Multimedia Search Reranking: A Literature Survey

Tao Mei, Microsoft Research Asia
Yong Rui, Microsoft Research Asia
Shipeng Li, Microsoft Research Asia
Qi Tian, University of Texas at San Antonio

The explosive growth and widespread accessibility of community contributed media content on the Internet have led to a surge of research activity in multimedia search. Approaches that apply text search techniques for multimedia search have achieved limited success as they entirely ignore visual content as a ranking signal. Multimedia search re-ranking, which reorders visual documents based on multimodal cues to improve initial text-only searches, has received increasing attention in recent years. Such a problem is challenging because the initial search results often have a great deal of noise. Discovering knowledge or visual patterns from such a noisy ranked list to guide the re-ranking process is difficult. Numerous techniques have been developed for visual search re-ranking. The purpose of this paper is to categorize and evaluate these algorithms. We also discuss relevant issues such as data collection, evaluation metrics, and benchmarking. We conclude with several promising directions for future research.

Categories and Subject Descriptors: H.5.1 [**Information Interfaces and Presentation**] Multimedia Information Systems; H.3.3 [**Information Search and Retrieval**] Retrieval models

General Terms: Algorithms, Experimentation, Human Factors.

Additional Key Words and Phrases: Multimedia information retrieval, visual search, search re-ranking, survey.

1. INTRODUCTION

The proliferation of capture devices and the explosive growth of online social media have led to the countless private image and video collections on local computing devices such as personal computers, cell phones, and personal digital assistants, as well as the huge yet increasing public media collections on the Internet [Boll 2007]. For example, the most popular photo sharing site—Flickr [Flickr], reached five billion photo uploads in 2011, as well as 3-5 million new photos uploaded daily [Kennedy et al. 2007]. Facebook held more than 60 billion photos shared by its communities as of 2011 [Facebook], while Youtube streams more than one billion videos per day worldwide [YouTube].

Such explosive growth and widespread accessibility of visual content have led to a surge of research activity in visual search. The key problem is retrieving visual documents (such as images, video clips, and Web pages containing images or videos) that are relevant to a given query or user search intention from a large-scale database. In the last decade, visual search has attracted a great deal of attention,

Tao Mei, Yong Rui, and Shipeng Li are with Microsoft Research Asia, Beijing 100080, P. R. China (email: {tmei, yongrui, spli}@microsoft.com). Qi Tian is with the University of Texas at San Antonio (email: qitian@cs.utsa.edu).

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

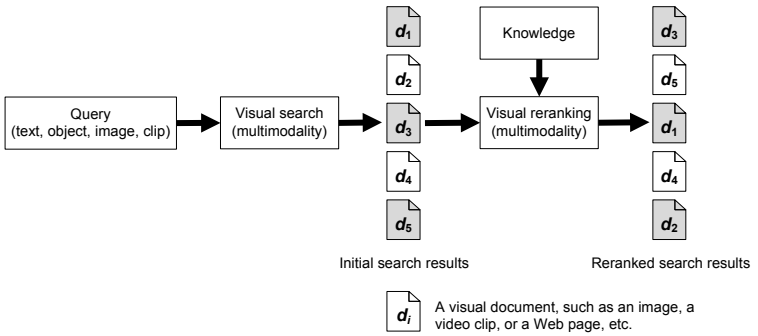

Fig. 1. A general process for multimedia search re-ranking, designed similar to the figure in [Hsu et al. 2007].

though it has been studied since the early 1990s (referred to as content-based image/video retrieval [Lew et al. 2006], [Li et al. 2007], [Rui et al. 1999]). Many research demonstrations and commercial applications have been developed. Due to the great success of text search, most popular image and video search engines, such as Google [Google], Bing [Bing], Yahoo! [Yahoo!], and so on, build upon text search techniques by using the non-visual information (such as surrounding text and user-provided tags) associated with visual content. This kind of multimedia search approach cannot always achieve satisfying results as it entirely ignores the visual content as a ranking signal [Chang et al. 2007], [Datta et al. 2008], [Hauptmann et al. 2008a], [Hsu et al. 2007], [Snoek and Worring 2009].

To address the problems of visual search approaches, multimedia search re-ranking has received increasing attention in recent years. It is defined as the reordering of visual documents based on the information manifested in the initial search results or a knowledge base to improve the search performance. This information actually consists of multimodal cues that can be the knowledge or specific patterns mined from the initial ranked list, query examples, or any available auxiliary knowledge. From another perspective, re-ranking can be viewed as a post-process of core search. Figure 1 shows a general process for multimedia search re-ranking. A visual document might be an image, a video clip, or a Web page containing images or videos. Given an initial ranked list of visual documents returned by any search approach, visual search re-ranking improves search performance by reordering these visual documents based on the multimodal cues. For example, in a real image search engine, the initial text search results can be re-ranked, according to the visual similarity to a given example [Bing], [Google] or color style (e.g., color or grey). Another example in an image retrieval system is to re-rank or filter the images accord-ing to some predefined categories [Fergus et al. 2004], [Wnuk and Soattoh 2008]. In the settings of object retrieval, a geometric verification step is usually introduced to re-rank the results returned from the bag-of-words (BoW) model based on checking the spatial configurations [Philbin et al. 2007], [Jegou et al. 2010]. The challenges associated with multimedia search re-ranking can be attributed to the following factors:

—*Unsatisfying initial search performance*. The initial search results usually contain a small portion of relevant documents. For example, the best automatic video search only achieves about 10% of the mean average precision (MAP) in TRECVID 2008 [TRECVID] [1]. The most popular BoW model (without re-ranking) for object retrieval can only achieve about 30%-70% of MAP on the Oxford build facade dataset, depending on codebook training and visual features [Chum et al. 2007], [Chum et al.

[1] This performance was conducted over 380 hours' video with 24 queries in total, which indicates that on average between 2 and 3 of the top 10 returned video clips are estimated to contain the desired video.

(a) Google (b) Bing (c) Yahoo! (d) Videosurf

Fig. 2. Examples of multimedia search re-ranking in some commercial search engines.

2011], [Philbin et al. 2007]. This leads to large visual variance and little relevant information within the initial search results. It is challenging to mine knowledge from such noisy data.

—*Lack of available knowledge or context for re-ranking*. Although we can design specific search interfaces to enable users to better formulate their queries, or collect demographic information (e.g., name, interest, location, etc.) or search logs, most search users are reluctant to provide their profiles or visual query examples (e.g., an image, a set of video keyframes, or a video clip).

—*Large-scale dataset*. Most existing approaches are not extensible for a large-scale dataset due to the algorithmic scalabilities and response time. As a result, only the top returned documents (e.g., top 1,000 images [Liu et al. 2009], or video shots [Hsu and Chang 2007], [Hsu et al. 2006]) are usually considered in the re-ranking process. An ideal re-ranking system would be able to handle all the documents in real time.

Many existing commercial search engines have developed different re-ranking schemes to improve the search experience. Figure 2 shows some examples of re-ranking designs. For example, Google [Google], [Jing and Baluja 2008a], [Jing and Baluja 2008b] and Bing [Bing], [Cui et al. 2008a], [Cui et al. 2008b] support retrieving similar or near-duplicate objects in their text-only image search results. Yahoo! [Yahoo!] and Bing [Bing], [Wang and Hua. 2011] integrate a set of content filters (also called attributes) such as “image size,” “layout,” “style,” and “popular queries,” in their image searches, while Videosurf [VIDEOSURF] uses a face detector as the filter for re-ranking keyframes. Although these content filters can facilitate some specific search tasks, most of them rely on simple features and do not directly represent the relevant information associated with the query. The key problem—the relevance between the search results and the query—still remains a challenging and open research problem. There is no generic approach that can deal with all kinds of query and search intent. On the other hand, “type less and search more” is the most desired feature in most search engines. It is thus not practical to let users spend considerable time and perform several rounds of interactions to look for desired search results. Therefore, it remains an open issue to re-rank the visual search results according to the query and user intent. In other words, there is a gap between the user search intent and the results from existing search systems.

While numerous approaches have been proposed for multimedia search re-ranking, we are unaware of any survey on this particular topic. Clearly, multimedia search re-ranking has been an important and hot topic in both academia and the industry. We have observed that almost all notepapers in recent TRECVID proceedings have adopted re-ranking techniques for improving search task performance [TRECVID], not to mention many works on image and object retrieval in the computer vision community where re-ranking has become a key post-processing step [Philbin et al. 2007], [Jegou et al. 2010], [Chum et al. 2011]. It is worthwhile to re-visit and categorize the current techniques and ex-

plore the potential benefits and challenges that both the multimedia and vision communities offer to visual search re-ranking and vice versa.

The research on visual search re-ranking has proceeded along four paradigms from the perspective of the knowledge exploited for mining relevant information: 1) *self-re-ranking* which only uses the initial search results; 2) *example-based-re-ranking* which leverages user-provided query examples; 3) *crowd-re-ranking* which explores the crowdsourcing knowledge available on the Internet, e.g., the multiple image and video search engines or sites, the user-contributed online encyclopedia like Wikipedia [Wikipedia], and so on; and 4) *interactive-re-ranking* which involves user interaction to guide the re-ranking process. The scope of this survey is as follows. We will first introduce typical multimedia information retrieval systems and the role of re-ranking in these systems. We will then review the re-ranking methodologies in terms of these dimensions. Since these data-driven methods rely heavily on the data sets and the corresponding knowledge mined from these data, we also discuss the data sets that are suitable for visual re-ranking. A related problem is how to evaluate the performance of a re-ranking method. Therefore, we review the performance metrics and evaluations for visual search re-ranking in this paper. The scope of this survey will cover the approaches inspired from multiple research fields such as computer vision, machine learning, text retrieval, and human-computer interaction.

As most re-ranking methods for visual search are highly motivated by the re-ranking and rank aggregation methods in the text domain, we first discuss the related techniques in text search, which can provide a comparable analysis of text and multimedia. However, as this paper focuses on the visual domain, we only briefly introduce some representative methods.

1.1 Re-ranking in Text Domain

Similar to visual search re-ranking, the research on text search re-ranking can be also categorized into the following paradigms.

—*Self-re-ranking*. Analogous to multimedia re-ranking, the self-re-ranking methods for text search also include: 1) clustering-based method [Lee et al. 2001], [Na et al. 2008], where the initially retrieved documents are grouped into different clusters, 2) pseudo relevance feedback [Cao et al. 2007], [Tseng et al. 2008], where the top ranked documents are regarded as "positive" when learning a ranking model, and 3) graph-based method [Bendersky and Kurland 2008], [Brin and Page 1998a], [Deng et al. 2009], [Kurland and Lee 2005], [Lin 2008], where a graph is built locally from the initial top ranked documents or globally from the entire document collection.

