manner.

There is a significant performance gap between the two oracle runs LF^* and Cir^*. Though both runs involve utilization of all the three single modalities, they are fundamentally different in the way that the performance of LF^* is as a result of combining different modalities, and Cir^* is by boosting each other through mutual reinforcement. As indicated by our results, allowing modality interaction can constantly lead to better performance gain than simple linear fusion. An interesting observation is that the performance gain of Cir^* tends to be large for the top few ranked results. The gain is however gradually decreased when going deeper into the list. In contrast, the improvement of LF^* is more obvious for the lower ranked items. Especially, the improvement is marginal if considering only the top-10 items in the ranked lists. This somewhat reveals the weakness of linear fusion, where when one of modalities is strong enough, further fusion will not lead to apparent improvement. In practice, the fusion may even degrade the performance of strong modality since fusion weights are query-dependent and not known in prior. Cir^*, in comparison, is benefited from the mechanism of modality interaction. The chance that a strong modality can learn from weaker modalities and lead to performance gain is better.

Compared to oracle run, the performance of Cir^+ which predicts modality importance by MAD approaches that of Cir^* along with the increasing of depth in the list. Therefore, using MAD as a prediction of modality importance is effective for web image reranking. In addition, Cir^+ performs consistently better than Cir^- at different depths of the list, which verifies the order analysis presented in Section IV-A as well. Figure 4 further shows the degree of improvement on three different modalities introduced by different versions of circular reranking. The result indicates that improvement can be generally expected, and larger degree of improvement is attained when modality importance could be correctly predicted.

Figure 5 details the performance across different queries. The NDCG values is normalized with respect to the best and worst results. Using circular reranking and MAD, Cir^+ exhibits better NDCG than single modalities run in 67 out of 68 queries. The improvement is observed in different types of queries. For instance, the images relevant to the queries "animal" and "flower" are diverse in appearance, resulting in poor performance by BoW or CM alone. Instead, text feature is found to be more helpful for these queries. Cir^+ successfully boosts up the normalized NDCG of both queries to 1. Similar degree of improvement is also found for queries such as "love" and "rose" where text description is highly noisy and visual features show better performance. In the extreme cases

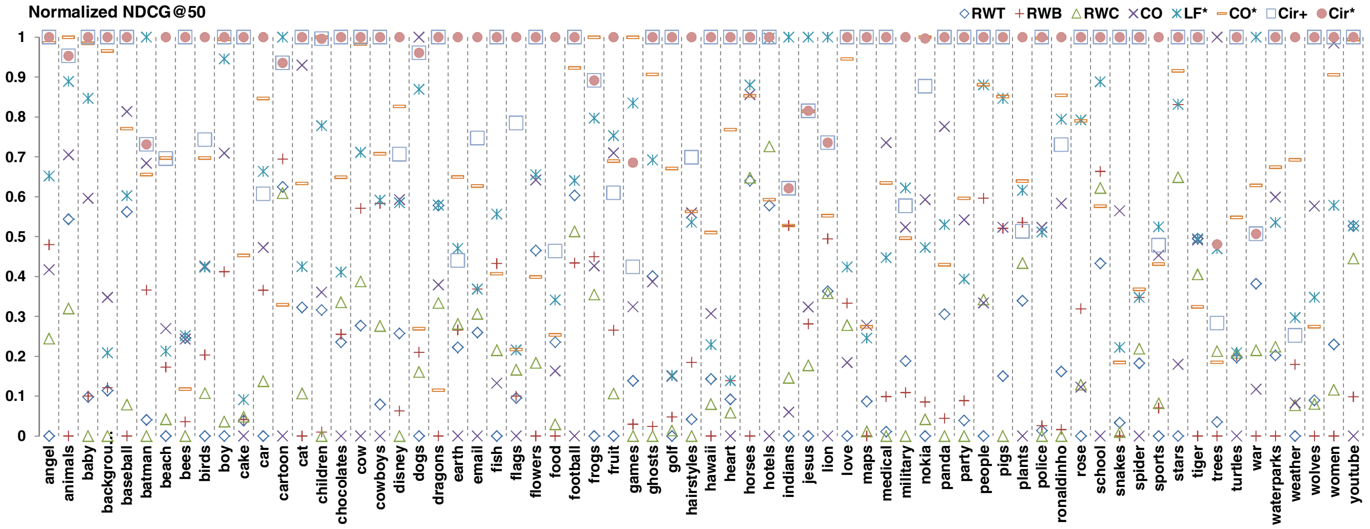

Fig. 5. Normalized NDCG@50 of different approaches across 68 queries. Note that NDCG is scaled with max–min normalization.

where all modalities are found to be less helpful, such as queries "baby" and "cake," improvement is also observed in Cir^+.