—*Example-based-re-ranking* [Bogers and Bosch 2009], [Zloof 1975a], [Zloof 1975b]. Query-by-Example (QBE) was first proposed by Zloof in the 1970s [Zloof 1975a], [Zloof 1975b]. The motivation is to parse a user's query into a structured statement expressed in a database manipulation language. Later, researchers investigate ways to understand the query provided by users using accompany examples. For example, Bogers *et al.* propose dividing the IR test collections into different subcollections, and applying a linear fusion of the search results from disparate baseline results [Bogers and Bosch 2009]. The weights for fusion are determined by the *authoritative* scores, which reflect the expertise between authors on certain topics.

—*Crowd-re-ranking* [Carbonell 1997], [Chen et al. 2011], [Dwork et al. 2001], [Liu et al. 2007], [Renda and Straccia 2003], [SavvySearch], [White et al. 2008]. For example, translingual information retrieval (TIR) is characterized by providing a query in one language and searching documents in one or more different languages [Carbonell 1997]. This is similar to the setting of TRECVID search tasks, where a video document is probably associated with different machine translated languages [TRECVID]. Metasearch is a prominent approach to combine the search result lists returned by

multiple search engines [Dwork et al. 2001], [Renda and Straccia 2003], [SavvySearch]. Each document in a returned list is ordered with respect to a search engine and a relevance score. Liu *et al.* propose to leverage user-labeled data to perform metasearch in a supervised manner [Liu et al. 2007], while Chen *et al.* suggest a semi-supervised approach to ranking aggregation by leveraging the large amounts of unlabeled data [Chen et al. 2011]. Different from metasearch where the final search results are the combination of multiple lists, White *et al.* propose to only provide the best single list from multiple search engines [White et al. 2008]. A learning-based approach is adopted to support switching between search engines.

—*Interactive-re-ranking* [Rohini and Varma 2007], [Yamamoto et al. 2007]. For example, a user is enabled to edit a part of the search results (i.e., delete and emphasis operations) in [Yamamoto et al. 2007]. The operations are then propagated to all of the results to re-rank them. Rohini *et al.* propose to learn the profiles of the users using machine learning techniques by making use of past browsing histories, and then re-rank the results based on collaborative filtering techniques [Rohini and Varma 2007].

In addition, text search re-ranking also involves the analysis of query logs [Teevan et al. 2007], [Zhuang and Cucerzan 2006], as well as the consideration of the diversity of search results [Carbonell and Goldstein 1998]. A more comprehensive study on the combination approach for information retrieval can be found in [Croft 2000].

With the aim of providing a comprehensive and critical survey of current approaches to visual search re-ranking, this paper is organized as follows. Section 2 gives a detailed review of techniques to visual search re-ranking, including general visual search framework, as well as an overview of re-ranking from the perspective of Bayesian formulation and methodologies for re-ranking. Moreover, we offer a brief survey on re-ranking for text search, from which visual search re-ranking is motivated. Section 3 discusses benchmarking datasets and evaluation criteria. We conclude this paper with a discussion of several promising directions in Section 4.

2. MULTIMEDIA SEARCH RE-RANKING: METHODOLOGIES

In this section, we review existing approaches to visual search re-ranking. We first position re-ranking as a key component in a typical visual search system and provide a Bayesian formulation for overviewing re-ranking problem. We then classify re-ranking approaches into four categories and discuss each of them in details. We also discuss other recent techniques for re-ranking, such as query suggestion and user interface, as well as the re-ranking methods in text domain.

2.1 The Role of Re-ranking in Multimedia Search

A typical visual search system consists of several components, including query analysis, an index module, uni-modal search (e.g., text, visual, and concept searches), and re-ranking. Figure 3 shows a generic visual search framework. Usually, the query in a visual search system consists of a piece of a textual query (e.g., "find shots in which a boat moves past") and/or probably a set of query examples (e.g., objects, images, or video keyframes or clips) [2]. Via query analysis, the meaningful or important keywords and their expanded related terms are obtained based on the textual query. Meanwhile, the visual examples can be mapped to some relevant high-level concepts by the pre-trained classifiers for

[2] The query includes a textual sentence and several query examples, which is the typical setting of automatic search in TRECVID [TRECVID]; while in computer vision community, the typical setting for query includes a single image example or an object, usually a rectangle region (e.g., a building facade, a landmark, an animal, and so on), within a query image.

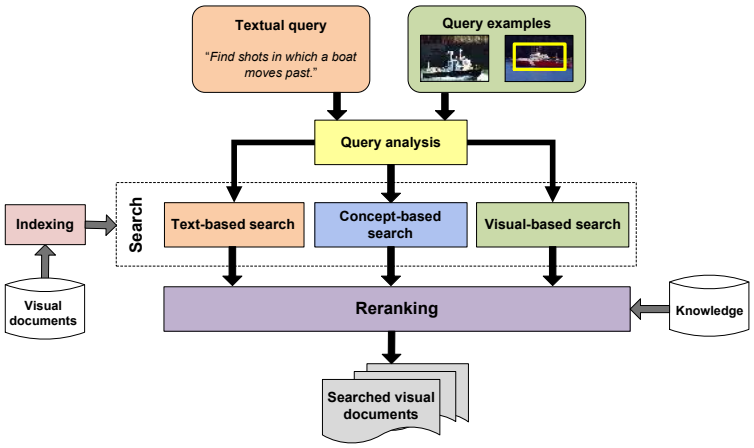

Fig. 3. A generic multimedia search framework.

concept-based search (e.g., boat, water, and outdoor). Specifically, the confidence scores from those classifiers can be treated as the weights for the corresponding concepts (i.e., hidden text), and further used in a text-alike search (e.g., inverted index based on term and document frequency) or used as a feature vector in a concept space for searching via query-by-example (QBE). Moreover, a set of low-level visual features (global and local features) is extracted to represent these query examples for visual-based search. These multimodal queries are fed into individual search models, such as text, concept, and visual-based searches, respectively. For example, a text-based search may use the speech recognition transcript, the closed caption available from the program channel, or the recognized captions embedded in video frames through Optical Character Recognition (OCR). The confidence vectors from concept detectors or low-level feature vectors can be used in same way as the QBE or tf-idf scheme [Baeza-Yates and Ribeiro-Neto 1999] for searching. More comprehensive introductions of content-based image and video retrieval can be found in [Datta et al. 2008], [Kennedy et al. 2008a], [Lew 2000], [Lew et al. 2006], [Smeulders et al. 2000], [Snoek and Worring 2009], [Yan and Hauptmann 2007b], [Philbin et al. 2007]. Based on these initial search results, as well as some knowledge, a re-ranking module is applied to aggregate the search results and reorder the initial document list to improve the search performance. In this paper, re-ranking is defined as improving the initial text-based search accuracy by considering the other search results and the prior knowledge. We can observe from Figure 3 that re-ranking plays a key in the visual search framework to improve the initial search performance.

2.2 A Bayesian View of Multimedia Search Re-ranking

Visual re-ranking problem can be traced back to the late 1990s when researchers started focusing on improving content-based image retrieval (CBIR) results via relevance feedback techniques [Benitez et al. 1998], [Rui et al. 1998], [Zhou and Huang 2002], [Zhou and Huang 2003]. It emerged as an independent research topic and attracted increasing intention beginning in the early 2000s. In the most common formulation, the re-ranking problem can be reduced to a problem of re-estimating relevance for each document that has been ranked in an initial search result (e.g., a ranked list of documents searched by a text-only approach, a ranked list of objects returned by a bag-of-words model, etc.). Intuitively, this estimation is usually based on some knowledge mined from the initial search results or the queries, prior knowledge from the Web, some domain-specific knowledge databases, or the interac-

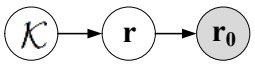

Fig. 4. Graphical model representation of $\mathcal{K}$, $\mathbf{r}_0$, and $\mathbf{r}$. The shaded node indicates the known values. $\mathbf{r}_0$ and $\mathcal{K}$ are assumed independent in this model.

tions of users. Once we can estimate the relevance of each document, we can re-rank these documents according to their relevance scores and obtain the best ranked list of these documents.

More formally, the re-ranking problem can be formulated as finding the optimal ranked list from the perspective of Bayesian theory as follows. Let $\mathcal{D}$ ($\mathcal{D} = \{d_i\}_{i=1}^{N}$) denote the collection of documents to be ranked or reranked, where d_i is a single document (such as an object, an image, a video, or a clip) and N is the number of documents in $\mathcal{D}$. Let $\mathbf{r}_0$ denote the initial ranked list and $\mathbf{r}$ the best ranked list. In these lists, each document d_i has a corresponding ranking order or relevance score r_i with respect to the query $\mathbf{q}$. Note that $\mathbf{q}$ may be a piece of terms, or visual examples (i.e., a set of objects/images or video clips), or any combination of them. Let $\mathcal{R}$ denote the set of all possible ranked lists ($\mathbf{r}_0$, $\mathbf{r} \in \mathcal{R}$) and $\mathbf{r} = [r_1, r_2, \ldots, r_N]^T$, where r_i ($0 \leqslant r_i \leqslant 1$) is the relevance score for the i-th visual document d_i. If we only consider the rank order of each document in the $\mathbf{r}$, then the space of $\mathcal{R}$ is $N!$, which can be huge if N is big enough. Let $\mathcal{K}$ denote the knowledge for guiding the re-ranking process. From a probabilistic view, given the initial ranked list $\mathbf{r}_0$ and prior knowledge $\mathcal{K}$, re-ranking can be formulated as to derive the optimal list $\mathbf{r}^*$ with the maximum a posterior probability,

$$\mathbf{r}^* = \arg\max_{\mathbf{r}\in\mathcal{R}} p(\mathbf{r}|\mathbf{r}_0, \mathcal{K}). \tag{1}$$

According to Bayes' formula, the posterior $p(\mathbf{r}|\mathbf{r}_0, \mathcal{K})$ is proportional to the product of the conditional prior probability and the likelihood,

$$p(\mathbf{r}|\mathbf{r}_0, \mathcal{K}) = \frac{p(\mathbf{r}|\mathcal{K})p(\mathbf{r}_0|\mathbf{r}, \mathcal{K})}{p(\mathbf{r}_0|\mathcal{K})} \propto p(\mathbf{r}|\mathcal{K})p(\mathbf{r}_0|\mathbf{r}, \mathcal{K}), \tag{2}$$

where $p(\mathbf{r}|\mathcal{K})$ indicates the conditional prior of the ranked list given the prior knowledge, and $p(\mathbf{r}_0|\mathbf{r}, \mathcal{K})$ indicates the likelihood of the initial ranked list given the "true" list. Intuitively, the conditional prior $p(\mathbf{r}|\mathcal{K})$ actually expresses how the "true" list $\mathbf{r}$ is consistent with the knowledge $\mathcal{K}$. In other words, the prior knowledge acts as the basis of re-ranking to ensure that there is the maximal consistency between the ranked list and itself. For example, $p(\mathbf{r}|\mathcal{K})$ can be modeled by the visual consistency between the reranked list and the knowledge in terms of some dominant patterns. The likelihood $p(\mathbf{r}_0|\mathbf{r}, \mathcal{K})$ expresses the probability that the initial ranked list aligns with the "true" list and the knowledge. For example, it can be modeled based on the disagreement between the reranked and initial lists. Note that $\mathbf{r}_0$ is given based on text-based search, thus we can assume that $\mathbf{r}_0$ and $\mathcal{K}$ are independent from each other [Tian et al. 2008]. Fig. 4 shows the graphic model representation of the relationship between $\mathcal{K}$, $\mathbf{r}$, and $\mathbf{r}_0$. Then we can rewrite it as:

$$p(\mathbf{r}_0|\mathbf{r}, \mathcal{K}) \triangleq p(\mathbf{r}_0|\mathbf{r}). \tag{3}$$

By plugging equation (3) into (2), we can obtain the formulation of re-ranking from the Bayesian perspective,

$$\mathbf{r}^* = \arg\max_{\mathbf{r}\in\mathcal{R}} p(\mathbf{r}|\mathcal{K})p(\mathbf{r}_0|\mathbf{r}). \tag{4}$$

Equation (4) indicates the optimization of the reranked list in terms of two somewhat conflicting objectives, i.e., maximizing the consistency to the knowledge and minimizing the disagreement with the initial ranked list. Thus, the central problem of re-ranking is the modeling of consistency between

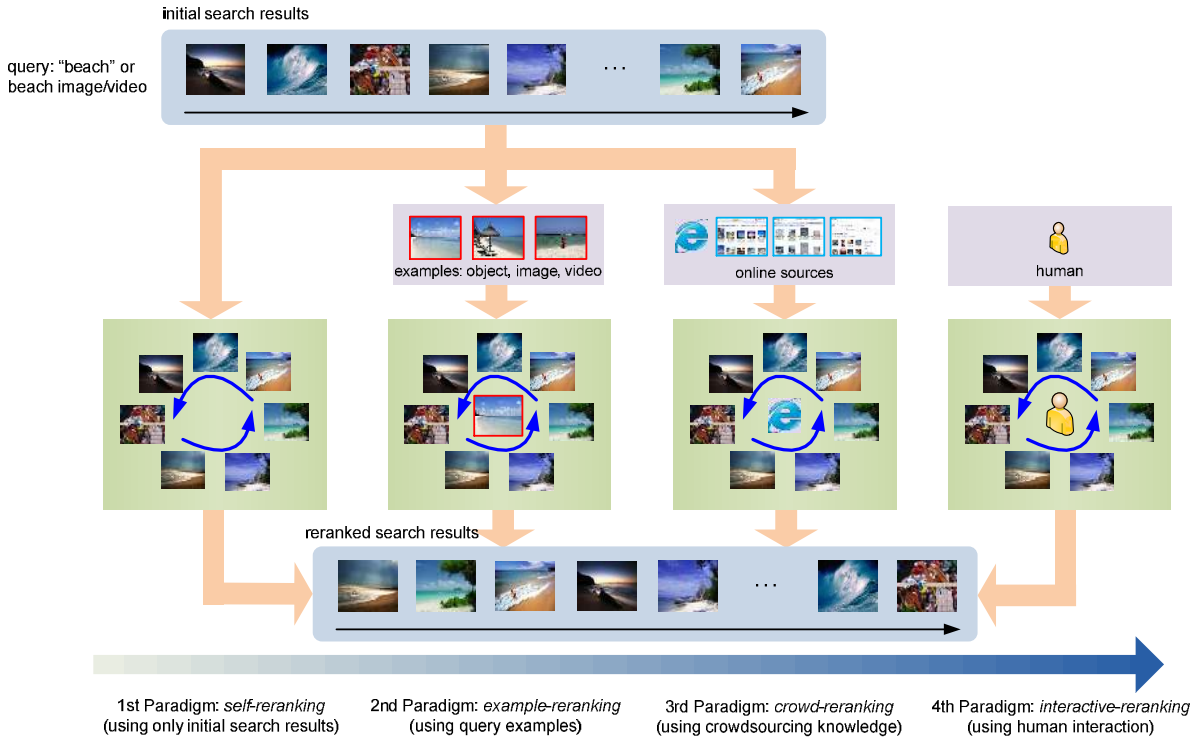

Fig. 5. Paradigms for multimedia search re-ranking.

$\mathbf{r}$ and $\mathcal{K}$, as well as the distance between $\mathbf{r}_0$ and $\mathbf{r}$. From the perspective of how the knowledge $\mathcal{K}$ is exploited, we can classify the approaches to re-ranking into the following four paradigms:

—*Self-re-ranking*, which focuses on mining the knowledge only from the initial ranked list $\mathbf{r}_0$, i.e., $\mathcal{K} = \phi$.

—*Example-based-re-ranking*, which leverages the user-provided visual query examples (i.e., objects in images, images, or video keyframes/clips) to detect the relevant patterns with respect to the given query $\mathbf{q}$, i.e., $\mathcal{K} = \mathbf{q}$.

—*Crowd-re-ranking*, which aims to mine the dominant visual patterns from the crowdsourced knowledge available on the Web or from domain-specific knowledge bases, i.e., $\mathcal{K}$ is mined from the knowledge on the Web.

—*Interactive-re-ranking*, which involves user interactions (i.e., human labor and feedbacks) to refine search results, i.e., $\mathcal{K}$ is mined from user interactions or human knowledge.

Figure 5 illustrates these paradigms for visual search re-ranking. In summary, most existing re-ranking approaches first mine a prior knowledge $\mathcal{K}$ (i.e., the dominant patterns which are relevant to the query), and then perform re-ranking based on three widely adopted assumptions: 1) the visual documents with the dominant patterns are expected to be ranked higher than others, 2) the visual documents with similar visual appearance are to be ranked closely, and 3) the top ranked documents in the initial list are expected to be ranked relatively higher than the other documents. We will introduce the methodologies for re-ranking in terms of the above four paradigms in the next section.

Table I. CATEGORIZATION OF RE-RANKING METHODOLOGIES AND REPRESENTATIVE APPROACHES

Paradigms	Representative approaches
Self-re-ranking	
– Clustering-based methods	Information Bottleneck Principle [Hsu et al. 2006]
– Pseudo relevance feedback	Pseudo-Relevance Feedback [Yan et al. 2003]
– Object recognition-based methods	Category filtering [Fergus et al. 2005], [Wnuk and Soattoh 2008]
– Graph-based methods	Random Walk [Hsu and Chang 2007], Visual Rank [Jing and Baluja 2008a], [Jing and Baluja 2008b]
Example-based-re-ranking	
– Concept-based methods	Text-like Multimedia Search [Li et al. 2007], Concept-based Fusion [Kennedy et al. 2008b]
– Linear multimodal fusion	Query-dependent Fusion [Chua et al. 2004], [Hauptmann et al. 2008b], [Yan et al. 2004]
– Query expansion	Total recall [Chum et al. 2011], [Chum et al. 2007]
– Geometric verification	Fast spatial matching [Philbin et al. 2007], [Jegou et al. 2010] Spatial coding [Zhou et al. 2010]
Crowd-re-ranking	Multiple Search Engines [Liu et al. 2009], Visual Query Suggestion [Zha et al. 2009]
Interactive-re-ranking	MediaMill [Snoek et al. 2006], CuZero [Zavesky and Chang: 2008] Color Map [Wang and Hua. 2011]

2.3 Methodologies for Multimedia Search Re-ranking

Table I summarizes algorithms and representative works for multimedia search re-ranking. In the next section, we will discuss the general approaches of each category. Please note that the sub-categories in one paradigm can still be applied to another paradigm.

2.3.1 *Self-re-ranking Methods.* In this paradigm, the re-ranking objective is to discover relevant visual patterns from the initial ranked list that can provide clues for re-ranking. Although they are quite noisy due to the unsatisfying text-only search performance, the initial search results, especially the top-ranked documents, can be regarded as the resource for mining some relevant information, since the analysis on click-through data from a very large search engine log shows that users are usually interested in the top-ranked portion of search results [Wei et al. 2009].

Based on how the relevant information is mined from the initial ranked list $\mathbf{r}_0$, the self-re-ranking methods can be further classified into the following categories:

—Clustering-based re-ranking [Ben-Haim et al. 2006], [Cai et al. 2004], [Hsu et al. 2006], [Jing et al. 2006], [Park et al. 2005], [Wang et al. 2007], [Wei et al. 2009]. The clustering-based methods assume that relevant documents tend to be more similar to each other than to irrelevant ones, so that clustering the retrieved visual documents may further separate the relevant from the irrelevant ones. The key problems are how to do clustering in the noisy initial ranked documents, as well as how to rank the clusters and the documents within each cluster.

—Pseudo relevance feedback (PRF) [Amir et al. 2005], [He et al. 2005], [Liu and Mei 2011], [Liu et al. 2008], [Liu et al. 2008], [Rudinac et al. 2009], [Yan et al. 2003]. 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 [Benitez et al. 1998], [Rui et al. 1998], [Zhou and Huang 2002], [Zhou and Huang 2003]. Those pseudo relevant samples can be further used in any learning method to classify the remaining documents into relevant or irrelevant classes, or be used as "query examples" to compute the distance to the remaining documents, or be the feedback to the system for query term re-weighting or re-formulation. Note that the assumption of pseudo relevance makes automatic re-ranking possible. The key problems include how to select the pseudo relevant

documents from the noisy initial ranked list and how to treat these documents for re-ranking. Note that we generalize the idea of PRF to any methods in which the top-ranked documents are regarded as "positive" in this paper.

—Object recognition-based re-ranking [Fergus et al. 2005], [Liu et al. 2009], [Wnuk and Soattoh 2008]. This kind of method is highly motivated from the success of object recognition in computer vision. The methods are more focused on queries that are related to object categories such as "car," "horse," "bottles," and so on. Observing that the visual documents related to these queries are typically visually similar, while those unrelated look different from each other, researchers in the computer vision community attempt to model this kind of visual consistency or object appearance with respect to the query or the initial ranked list, so that they can re-rank visual search results according to the fitness of the models (i.e., likelihood). Most of the models are generative and probabilistic, characterized by the scale and orientation invariance, as well as the simultaneous consideration of shape, appearance, and spatial layout. However, it is significantly different from the classical setting of visual recognition where there is usually a clear training set consisting of carefully labeled "positive" and "negative" examples. The challenges lie in that: 1) the training set (i.e., the initial ranked list) is not labeled, and moreover, only contains a minority of "good" examples; 2) the modeling task is to sort the "training" set rather than to classify the fresh "testing" data; and 3) the model has to deal with the heterogeneous features since even "good" visual documents in the training set have high visual variance.