3) Parameter Sensitivity: A common problem with random walk based algorithm is the need to set the parameters to tradeoff the initial and updated scores. In the previous experiments, these values were optimally set in order to examine the performance of circular reranking irrespective of the parameter influence. Furthermore, we conducted experiments to test the sensitivity of ω_n parameters towards search performance. The result shows that the performance surface is always convex for most queries. With NDCG@50 as an example, the performance fluctuates within the range of 0.02, when the values of parameters are set from 0.1 to 0.9. According to our analysis, the values of different parameters at an optimal point reflect the performance of their initial ranked lists. This gives clue that the values should be set according to the relative effectiveness of modalities in search for a given query, though the exact values vary across different query types. When training examples are available, the values of parameters can be estimated. In Section V-C, we will further show how the values are learnt based on query-class dependent fusion.

C. Video Search Reranking

The third experiment was conducted on TRECVID 2007 (TV07) and 2008 (TV08) video benchmarks [22]. The datasets consist of 50 and 100 hours of Dutch videos respectively from the Netherland Institute for Sound and Vision. TV07 has 18,142 shots and TV08 has 35,766 shots. There are 24 multimedia queries in TV07 labeled with id from 197 to 220, and 48 queries in TV08 labeled with id from 221 to 268. Each query consists of a short text description, and a few image and video examples. Figure 6 shows examples of queries in TV07 and TV08. The task of video search is to retrieve and rank top 1,000 relevant shots for each query. Following TRECVID, we used average precision (AP) to measure the reranking performance for each query. The final result is presented with mean AP (MAP) which averages AP over all the queries.

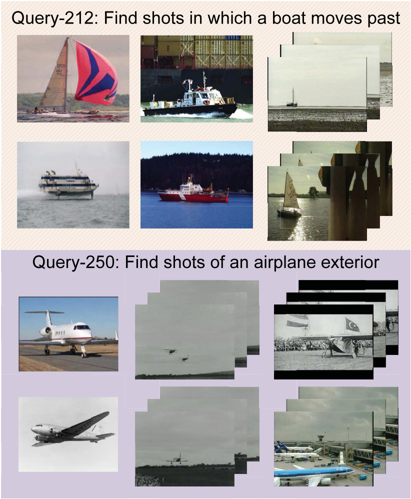
Query-212: Find shots in which a boat moves past
Query-250: Find shots of an airplane exterior

Fig. 6. Examples of textual and visual queries in TRECVID datasets.

We compared the following approaches, with the main focus is to examine how the tradeoff parameters ω_n learnt from query-class-dependent fusion for circular reranking will impact the performance.

1) Single modality search. We designed three baseline runs using Text, CM and BoW respectively. Text run uses the provided text descriptions of queries for search. The queries are matched against the speech transcripts of video shots, and the similarity is measured by BM25 [11]. CM and BoW use the image and video examples of queries to match against the keyframes in video shots. The query-shot relevancy is measured by cosine similarity. In the case when there are multiple image and video examples as query, the score of a shot is decided upon the example which is most similar to the shot.
2) Circular reranking. We designed five runs: ET, EB, EC, QC and Cir^*. The tradeoff parameters of former