—Graph-based re-ranking [Hsu and Chang 2007], [Jing and Baluja 2008a], [Jing and Baluja 2008b], [Liu et al. 2007], [Zitouni et al. 2008]. The methods in this category are highly motivated by the well-known PageRank technique for text search [Brin and Page 1998a], in which the relevance of a document is propagated throughout the link structure among a large number of documents. A graph $\mathcal{G} =< V, E >$ can be built over the initial ranked list, in which each node v ($v \in V$) corresponds a visual document, and the edge e ($e \in E$) corresponds to the multimodal similarities between two documents. 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 two objectives in equation (4): maximizing a global consistency $p(\mathbf{r}|\mathcal{K})$ and minimizing a distance $p(\mathbf{r}_0|\mathbf{r})$ to compromise the reranked list to the initial one. Usually, after differentiating and simplifying, this optimization problem can be solved by some close-form solution or in an iterative way. In fact, graph-based methods can be viewed as a non-linear fusion of heterogeneous ranked lists. The graph representation includes PageRank [Jing and Baluja 2008a], [Jing and Baluja 2008b], [Liu et al. 2007], [Wang et al. 2009], [Zitouni et al. 2008], Random Walk [Hsu and Chang 2007], [Hsu et al. 2007], Bayesian formulation [Tian et al. 2008], and multi-level graph [Hoi and Lyu 2007].

Different self-re-ranking approaches may have different conditions to work. In general, self-re-ranking highly depends on the initial search results, since the only information for mining relevant visual pattern $\mathcal{K}$ is the initial ranked list or the top of this list. In other words, self-re-ranking may not work well or even downgrade the performance if the initial search results are not relevant at all. Within the self-re-ranking paradigm, clustering-based approaches are characterized by their high effectiveness, as they are very intuitive and we only need to conduct several rounds of clustering within a part of the initial ranked list. PRF-based approaches are more expensive than clustering-based approaches, as they need to build the query-dependent classification or ranking models on the fly [Jain and Varma 2011]. As a result, PRF-based re-ranking is not very practical for real-time requirement. Researchers are now investigating learning-to-rank methods to build query-independent PRF re-ranking schemes

[Liu 2009]. Moreover, PRF-based approaches highly depend on the performance of the initial search, since they assume that the top ranked documents are more relevant than those ranked in the bottom. Object recognition-based re-ranking approaches are related to the performance of object recognition, as well as how the query can be represented or related to the recognized object categories. They are not that expensive as PRF-based approaches as the prediction of object categories in object recognition-based approaches are faster than building ranking models in PRF. But still, the approaches need to conduct feature extraction on the fly or access the features that might be extracted offline and stored in backend files. The graph-based re-ranking methods are highly effective as the graph can be built offline only once if the graph includes all the documents to be searched. Then, the re-ranking problem can be transferred into the traditional "PageRank" framework and efficiently solved in an iterative way.

Note that although we have the above categories, one single paper may use re-ranking methods from more than two categories. In the next section, we introduce representative works for each category.

1) **Clustering-based methods.**

Hsu *et al.* propose to first rank image clusters and then rank the images within each cluster. They first obtain the optimal clustering of the top text-based search results by preserving the maximal mutual information about the search relevance, and then order the images within each cluster by the local feature density [Hsu et al. 2006]. The visually consistent images, which occur more frequently within the clusters with higher relevance, will be reranked higher. The approach first estimates the soft-pseudo-relevance label y of an image x, denoted by the post-probability of relevance $p(y|x)$. The top ranked images, together with a set of sampled negative images, are used to compute the joint probability $p(x, y)$ in a high dimensional feature space. Then, the Information Bottleneck (IB) principle is employed to cluster both the top-ranked and negative images based on the joint probability. Finally, the cluster c is ranked according to the conditional probability $p(y|c)$ and the images within each cluster c are reranked according to the feature density $p(x|c)$. Similarly, the approach in [Ben-Haim et al. 2006] first segments each image in the search results into several regions, clusters these regions based on the color histogram representation and the mean shift algorithm, and then detects the "significant" cluster (with the largest number of regions). The similarity between each image with the "significant" cluster is used as a re-ranking signal. In [Park et al. 2005], the Hierarchical Agglomerative Clustering (HAC) algorithm is used to cluster the text-based image search results. Then images are reranked according to the distance of a cluster from a query.

Jing *et al.* employ image clustering techniques to identify semantically relevant clusters to a query, and design an efficient user interface for browsing image search results by considering both the image and textual (i.e., title) thumbnails for visual representation [Jing et al. 2006], [Wang et al. 2007]. Cai *et al.* incorporate a vision-based page segmentation algorithm to partition a web page (usually containing images) into blocks, and then represent the web images by using visual, textual, and link information based on the block-level link analysis [Cai et al. 2004]. Spectral techniques are applied to cluster the search results into different semantic categories, in which several images are selected as representative images for quick browsing.

2) **Pseudo relevance feedback methods.**

In [Amir et al. 2005], [Yan et al. 2003], the pseudo-negative images are sampled from the lowest rank of the initial text-based search results, while the query videos and images are taken as the positive examples. The re-ranking is then formulated as a classification problem where multiple discriminative classifiers are trained with these pseudo- negative and positive examples. The visual documents are finally ordered according to the confidence scores output from the classifiers. This approach improves the search performance by 7.5% gain in terms of MAP in TRECVID 2003 [Yan et al. 2003]. The assumption of conventional pseudo-relevance feedback that most top-ranked documents are relevant to the

given query was relaxed in [Yan and Hauptmann 2007a], where it only required the top-ranked documents contain more relevance documents than the bottom-ranked document. The authors proposed a probabilistic local feedback (PLF) model based on a discriminative probabilistic retrieval framework.

Liu *et al.* claim that the best ranked list cannot be obtained until any two arbitrary documents from the list are correctly ranked in terms of relevance [Liu et al. 2008]. This is different from conventional ranking problem where a document is classified as relevant or not independently. They first cluster the initial search results. Then, they propose to incrementally discover the so-called "pseudo preference pairs" from the initial search results by considering both the cluster typicality and the local typicality within the cluster. Here, *typicality* (i.e., the visual representativeness of a visual document with respect to a query) is a higher-level definition than *relevance*. For example, an image with a "boat" may be relevant to the query "find the images with boat," but may not be typical as the boat object is quite small in the image. The Ranking Support Vector Machine (RSVM) is then employed to perform pairwise classification [Herbrich et al. 2000]. The documents are finally reranked by predicting the probabilities of the RSVM. In [Liu et al. 2008], the optimal pairs are identified purely based on the low-level visual features from the initial search results. Later, the authors observe that leveraging concept detectors to associate a set of relevant high-level concepts to each document will improve the discovery of the optimal pairs [Liu and Mei 2011], [Liu et al. 2008].

Rudinac *et al.* further incorporate the visual representativeness of the returned documents to the conventional PRF [Rudinac et al. 2009]. They propose the Average Item Distance (AID) to measure the visual representativeness. Intuitively, within the top returned results in the conventional PRF, the documents that best typify the initial search results (i.e., with high AID scores) will be reranked higher. He *et al.* use the multi-view learning to identify the relevant documents by combining the results from two complementary yet independent learners [He et al. 2005]

3) **Object recognition-based re-ranking.**

Wnuk *et al.* propose an approach to filtering the "strange" images from the noisy text-based search results by considering the visual consistency [Wnuk and Soattoh 2008]. They claim that the remaining images after filtering could be used for building models for object recognition, and further for reranking the initial search results. Similarly, the works in [Fergus et al. 2005], [Liu et al. 2009] find that it is reasonable to learn the object category models from the noisy image research results by exploiting the visual consistency in a unsupervised or semi-supervised way.

4) **Graph-based re-ranking.**

Zitouni *et al.* find that only a subset of text-based image search results contains relevance images and this subset usually forms a dense component in a full-connected graph [Zitouni et al. 2008]. Based on this observation, they present the similarities of the top-ranked images (based on the local descriptors) in a graph structure, find the densest component in the graph, and then assign higher ranks to the images in the densest component while low ranks to the others.

Jing *et al.* apply the PageRank to product image search and designed the VisualRank algorithm for ranking/re-ranking [Jing and Baluja 2008a], [Jing and Baluja 2008b]. The VisualRank employs the Random Walk intuition to rank images according to the visual hyperlinks among images. Intuitively, if a user is viewing an image, and there are other (visually similar) images, then there is a probability that the user will jump from this image to another similar one. This is analogous to PageRank where the importance of a web page is usually measured based on the link structure [Brin and Page 1998b], [Brinkmeier 2006]. In the VisualRank, the ranking score $\mathbf{r}$ is defined as follows:

$$\mathbf{r} = d\mathbf{S} \times \mathbf{r} + (1-d)\mathbf{p}, \text{ where } \mathbf{p} = \left[\frac{1}{N}\right]_{N\times 1}, \tag{5}$$

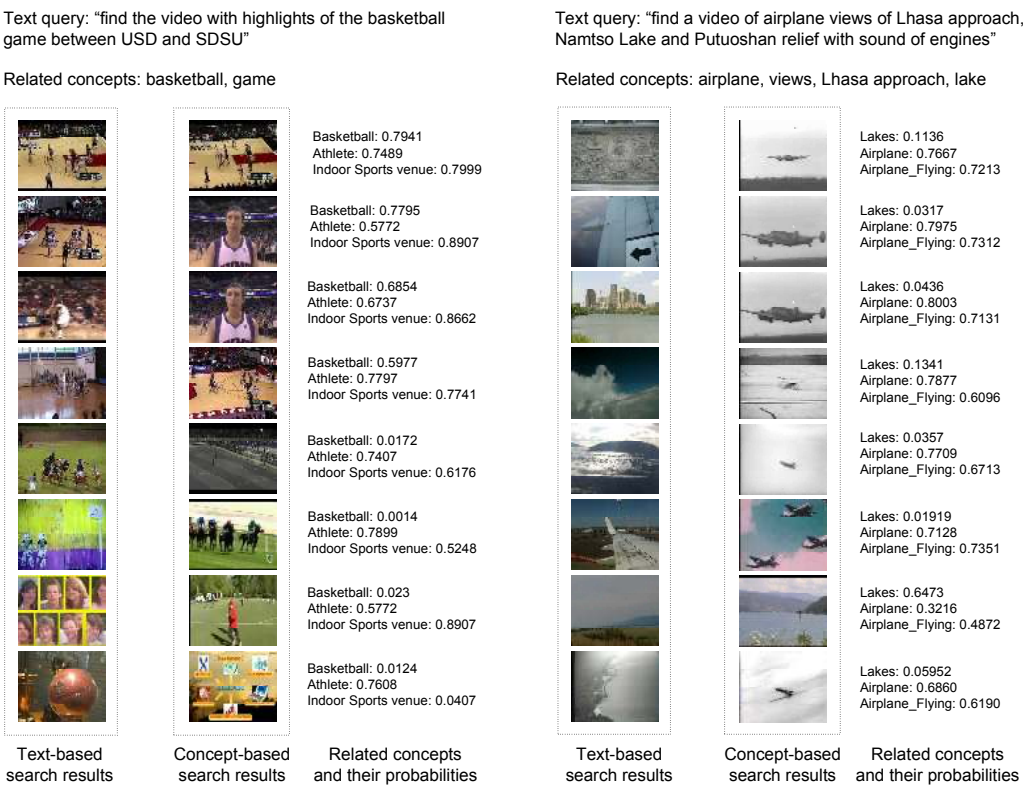

Fig. 6. Examples of search results by text and concept based approaches. The result examples are from TRECVID 2011 [Ngo et al. 2011].

where $\mathbf{S}$ is the normalized adjacency matrix whose element S_{ij} measures the visual similarity between a pair of images (i, j), $\mathbf{p}$ is called the personalization value for each image, and d is a damping factor for creating a small probability for a random walk traveling to some other images in the graph. $\mathbf{p}$ can be set equally for all images or set as the initial ranking scores $\mathbf{r}_0$. The VisualRank was evaluated on a collection of the most popular $2,000$ product image queries, and it demonstrated superior performance to Google image search results.

To reveal the hidden links between video stories, video re-ranking is formulated as a random walk over a context graph, where the nodes represent the video documents, connected by the edges weighted with multimodal similarities (i.e., the textual and visual similarities) [Hsu and Chang 2007], [Wang et al. 2009]. The random walk is biased with the preference towards the initial text search results, while its stationary probability is regarded as the final relevance score for re-ranking. Similar ideas can be found in [Liu et al. 2007], where the video search re-ranking problem is investigated in a “PageRank” fashion. The video shots are taken as web pages, while the multimodal similarities between the video shots are taken as hyperlinks. The topic-sensitive PageRank algorithm is then employed to propagate the relevance score of video shots through these hyperlinks [Haveliwala 2002].

2.3.2 *Example-based-re-ranking Methods.* The *self-re-ranking* methods provide a straightforward yet efficient way for re-ranking as they do not require any other information besides the textual query and the initial search results. However, it is well known that text-based multimedia search engines

do not always have satisfying performance, mainly because of the noisy or missing surrounding texts. The initial ranked list usually cannot provide enough cues to detect recurrent patterns for re-ranking. To address this issue, the second paradigm, i.e., *example-based-re-ranking* leverages a few query examples (e.g., a set of objects, images, or video clips), which are provided along with the textual query by the user, to train the re-ranking models. This is a typical setting in traditional content-based visual retrieval where a user is required to provide a visual example [Chang et al. 2007], [Datta et al. 2008], [Hua and Tian 2009], [Philbin et al. 2007]. The search performance can be improved due to the relevant information derived from these query examples.

There are many ways of using these query examples. The early research in CBIR has focused on the problem of query-by-example for decades [Datta et al. 2008], [Lew et al. 2006], [Rui et al. 1999], [Smeulders et al. 2000], where the key issues are low-level feature representation for image content and distance metric for computing image similarities [Mei and Rui 2009]. Intuitively, the documents that are visually similar to the query examples will be ranked higher than others [Zloof 1975a]. Another method is to represent query examples using a set of or a distribution of predefined high-level concepts. The concepts can be automatically discovered by model-based classifiers. For example, a query image with the "Great Wall" could be associated with the concept "mountain," "sky," "China," and their corresponding probabilities (e.g., "mountain" = 0.65, "sky" = 0.30) [Natsev et al. 2007], [Snoek and Worring 2009]. The associated concepts are also called attributes in the vision community [Li et al. 2010], [Li et al. 2010]. Then, the related concepts are used to either expand the initial textual query (by adding the concepts as query terms) or re-rank the visual documents based on the concept-based distance. Once we have multiple reranked lists from different re-ranking methods, we can use the simple linear fusion (either query-independent or query-dependent) or non-linear fusion to combine these results. Figure 6 shows the search result examples from a text and concept based approach based on TRECVID data. It shows that by using concept detectors on the query examples, the latent semantics can be mined to further boost the search results.

Another example is object retrieval, where a user provides an query image or indicates a region-of-interest in a query image, and the search system returns the images that are similar to the query image or contain the same objects (from different views) [Philbin et al. 2007], [Jegou et al. 2010], [Turcot and Lowe 2009]. A geometric verification process is introduced to re-rank the returned visual documents by checking the spatial constraints between the query object/image and each returned result. Alternatively, the top returned results can be leveraged to build a latent feature model which can be further used to expand existing queries for re-ranking [Chum et al. 2007], [Chum et al. 2011]. Example-based object retrieval is a typical setting in the computer vision community. There is a great deal of research on re-ranking query-example-based object retrieval results by using geometric verification, codebook learning, query expansion, and so on. However, as this is not the focus of this survey, we only briefly introduce the typical methods for re-ranking object retrieval results.

We now introduce these methods in the *example-based-re-ranking* paradigm.

—Concept-based re-ranking. Using many intermediate semantic concepts has the potential to bridge the semantic gap between what a low-level feature-based image analysis can extract and what users really want to find (using a piece of text description) [Hauptmann et al. 2008b], [Hauptmann et al. 2007], [Kennedy and Chang 2007], [Li et al. 2007], [Naphade et al. 2006], [Ngo 2009]. Meanwhile, the rich set of predefined concepts and their corresponding training and testing samples available in the community has made it possible to explore the semantic description of a query in a large concept space. For example, 101 concepts are defined in MediaMill [Snoek et al. 2006], 374 in LSCOM-Light [Yanagawa et al. 2007], 834 in LSCOM [Naphade et al. 2006], 17,624 in ImageNet [Deng et al. 2009], [ImageNet], and so on. The basic idea of concept-based methods is to utilize the results from

concept detection to aid search, thereby leveraging human annotation on a finite concept lexicon to help answer infinite search queries. For example, a search query like "Find shots of boats" could be handled by searching against the transcript to find occurrencec of "boat," but also by giving positive weight to shots that are positive for the concepts of "boat" and "water" (since "boat" and "water" are highly correlated concepts) and negative for "indoor." This is similar to traditional query expansion in the information retrieval, where a textual query is expanded to associate with more words that are descriptive. The key problems here are how to select relevant concepts from a predefined lexicon with hundreds of concepts and how to leverage these concept detectors for re-ranking. The challenge arises from the very limited performance of concept detection. For example, the best system for high-level feature extraction only achieved an accuracy of less than 23% in terms of MAP over 20 concepts in TRECVID 2009 [TRECVID] [3]. However, it is found that even with such kind of accuracy from concept detection, we can still achieve "good" search results using a few thousand concepts [Hauptmann et al. 2008b], [Hauptmann et al. 2007]. For example, Hauptmann *et al.* conclude that a concept-based video retrieval system with fewer than 5,000 concepts (detected with a minimal accuracy of 10% MAP) is likely to provide high accuracy search results in the broad-cast news domain (e.g., 37% MAP when using the full LSCOM set of concepts [Naphade et al. 2006]) [Hauptmann et al. 2007]. They also suggest a way to select "helpful" concepts based on the mutual information measure. Ngo *et al.* present a concept-driven multimodality fusion approach in their automatic video search system [Ngo 2009]. They first generate a set of concepts for a given query. To obtain the optimal weight for combining the search results based on each concept, they conduct a simulated search evaluation, in which a concept is treated as a simulated query associated with concepts and ten randomly chosen positive visual samples. Then, the uni-modal search performance for the concept and its related visual samples against a training dataset is manually labeled. With the simulated search evaluation, given a testing query, they estimate the concept-based fusion weights by jointly considering query-concept relatedness and the simulated search performance of all concepts. For a comprehensive survey of concept-based video search, please refer to [Snoek and Worring 2009].

—Linear multimodal fusion (LMF). This is the most straightforward and easy-to-implement method for re-ranking. Suppose we have multiple ranked lists in addition to the initial text-only list, each of which can be obtained by any single modality-based method introduced in Section 2.3. For example, Figure 3 illustrates three kinds of search results from text, visual, and concept modalities. To fuse different ranked lists, we can linearly combine the relevance scores for each document and re-rank the documents according to their combined scores, i.e.,

$$\mathbf{r} = \sum\nolimits_i \omega_i \mathbf{r}_i, \tag{6}$$

where $\mathbf{r}_i$ is the i-th ranked list, and ω_i $(0 \leqslant \omega_i \leqslant 1)$ is the weight for fusion. The key problem is then how to decide the weights for fusion. The weights could be obtained based on query-independent rules [4] (e.g., *average*, *max*, *min*, *Borda count* operations, or learning-based) [Chua et al. 2004], [Haubold et al. 2006], [Mei et al. 2007], [Snoek et al. 2006], [Worring et al. 2007], [Yan and Hauptmann 2004], or query-dependent analysis (e.g., cross-validation and learning approaches, such as logistic regression, boosting, etc.) [Chang et al. 2006], [Chua et al. 2004], [Haubold et al. 2006], [Kennedy et al. 2008b], [Liu et al. 2008], [Mei et al. 2007], [Natsev et al. 2007], [R.Yan and Hauptmann 2003], [Wang et al. 2009], [Wilkins et al. 2006], [Wilkins et al. 2010], [Yan and Hauptmann 2006], [Yan and Hauptmann 2007a], [Yan et al. 2004]. The query-independent approaches might have limited effectiveness because the optimal combination strategies usually vary considerably for

[3] MAP: mean average precision.
[4] Linear query-independent fusion is a popular fusion strategy in many notepapers in TRECVID

different query topics, while query-dependent approaches map the query space into a mixture of predefined query classes and learn the best combination strategy for each query class. In the latter, the weight $\omega_i(\mathbf{q})$ for each ranking feature is related to the associated query classes or the ranking score distribution of the documents. The problem in the query-dependent approach is then how to obtain the optimal weights $\omega_i(\mathbf{q})$ with respect to query $\mathbf{q}$. Kennedy *et al.* review different query-dependent fusion methods for multimedia search [Kennedy et al. 2008b]. As discussed in [Snoek et al. 2005], there are *early* and *late* fusion schemes for semantic video analysis, where the *early* fusion occurs within a single modality before the stage of concept learning, while the *late* fusion occurs in a later stage of combining across individual modalities. As most linear fusion schemes for re-ranking focus on the combination of different search results (which usually come from a single modality), the LMF discussed in this paper could be regarded as a *late* fusion.

—Geometric verification (GV). A naive and inefficient solution to example-based visual search, in the early years of 1990s, would be to formulate a ranking function based on some visual features and distance metrics, and apply it to every visual document in the dataset to return a ranked list. This is very computationally expensive as the complexity is linear to the size of the whole corpora. The standard method in text retrieval is to use a bag-of-words (BoW) model, which is efficiently implemented as an inverted file structure [Baeza-Yates and Ribeiro-Neto 1999]. Recent research in visual search has mimicked simple text retrieval by using the analogy of "visual words" [Sivic and Zisserman 2003], [Nister and Stewenius 2006]. Images are scanned for "salient" regions and a high-dimensional descriptor, for example, the Scale Invariant Feature Transform (SIFT) [Lowe 2004], is computed for each region. These descriptors are quantized or clustered into a vocabulary of visual words with the k-means algorithm. An image is then represented by the frequency histogram of visual words (i.e., bag-of-visual-words) obtained by assigning each descriptor of the image to the closest visual word. The visual words are indexed in an inverted file similar with the tf-idf in text retrieval. Although the above bag-of-visual-words model for example-based visual search is effective, a re-ranking procedure is vital for improving the search performance by the nature of a query image: 1) an image query is not like a textual query in that the image query is usually much longer (high-dimensional feature vector) and much noisier, and 2) an image query has its own distinct characteristics, for example, spatial structure information. When following the bag-of-visual-words model, the search performance is significantly improved if a post-processing with geometric verification is introduced after the first round of retrieval [Philbin et al. 2007], [Jegou et al. 2010], [Turcot and Lowe 2009]. The geometric verification step ensures that the top returned images and the query not only contain similar visual features, but also that the features occur in compatible spatial configurations. This is usually conducted only on a subset of the initial ranked list, e.g., the top 1,000 results.