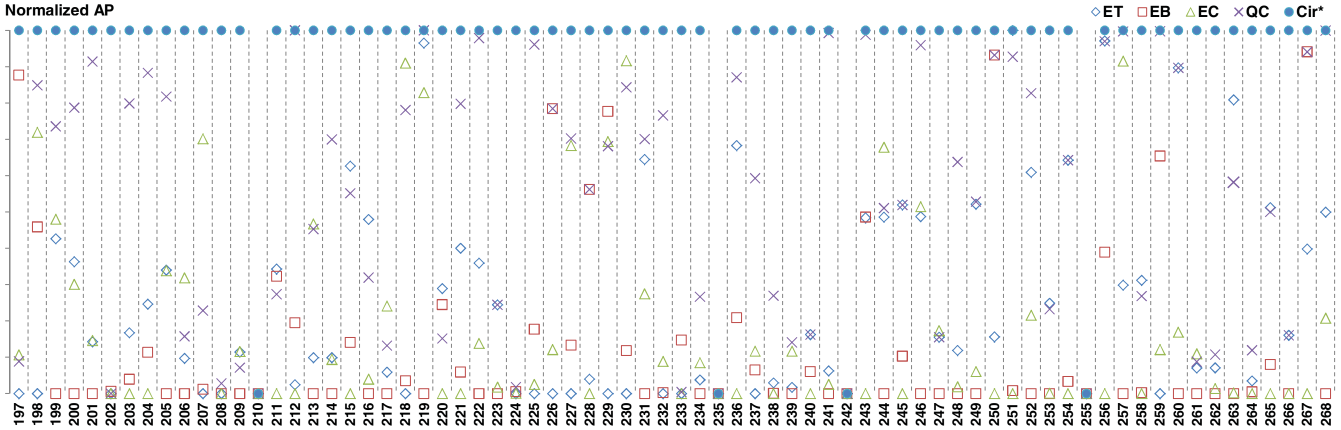

Fig. 7. Per query AP performance of different runs on TRECVID datasets. Note that AP (y-axis) is scaled with max-min normalization.

four runs are learned using training examples, where TV07 is used for training and TV08 for testing. The experiment is further repeated by using TV08 for training and TV07 for testing. In ET, EB and EC, the parameters are learnt and then fixed for all the queries for text, BoW and CM respectively. The learning is based on the exhaustive search of optimal parameters which offer the best MAP performance in the training set. These three runs are collectively called Fix. In QC, parameters are learnt based on query-class-dependent fusion [4], [29], where each query is mapped to a class defined by TRECVID. There are four classes defined for the queries: named entities (NE), person-thing (PT), event (E) and place (P). For each query class, the optimal values for tradeoff parameters are exhaustively search for each possible order of modalities. During search, a query is mapped to one of the classes. MAD was then employed to determine the order of modalities in circular reranking, and the learnt values of tradeoff parameters are applied. If a query belongs to multiple classes, we used the average of corresponding optimal parameters. Finally, $Cir*$ is an oracle run where the optimal parameters are exhaustively search for each query. This run represent the best possible performance that could be achieved by circular reranking.

Table II shows the MAP performance of different runs. Circular reranking consistently outperforms single modality runs in two datasets. In addition, QC, which adapts parameters at query class level, also exhibits better performance than Fix. Note that because TV07 has much more training queries than TV08, it is not surprise that the MAP of TV07 is better than TV08. Figure 7 further details the AP performance for all the 72 queries. Basically different modalities respond differently to queries. For instance, query id-212: “find shots in which a boat moves past,” is better retrieved with BoW. On the other hand, the query id-219: “find shots that contain the Cook character in the Klokhuis series” shows much better result with text modality. In the experiment, circular reranking successfully brings up the MAP performance of these queries. Among all the queries, QC outperforms Fix in 50 out of 72 queries. As TRECVID only defines four general query classes, the mapping from query to class could be ambiguous in some cases. The learning of optimal parameters is also practically not an easy task when the mapping is ambiguous. Though QC shows better MAP than Fix, queries such as id-207: “find shots of waterfront with water and building” and id-229: “find shots of one or more people where a body of water can be seen” show worse performance than Fix. The fact that many queries in TRECVID are categorized as Person-thing (PT) has made the searching of optimal weights difficult for this class. This in turn affects the overall performance stability. Among the 72 queries, compared to the methods other than $Cir*$, QC achieves the best performance for 46 queries, followed by ET for 12 queries, EC for 9 queries and EB for 5 queries. There are 26 queries where the performance of QC is close to that of $Cir*$. Among them, 23 queries share similar values of tradeoff parameters as $Cir*$. The general observation is that, as long as the estimated values do not deviate too much from the optimal ones, the performance of QC is expected to be close to $Cir*$.

TABLE II
MAP OF DIFFERENT RUNS FOR VIDEO RERANKING ON TRECVID 2007 AND 2008 DATASETS

TV-	Text	BoW	CM	ET	EB	EC	QC	$Cir*$
07	.029	.009	.024	.039	.016	.035	.045	.053
08	.010	.004	.005	.011	.007	.006	.015	.021

TABLE III
SIGNIFICANCE TEST AT 0.05 LEVEL ($X \gg Y$ INDICATES THAT X IS SIGNIFICANTLY BETTER THAN Y)

TV-	Reranking Methods
07	$Cir* \gg QC \gg ET \gg EB, EC$
08	$Cir* \gg QC \gg ET \gg EB, EC$

To verify that the performance of different runs presented in Table II is not by chance, we also conducted significance test using the randomization test [18] suggested by TRECVID. The target number of iterations used in the randomization is 100,000. The results at 0.05 significance level are shown in Table III. In both datasets TV07 and TV08, $Cir*$ is found to be significantly better than QC, and QC is in turns better than other runs.

VI. Conclusion

We have presented circular reranking which explores information exchange and reinforcement for visual search reranking. Particularly, we analyze the placement of modalities in the circular framework which could lead to the highest possible retrieval gain in theory for search reranking. To verify our claim, we have presented approaches based on the existing works in the literature for predicting the modality importance to sort and weight the modalities accordingly for circular reranking. Experiments conducted for image and video retrieval basically validate our proposal and analysis. Performance improvement is also observed when comparing to other reranking techniques such as linear fusion based on oracle setting and fixed weights learnt from training examples. The degree of improvement, though, is limited by how accurate the modality importance and fusion weights can be estimated, which could be noticed from our empirical results when comparing to the oracle setting of circular reranking. Thus, our future works include more in-depth studies of how fusion weights could be determined to boost the effectiveness of circular reranking.

References

[1] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli. "Multimodal fusion for multimedia analysis: A survey," *Springer Multimedia Syst. J.*, vol. 16, no. 6, pp. 345–379, 2010.

[2] C. A. Bhatt and M. S. Kankanhalli, "Multimedia data mining: State of the art and challenges," *Multimedia Tools Appl.*, vol. 51, no. 1, pp. 35–76, 2011.

[3] *Bing*. (2009) [Online]. Available: http://www.bing.com/

[4] T.-S. Chua, S.-Y. Neo, H.-K. Goh, M. Zhao, Y. Xiao, G. Wang, S. Gao, K. Chen, Q. Sun, T. Qi, "TRECVID 2005 by NUS PRIS," in *Proc. NIST TRECVID Workshop*, 2005, pp. 1–11.

[5] K. M. Donald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *Proc. ACM Int. Conf. Image Video Retr.*, Jul. 2005, pp. 61–70.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1816–1823.

[7] W. Hsu, L. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 35–44.

[8] W. Hsu, L. Kennedy, and S.-F. Chang, "Reranking methods for visual search," *IEEE Multimedia*, vol. 14, no. 3, pp. 14–22, Jul.–Sep. 2007.

[9] W. Hsu, L. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 971–980.

[10] Y. S. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, Nov. 2008.

[11] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments," *Inf. Process. Manage.*, vol. 36, no. 6, pp. 779–840, 2000.

[12] L. Kennedy, S. F. Chang, and A. Natsev, "Query-adaptive fusion for multimodal search," *Proc. IEEE*, vol. 96, no. 4, pp. 567–588, Apr. 2008.

[13] L. Kennedy, A. Natsev, and S. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 882–891.

[14] Y. Liu, T. Mei, and X.-S. Hua, "CrowdReranking: Exploring multiple search engines for visual search reranking," in *Proc. ACM Special Interest Group Inf. Retr.*, 2009, pp. 500–507.

[15] Y. Liu, T. Mei, X. Wu, and X.-S. Hua, "Optimizing video search reranking via minimum incremental information loss," in *Proc. ACM Int. Workshop Multimedia Inf. Retr.*, 2008, pp. 253–259.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[17] F. Richter, S. Romberg, E. Hörster, and R. Lienhart, "Multimodal ranking for image search on community databases," in *Proc. ACM SIGMM Int. Workshop Multimedia Inf. Retr.*, 2010, pp. 63–72.

[18] J. P. Romano, "On the behavior of randomization tests without a group invarianceassumption," *J. Amer. Statistical Assoc.*, vol. 85, no. 411, pp. 686–692, 1990.

[19] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proc. ACM SIGMM Int. Workshop Multimedia Inf. Retr.*, 2006, pp. 321–330.

[20] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, J. R. R. Uijlings, J. He, X. Li, I. Everts, V. Nedovic, M. van Liempt, R. van Balen, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J. M. Geusebroek, T. Gevers, M. Worring, A. W. M. Smeulders, and D. C. Koelma, "The mediamill trecvid 2008 semantic video search engine," in *Proc. NIST TRECVID Workshop*, 2008, pp. 1–14.