—Query expansion (QE). In text retrieval, a well-known technique is query expansion, in which a new query is generated by using the highly ranked documents [Baeza-Yates and Ribeiro-Neto 1999]. This is similar with pseudo relevant feedback introduced in the self-re-ranking paradigm. In image retrieval, query expansion is useful as the visual words used to index images are a synthetic projection from a high-dimensional descriptor space and therefore suffer from substantial noise and drop-outs. Using query expansion, a laten model can be built to form the union of features common to a transitive closure of the returned images [Chum et al. 2011], [Chum et al. 2007]. In general, query expansion in multimedia search re-ranking works in the following way: 1) given a query image or object, search the image database and retrieve a set of image regions that match the query, 2) combine the retrieved regions, along with the original query, to form a richer latent model of the object, 3) re-query the dataset using this expanded model to retrieve an expanded set of matching regions, and 4) repeat this process as necessary.

Note that although concept-based methods for re-ranking do not solely rely on the initial search results as they leverage concept detectors which involve human labor for annotating and classifiers for training and predicting, we still list them in the *self-re-ranking* paradigm. This is because: 1) the concept detectors share the same dataset as search and re-ranking, and 2) this kind of knowledge (i.e., concept detectors) is carefully excerpted by a training process from a relatively clean dataset, which is very different from the so-called crowdsourced and noisy knowledge on the Web. Therefore, it is more reasonable to categorize this kind of method into *self-re-ranking*. A survey on how concepts can be used in video retrieval can be found in [Hauptmann et al. 2008b].

The concept-based re-ranking approaches highly depend on the performance of concept detection, which is still a challenging problem in both the multimedia and computer vision communities. Furthermore, the concept lexicon is usually fixed and thus not scalable to unlimited concepts. These approaches need the concept models built from a large-scale of training data. Because of these reasons, such a model-based re-ranking scheme is limited to very few query words and still in its infancy. Geometric verification, however, is highly effective and has been adopted in many commercial search engines. Spatial verification based on the k-d tree structure and the high ranked documents is fast and easy to be implemented in parallel. This scheme is in particular suitable for searching images with strong spatial configuration, such as landmarks and building facades. Query expansion methods are more expensive than geometric verification and can be used as a post-processing step to further improve the re-ranking performance.

We now introduce some representative research for each category.

1) **Concept-based re-ranking.**

Li *et al.* have shown that, when provided with a visual query example, searching through the concept space is a good complement to searching in the text and low-level feature spaces [Li et al. 2007]. They first build a concept space (with 311 concepts) over the whole dataset, where each document is associated with multiple relevant concepts (called "visual terms"). Given a query, they employ concept detectors over the query example to obtain the presence of concepts, and then adopt *c-tf-idf*, a *tf-idf*-like scheme, to measure the informativeness of the concepts to the query. The *c-tf-idf* is used in a traditional text-based search pipeline, e.g., a vector model or a language model [Baeza-Yates and Ribeiro-Neto 1999], [Salton et al. 1975], to measure the relevance between the given query and a document. These concept-based search results are finally combined with those from other modalities (e.g., text and visual) in a linear way.

Kennedy *et al.* leverage a large pool of 374 concept detectors for unsupervised search re-ranking [Kennedy and Chang 2007]. Each document in the database is represented by a vector of concept confidence scores by running the 374 concept detectors. Given a query, the initial search results are used to discover the pseudo-positive concepts (corresponding to the high-scored documents) and pseudo-negative concepts (corresponding to the randomly sampled documents). The key problem is to find a subset of concepts in the lexicon that have a strong relationship with the query. They employ the mutual information criteria to select the related concepts. Then, they form a feature vector for each document consisting of the related concept confidence scores, and train an SVM model based on the pseudo- positive and negative samples from the initial search results. The SVM testing results are finally combined with the initial ranking scores on average to re-rank the documents. They achieve an improvement of 15%–30% in MAP on the TRECVID 2005/6 search tasks. This approach is similar to PRF, except that the document is represented by the semantic concepts rather than the low level-features in PRF.

2) **Linear multimodal fusion for re-ranking.**

As recent query-dependent fusion methods consistently outperform query-independent methods in recent years' video search task in TRECVID [Hauptmann et al. 2008b], we only present exemplary

systems based on query-dependent re-ranking methods. In the video search system developed by NUS [Chua et al. 2004] and CMU [Hauptmann et al. 2004], [Yan et al. 2004], a textual query is first classified into a set of predefined query classes (i.e., "Person," "Sports," "Finance," "Weather," "Disaster," and "General" in [Chua et al. 2004]; "Named Persons," "Named Object," "General Object," "Scene," and "Sports" in [Hauptmann et al. 2004], [Yan et al. 2004]) according to some rule-based classification or machine learning techniques based on the automatically extracted entities. As a result, a query like "find shots of one or more buildings with flood waters around it/them" will be classified as "Disaster" category. Then, different combination weights are employed for different query classes in the raranking process. In the re-ranking for the above query in the "Disaster" category, the weights of the visual concept detectors such as "water-body" and "fire" will be given with higher weights.

Similarly, the automatic video retrieval system developed by IBM adopts query-class dependent and query-component-dependent weighting schemes [Ebadollahi et al. 2006]. The former scheme first assigns each textual query into one of seven predefined classes, and then optimally sets the linear fusion weights within each class by a cross-validation, i.e., the weights for each class are taken as the set that maximized the average performance for all training queries in this class. The query-component dependent scheme extends the former by allowing overlap in the seven query classes. The weights are similarly learned over the set of training queries with each component by maximizing the average performance. Note that the unimodal retrieval results in equation (5) are obtained from text, concept, content, and visual modalities separately in [Ebadollahi et al. 2006]. They find that the query-class dependent and query-component-dependent schemes are able to yield 14% and 13% improvement from query-independent fusion, respectively.

Donald *et al.* comprehensively study the linear fusion strategies for video shot retrieval [Donald and Smeaton 2005]. They combine the search results for the text and visual features, using variations of fusion methods (e.g., *sum*, *max*, *min*, *weighted*). Through evaluations on TRECVID 2002–2004 data, they observe that 1) simply adding the normalized relevance scores r_i of the top searched results can consistently achieve the best performance for combining a single visual feature over multiple examples, and 2) using weighted linear fusion is the best for combining text and visual results. They also suggest that for multi-query-example multi-feature multimedia search, features should first be fused for the examples and then the scores from these features can be fused linearly.

One of the key problems in re-ranking is how to make use of the ranking features that have no explicit correlation with the textual query. For example, if we provide a query "finding a building" to a system with 10 different ranking features available (such as "outdoor," "building," "indoor," "people," and so on), initially, the system only accounts for the concept of "building" while neglecting the other concepts. The probabilistic local feedback (PLF) in [Hauptmann et al. 2008b], [Yan and Hauptmann 2006], [Yan and Hauptmann 2007a] is a discriminative model for combining multiple search results. It automatically expands additional ranking features that are closely related to the original query by treating the weights of unweighted ranking features as latent variables rather than simply setting them to zero.

Specifically, given a query $\mathbf{q}$ provided by users, let $y_i \in \{+1, -1\}$ indicate if the document d_i is relevant or not, and $f_j(d_i)$ indicate a bag of ranking features or ranked results. For example, $f_j(d_i)$ can be the unimodal search results of text or visual modalities, or the detection results of predefined concepts. The ranking problem is to estimate a posterior probability of the relevance $p(y_i|d_i, \mathbf{q})$. This probability is modeled as a logistic function on a linear combination of ranking features in [Hauptmann et al. 2008b], [Yan and Hauptmann 2006], [Yan and Hauptmann 2007a], as follows:

$$p(\mathbf{y}|\omega, \mathcal{D}) \propto \prod_{i=1}^{N} \exp\left(y_i \sum_{j=1}^{N_j} \omega_j f_j(d_i)\right) \tag{7}$$

where N_j is the number of ranking features and ω_j is the linear weights for fusion. We drop the query $\mathbf{q}$ in the above equation as it is always given. However, we only have the initial weight ω_j for the known ranking features. To introduce the unweighted or unknown ranking features, a new latent weight $\mathbf{v}$ ($\mathbf{v} = \{v_\ell\}$) is introduced for each unweighted ranking features (e.g., concept or textual searches):

$$p(\mathbf{y}, \mathbf{v}|\omega, \mathcal{D}) \propto \prod_{\ell} p_0(v_\ell) \prod_{i=1}^{N} \exp\Big(y_i \sum_{j \in W} \omega_j f_j(d_i) + y_i \sum_{\ell \in U} v_\ell f_j(d_i)\Big) \tag{8}$$

where v_ℓ is the latent fusion weight for the ℓ-th unweighted concept, $W = \{j : \omega_j \neq 0\}$ contains the indexes of initial weighted concepts and $U = \{j : \omega_j \neq 0\}$ contains those of unweighted concepts, and $p_0(v_\ell)$ represents how likely an unweighted ranking feature is relevant. The probability in equation (8) can be practically inferred by the mean field approximation in an iterative way [Hauptmann et al. 2008b], [Yan and Hauptmann 2007a].

2.3.3 *Crowd-re-ranking Methods.* In contrast to the *self-re-ranking* and *example-based-re-ranking* paradigms, *crowd-re-ranking* methods are characterized by mining relevant patterns from the crowdsourced knowledge available on the Web, e.g., the common patterns mined from the image search results available from multiple search engines [Liu et al. 2009], the object model learned from the image search results from any existing search engine [Fergus et al. 2004], [Olivares et al. 2008], [Wang and Forsyth 2008], and the suggested queries augmented from the image collection on the Web [Zha et al. 2009], [Zha et al. 2010]. The motivations for using the crowdsourced knowledge for re-ranking are that the rich information from the crowdsourced data that are relevant to the given query can inform or complement to the information mined from the initial search results. Thus, more relevant information or common patterns can be discovered from the crowdsourced knowledge.

The CrowdReranking mines relevant visual patterns from image search results of multiple search engines, rather than only from the initial search result [Liu et al. 2009]. The principles behind CrowdReranking are that: 1) Using search results from different search engines can inform and complement the relevant visual information for each other, since they might have different data sources and indexing/ranking approaches; and 2) there are common visual patterns across different search results for a given query. Therefore, the basis of CrowdReranking is then to find the representative visual patterns, as well as their relations in multiple search results. First, a textual query is fed into multiple search engines to obtain lists of initial search results. Then, the representative visual words are constructed based on the local image patches from these search results. There are two explicit visual patterns detected from the visual words through a graph propagation process, i.e., *salient* and *concurrent* patterns. The former pattern indicates the importance of each visual word, while the latter expresses the interdependence among those visual words. Intuitively, a visual word with a strong *salient* pattern for a given query indicates that other concurring words (i.e., with strong *concurrent* patterns) would be prioritized. The re-ranking is then formalized as an optimization problem on the basis of the mined visual patterns and the Bag-of-Words (BoW) representation of the initial ranked list. The objective is to maximize the consistence $Cons(\mathbf{r}, \mathcal{K})$ between the learned visual patterns $\mathcal{K}$ and the reranked list $\mathbf{r}$, as well as minimizing the disagreement $Dist(\mathbf{r}_0, \mathbf{r})$ between the initial ranked list $\mathbf{r}_0$ and the reranked list $\mathbf{r}$ as follows

$$\mathbf{r}^* = \arg\min_{\mathbf{r}} \big\{ Dist(\mathbf{r}_0, \mathbf{r}) - \lambda Cons(\mathbf{r}, \mathcal{K}) \big\} \tag{9}$$

where λ tunes the contribution of the learned knowledge $\mathcal{K}$ to the reranked list. The distance function could be formalized as either pointwise or pairwise distance, while the consistence is defined as the

Fig. 7. Visual query suggestion for re-ranking [Zha et al. 2009] ©ACM 2010.

cosine similarity between a document and the mined visual patterns. They showed the improvement of 16% in terms of $NDCG@1$ in the MSRA-MM dataset [Li et al. 2009] [5].

The computer vision researcher leverage the visual search results available from existing commercial search engines to learn relevant object models for re-ranking. Fergus *et al.* use the Google image search results as pseudo-positives and utilized a parts-based approach to learn the object model, and then re-rank the search baseline images based on that model [Fergus et al. 2005], [Fergus et al. 2004]. This is similar to the PRF. Their experiment was limited to specific queries of simple objects such as bottles, cars, that is, instead of natural language queries as those in TRECVID and general web search.

Another crowd-re-ranking approach is query expansion. Query expansion is a popular approach in information retrieval (IR), which reformulates a textual query by introducing more relevant query terms or associated examples to the initial query to match additional documents. The automatically discovered new query terms or examples are obtained from the initial search results or the entire document collection. This re-ranking method is similar to object recognition-based re-ranking, except that: 1) there is usually no predefined categories in the query expansion-based approach, and 2) the initial query is enriched or expanded with additional terms in the query expansion-based approaches rather than mapped into a set of relevant categories in the object recognition-based approaches.

For example, Zha *et al.* propose a joint text and image query suggestion for re-ranking [Zha et al. 2009], [Zha et al. 2010]. Figure 7 shows an example of query suggestion for the initial textual query "Lamborghini." The search engine not only provides a textual query suggestion, but also the representative images for each suggestion. This kind of joint query suggestion is mined from a knowledge base, e.g., Flickr [Flickr], by selecting the most relevant and informative keywords, as well as the most visually representative images for each keyword. If a user selects one of the suggestions, the corresponding keyword will be added to enrich the original textual query, while the image collection associated with this suggestion will be formulated as the visual query. The visual query is then used as the query example to re-rank the search results (based on visual similarities) which are returned by using the suggested keywords as a new query. The visual query suggestion is very helpful and intuitive to the end users as it provides a more vivid suggestion with visual examples. However, the key challenge is how to make it respond in real-time (i.e., as soon as users input a query) and scalable to any query.

Crowd-re-ranking is still not practical, as it needs to mine knowledge from multiple resources on the Web. Unless the re-ranking model is query-independent, crowd-re-ranking will still be in its infancy. Another key issue is the noisy nature of Web knowledge in crowd-re-ranking.

2.3.4 *Interactive re-ranking Methods.* Although automatic re-ranking methods have achieved improvement over the initial search results, multimedia search systems with a human user in the loop have consistently outperformed fully automatic search systems. This has been validated for every year

[5] Please see Section 3 for the definition of $NDCG$ and the details of MSRA-MM dataset.

of search performance evaluation in TRECVID video search evaluation forum [TRECVID], as human beings can provide more concise information to guide the ranking procedure. In the interactive re-ranking procedure, a user is provided with the initial search results, interactively works through the results, and then either issues additional queries or annotates whether a part of the initial results are relevant or not. This user-provided information will be fed into the search system to further refine the search model for a better performance. The pioneer work of interactive multimedia search is relevance feedback, where users are required to annotate whether a subset of initial search results is relevant or not at each iteration [Benitez et al. 1998], [Rui et al. 1998], [Zhou and Huang 2002], [Zhou and Huang 2003]. As indicated in [Hauptmann et al. 2008b], there are three reasons that an interactive process is ideally suited for searching visual data: 1) the automatic content-based indexing of visual data is still behind progress made in the textual domain, and thus needs to keep humans in the loop; 2) while providing a good visual query might be difficult, it is more convenient for users to express the visual aspects of information need through interaction; and 3) visual data is particularly suited for interaction as a user can quickly grasp the vivid visual information and thus judge the relevance at a quick glance. Please note that interactive re-ranking is optional for all three re-ranking paradigms.

In the next section, we will introduce several exemplary interactive multimedia search systems. Since these systems keep users in the loop, a user-friendly interface for interaction is the key to achieving an efficient search. Thus, we also present their user interfaces for discussion.

The works in [T. L. Berg 2006], [Ding et al. 2008], [Hu et al. 2007], [Smith et al. 2003], [Tian et al. 2010], leverage relevance feedback to identify the relevant clusters for improving browsing efficiency. They first employ clustering techniques such as Bregman Bubble Clustering (BBC) [Ding et al. 2008], [Hu et al. 2007], and Latent Dirichlet Allocation (LDA) [T. L. Berg 2006] to cluster the top image search results, and then ask the users to label the relevance of those clusters. The user intent in the visual feature space is localized through a discriminative dimension reduction algorithm [Tian et al. 2010]. The images within the clusters are then ranked according to their similarities to the cluster centers. This is similar to the clustering-based re-ranking except that it involves human interaction to annotate the clusters. Hauptmann *et al.* propose to employ active learning technique to exploit both the human bandwidth and machine capability for video search re-ranking [Hauptmann et al. 2006]. At each iteration, the system returns the most informative or the best retrieved documents (obtained in the previous iteration) to users, among which users select the most relevant retrieved documents through the rapid serial visual presentation of search results. The system then reranks the previous search results based on the user relevance feedback via active learning. The re-ranking process runs iteratively until reaching certain steps or examining certain a number of documents.

The MediaMill system introduces the concept of video threads to visualize the reranked video search results [Snoek et al. 2006], [Worring et al. 2007]. A thread is a linked sequence of shots in a specific order type. These types include: 1) query result thread: the search results returned by the query are linked in terms of relevance; 2) visual thread: the shots are linked via visual similarity; 3) semantic thread: the shots with common concepts are linked; 4) top-rank thread: the top ranked shots from each concept are linked; 5) textual thread: the shots with similar ASR words are linked; and 6) time thread: the shots are linked following the timeline of a video. The system further supports two models for displaying threads, namely CrossBrowser and RotorBrower. The former is limited to show only two fixed threads—the query result and time threads, while the latter shows all possible relevant threads for each retrieved shot to users. Users can browse along any thread that catches their interest. Figure 8 shows these two browsing interfaces.

The real-time search re-ranking system developed by Cui *et al.* deals with the following three problems in a typical search engine [Cui et al. 2008a], [Cui et al. 2008b]: 1) how to make a user specify search intent, 2) how to define visual similarity for different search intents (since the similarities de-

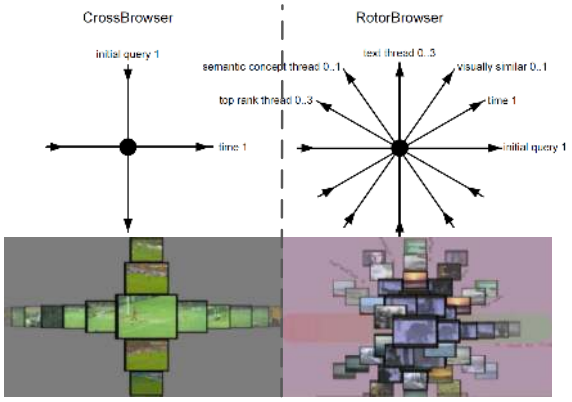

Fig. 8. Interactive search in MediaMill [Snoek et al. 2006] (with permission of Snoek et al.).

Fig. 9. Multimedia search re-ranking by (a) color map [Wang and Hua. 2011], (b) CuZero [Zavesky and Chang: 2008], and (c) Concept Map [Xu et al. 2010].

fined on different low-level features will lead to different results), and 3) how to make the re-ranking process real-time. The system enables users to first indicate a query image from the initial search results, and then classifies the query image into one of several predefined categories (e.g., "General Object," "Object with Simple Background," "Portrait," and "People") by the C4.5 decision tree. The feature weights for ensembling a visual similarity within each category are learned by minimizing a rank loss for all query images in a training set through RankBoost. The re-ranking process is fast since the average time for computing the visual similarity between each image pair is 0.01ms. The system is designed similar with the CueFlik, where users can provide example images and create their own rules for search via a few interactions [Fogarty et al. 2007].

Different from conventional sketch- or shape- based systems which retrieve images with similar sketch or shape [Cao et al. 2011], [Cao et al. 2010], [Eitz et al. 2011], the "search by color map" system enables users to indicate how their preferred colors are spatially distributed in the desired image, by scribbling a few color strokes, or dragging an image and highlighting a few regions of interest in an intuitive way [Wang and Hua. 2011]. Figure 9(a) shows the initial search results and the reranked search results by using a color map palette to specify search intent. For example, the search intent like "purple flower with green leaf around" usually cannot be satisfied via issuing a piece of textual query in an image search engine. However, in this system, based on the initial image search results about "flower," users can merely scribble the purple and green color strokes in the palette. Then, the system formulates a color map query by extracting the dominant colors and their corresponding positions from the palette, and reranks the top 1,000 search results according to color similarity.

The concept-based interactive multimedia search system leverages human interaction to re-formulate a new query, leading to a better search performance based on the spatial layout of concepts [Xu et al. 2010], [Zavesky and Chang: 2008]. Given a textual query, CuZero developed by the Columbia University is able to automatically discover relevant visual concepts in real time, and allows users to navigate seamlessly in the concept space at-will [Zavesky and Chang: 2008]. The search results will be reranked according to the arbitrary permutations of multiple concepts given by users. Figure 9(b) shows the user interface of CuZero system, in which users are provided with the automatic concept-based query suggestion and can adjust the weight of each concept by changing the size of the concept rectangle on the query panel. Such a navigation system allows efficient exploration of different types of new queries with zero effort. Similarly, the "image search by concept map" system enables users to indicate not only which semantic concepts are expected in the query, but also how these concepts are spatially distributed in the desired images [Xu et al. 2010]. Figure 9(c) shows the user interface, in which users can not only adjust the size and position of each concept on the query panel, but also select an exemplary query image for each concept. Similar ideas have been implemented for mobile visual search, where users can leverage multimodal input such as voice and image, as well as the advanced multi-touch function to conduct visual search on mobile devices [Wang et al. 2011].