[21] H.-K. Tan and C.-W. Ngo, "Fusing heterogeneous modalities for video and image re-ranking," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2011, p. 15.

[22] *TRECVID*. (2010) [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[23] M. Wang, L. Yang, and X.-S. Hua, "MSRA-MM: Bridging research and industrial societies for multimedia information retrieval," Microsoft, Washington, DC, Tech. Rep. MSR-TR-2009-30, 2009.

[24] X.-Y. Wei, Y.-G. Jiang, and C.-W. Ngo, "Concept-driven multi-modality fusion for video search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 1, pp. 62–73, Jan. 2011.

[25] P. Wilkins, T. Adamek, G. J. F. Jones, N. E. O'Connor, and A. F. Smeaton "Trecvid 2007 experiments at dublin city university," in *Proc. NIST TRECVID Workshop*, 2007, pp. 1–8.

[26] P. Wilkins, P. Ferguson, and A. F. Smeaton, "Using score distributions for query-time fusion in multimediaretrieval," in *Proc. ACM SIGMM Int. Workshop Multimedia Inf. Retr.*, 2006, pp. 51–60.

[27] P. Wilkins and A. F. Smeaton, "Weighted data fusion for content-based multimedia information retrieval," in *Proc. Internet Multimedia Search Min.*, Jan. 2010, pp. 1–45.

[28] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. ACM Int. Conf. Image Video Retr.*, 2003, pp. 238–247.

[29] R. Yan, J. Yang, and A. Hauptmann, "Learing query-class dependent weights in automatic video retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2004, pp. 548–555.

[30] T. Yao, T. Mei, and C.-W. Ngo, "Co-reranking by mutual reinforcement for image search," in *Proc. ACM Int. Conf. Image Video Retr.*, 2010, pp. 34–41.

[31] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2007, pp. 739–744.

Ting Yao received the B.Sc. degree in theoretical and applied mechanics, the B.Eng. degree in electronic information engineering, and the M.Eng. degree in signal and information processing from the University of Science and Technology of China, Hefei, China, in 2004, 2004, and 2008, respectively. He is currently pursuing the Ph.D. degree in computer science with the City University of Hong Kong, Kowloon, Hong Kong.

He was a Software Engineer with the Alibaba Company, Beijing, China, from 2008 to 2010. His current research interests include multimedia content analysis, large-scale search, and transfer learning.

Chong-Wah Ngo (M'02) received the B.Sc. and M.Sc. degrees in computer engineering from Nanyang Technological University, Singapore, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 1994, 1996, and 2000, respectively.

He is currently an Associate Professor with the Department of Computer Science, City University of Hong Kong, Kowloon. He was a Post-Doctoral Scholar with the Beckman Institute, University of Illinois in Urbana-Champaign. He was a Visiting Researcher with Microsoft Research Asia. From 2008 to 2009, he was the Chairman of ACM (Hong Kong Chapter). His current research interests include large-scale multimedia information retrieval, video computing, and multimedia mining.

Dr. Ngo is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA. He is also the Program Co-Chair of the 2012 ACM Multimedia Modeling and the 2012 International Conference on Multimedia Retrieval.

Tao Mei (M'07–SM'11) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He is currently a Researcher with Microsoft Research Asia, Beijing, China. He has authored or co-authored over 130 papers in journals and conferences, and has authored eight book chapters. He holds four U.S. patents and more than 30 in pending. His current research interests include multimedia information retrieval and computer vision.

Dr. Mei was a recipient of several paper awards from prestigious multimedia conferences, including the Best Paper Award and the Best Demonstration Award at ACM Multimedia in 2007, the Best Poster Paper Award at the IEEE MMSP in 2008, the Best Paper Award at ACM Multimedia in 2009, the Top 10% Paper Award at the IEEE MMSP in 2012, the Best Paper Award at ACM ICIMCS in 2012, and the Best Student Paper Award at the IEEE VCIP in 2012. He received Microsoft Gold Star Award in 2010. He is an Associate Editor of Neurocomputing and the *Journal of Multimedia*, a Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the *IEEE Multimedia Magazine*, the *ACM/Springer Multimedia Systems*, and the *Journal of Visual Communication and Image Representation*. He is the Program Co-Chair of MMM 2013, and the General Co-Chair of ACM ICIMCS 2013. He is a Senior Member of the ACM.