Generally, interactive re-ranking can achieve better search performance as it involves human knowledge during re-ranking processes. However, since human subjects are usually reluctant to interact too much with the search engines, it is therefore critical to design natural re-ranking user interface to maintain user experience while minimizing users' interacting time. It is also important to update re-ranking model instantly at each round of interaction.

3. DATASETS AND PERFORMANCE EVALUATION

3.1 Datasets for Multimedia Search Re-ranking

Since different papers report experimental results using different datasets, to compare methods fairly on a reasonable scale, a few benchmark datasets have recently been compiled. We review the following representative datasets because: 1) they were collected from real-world data sources, 2) they are large-scale and contain enough modality information for re-ranking (i.e., visual contents and textual descriptions), 3) they are open to research communities, and 4) they have been widely used in related research areas.

—Oxford 5K. The University of Oxford provides about 5,062 landmark images with different sizes of distractors from Flickr to evaluate object retrieval performance [Philbin et al. 2007]. A set of images comprising 11 different Oxford "landmarks" were collected. Images for each landmark were retrieved from Flickr [Flickr], using queries such as "Oxford Christ Church" and "Oxford Radcliffe Camera." They also retrieved further distractor images by searching on "Oxford" alone. The entire dataset consists of 5,062 high-resolution (1024×768) images. Each image is assigned one of four possible labels: (1) good—a nice, clear picture of the object/building, (2) OK—more than 25% of the object is clearly visible, (3) junk—less than 25% of the object is visible, or there is a very high level of occlusion or distortion, and (4) absent—the object is not present. They include two collections of distractors together with the Oxford 5K dataset: 1) the 100K dataset crawled from Flickr's 145 most popular tags and consisting of 99,782 high resolution images, and 2) the 1M dataset crawled from Flickr's 450 most popular tags and consisting of 1,040,801 medium resolution images. The Oxford 5K dataset provides a good test bed for evaluating object retrieval (belonging to example-based-re-ranking) in the computer vision community [Philbin et al. 2007], [Jegou et al. 2010].

—MSRA-MM. The MSRA-MM dataset was first released by Microsoft Research Asia in 2009 and further enriched in 2010. It contains an image set with one million images and a video set with 23,517

videos [Li et al. 2009], [Wang et al. 2009]. These data were collected from a commercial visual search engine by issuing representative 1,165 representative images and 217 video search queries (textual queries for image and video search), selected from the real-world query log. Each image and video keyframe is associated with a set of global visual features (e.g., color histogram, edge, texture, etc.), surrounding text such as the title and URL, and the corresponding categories (manually annotated 100 categories for image and 9 for video). The most distinctive feature of this dataset is that the original ranked order of each image and video is kept for each query. Therefore, it is natural and convenient for researchers to apply multimedia search re-ranking on this dataset. The MSRA-MM dataset has been used in many re-ranking and annotation tasks [Hong et al. 2010], [Wang and Hua. 2011], [Wang et al. 2011], [Wang et al. 2012].

—TRECVID. The main forum for studying video retrieval in the last few years, has been organized by the National Institute of Standards and Technology (NIST) in the form of the TRECVID video retrieval evaluations [TRECVID]. In 2001, NIST started the TREC Video Track (now referred to as TRECVID [Smeaton et al. 2006], [Smeaton et al. 2002]) to promote progress in content-based video retrieval via an open and metrics-based evaluation. The video corpora in TRECVID have comprised documentaries, movies, multilingual broadcast news, and surveillance videos, with international participation growing from 12 to 73 companies and academic institutions, as well as an increase of archived video materials from 12 to over 800 hours, from 2001 to 2010. Each video is associated with the textual description from a transcript, optical character recognition (OCR) results, or machine translation from multi-language. A large-scale concept ontology, which contains 834 concepts with broad categories (such as objects, activities/events, scenes/locations, people, and graphics), is also provided for video classification and concept-based search [Naphade et al. 2006]. As one of the largest video collections with manual annotations available to the research community, the TRECVID collections have become the standard large-scale test beds for the task of multimedia retrieval. These evaluations provide a standard collection available to all participants. A set of retrieval queries is also made available. The TRECVID query topics include requests for specific items or people and general instances of locations and events. Exemplary queries include "find shots with one or more people leaving or entering a vehicle," and so on. The test queries include text plus optionally example imagery, example video, and audio. Participants must return answers to these queries from the test video collection to NIST for official relevance judgments within a fixed timeframe of a few weeks after the release of the queries. In video retrieval, a broadcast video is commonly decomposed into numerous shots, with each shot represented by one or multiple keyframes. The numerous keyframes can then be subjected to image retrieval strategies. There is extensive research working on video search re-ranking based on TRECVID video collection, either in an automatic or interactive way [6].

—NUS-WIDE. To facilitate the research of mining community-contributed images and tags, the National University of Singapore presented NUS-WIDE as an open benchmark dataset in 2009 [Chua et al. 2009], [NUS-WIDE]. The dataset includes: 1) 269,648 images and 5,018 associated unique tags, which are randomly crawled from Flickr [Flickr]; 2) a set of low-level visual features for each image (such as color, edge, texture, and visual words based on local descriptors); and 3) ground-truth for 81 manually annotated concepts. Since each image has its tags, as well as visual descriptors and associated concepts, it is natural to apply multimedia search re-ranking, especially clustering and concept based raranking methods, to NUS-WIDE, e.g., [Li et al. 2010], [Zhu et al. 2010], [Yang et al. 2012].

[6] A list of recent publications that conduct search evaluations on the TRECVID data collection can be found at http://trecvid.nist.gov/trecvid.bibliography.txt.

Table II. DATASETS FOR MULTIMEDIA SEARCH RE-RANKING

Dataset	Description	Metadata	Link
Oxford 5K [Philbin et al. 2007]	5,062 Oxford landmark images	Labels of 11 landmarks	http://www.robots.ox.ac.uk/ ∼ vgg/data/oxbuildings/
MSRA-MM [Li et al. 2009]	23,517 web videos	Keyframes, visual descriptors, 109 categories, ranking	http://research.microsoft.com/ en-us/projects/msrammdata/
TRECVID [TRECVID]	8,000 Internet videos (200 hrs) TV videos (380 hrs, 129,668 shots)	Title/keywords/description shots/SR/topic/concept	http://www-nlpir.nist.gov/projects/ tv2011/tv2011.html
NUS-WIDE [NUS-WIDE]	269,648 Flickr images with 5,018 tags	Visual descriptors, 81 concepts	http://lms.comp.nus.edu.sg/ research/NUS-WIDE.htm
MCG-WEBV [Cao et al. 2009]	90,031 Youtube videos with 6,392 tags	Keyframes, visual descriptors, 81 topics and 15 categories	http://mcg.ict.ac.cn/mcg-webv.htm
ImageNet [ImageNet]	12,184,113 web images with 17,624 synsets	SIFT descriptors, annotated bounding box (subset)	http://www.image-net.org/
Tiny Image [Torralba 2008]	79,302,017 web images (32×32)	GIST descriptors, filename, ranking in search results	http://horatio.cs.nyu.edu/mit/ tiny/data/index.html
ImageCLEF [ImageCLEF]	237,434 images from Wikipedia	User-supplied annotations	http://www.imageclef.org/
PASCAL [PASCAL]	11,530 images with 20 classes, 27,450 objects and 6,929 segmentations	image classes, annotated objects and segmentations	http://pascallin.ecs.soton.ac.uk/ challenges/VOC/

—MCG-WEBV. The Institute of Computing Technology, Chinese Academy of Sciences, released a web video benchmark dataset in 2009 [Cao et al. 2009]. The dataset, called MCG-WEBV, consists of 90,031 most viewed videos and 6,392 user-provided tags from Youtube [YouTube], during the period from Dec. 2008 to Feb. 2009. To facilitate multimedia application and algorithm evaluation, 1.4 million shots and 2.1 million keyframes are provided. Each video is associated with five types of features, including the available metadata from Youtube (such as uploader, category, rating, title and description, related videos, number of views and comments, etc.), nine visual descriptors (such as color, edge, texture, face, and visual words), as well as textual and audio features. The core set of 3,283 videos, which are the seed videos for collecting related videos in MCG-WEBV, are further annotated with 81 topics and 15 Youtube categories by human. MCG-WEBV is characterized by its socialized features, i.e., human rating, related videos, number of views and comments. These community-based features can facilitate multimedia search re-ranking from the perspective of social networking and collaborative filtering.

There are some other datasets that could be used for visual re-ranking. For example, the Stanford Vision Lab released the ImageNet as an open image dataset organized ac-cording to the WordNet Hierarchy [Deng et al. 2009], [ImageNet]. ImageNet uses "synset" to denote a meaningful concept in WordNet, which is possibly described by multiple words or word phrases. There are 17,624 non-empty synsets in the ImageNet, each with 690 images on average. Although ImageNet is mainly designed for vision tasks (e.g., object recognition, image classification, indexing, and retrieval) rather than multimedia search re-ranking, this clean dataset with large-scale of images and rich concepts provides a good data source for building a concept hierarchy, and is thus suitable for concept and learning based re-ranking. However, no ground truth of initial search results is provided in the ImageNet. TinyImage is a dataset of 80 million 32×32 low resolution images [Torralba 2008]. Similar to ImageNet, the images in TinyImage were collected from the Internet by querying all words in WordNet in several image search engines. The difference is that only 10%-25% of the images in each synset are clean in TinyImage. MIRFLICKR is an image dataset released by the Leiden University [Huiskes and Lew 2008], [Huiskes et al. 2010]. It first contain 25,000 images collected from Flickr and then is extended to one million. Similar to NUS-WIDE [Chua et al. 2009], each image in MIRFLICKR is associated with user-provided tags, relevance score, and visual descriptors. In addition, the EXIF (EXchangable Image File Format) metadata is also provided. ImageCLEF recently released a Wikipedia collection for image retrieval, which consists of 237,434 images and user-supplied annotations [ImageCLEF], [Myoupo et al. 2009]. The images and their associated articles were collected from the well-known structured data source

Wikipedia [Wikipedia]. The INRIA introduces a new public dataset of 353 image search queries, called Web queries [Krapac et al. 2010]. For each query, the dataset includes the original textual query, the top-ranked images returned by a web search engine, and an annotation file (i.e., relevant or not with respect to the query, image URL, page title, the alternative text of the image, and the surrounding text of the image). In total, there are 71,478 images. The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection [PASCAL], [Everingham et al. 2010]. Organized from 2005, there have been 11,530 images with 20 classes, containing 27,450 region-of-interest (ROI) annotated objects and 6,929 segmentations. Although its main objective is to evaluate object detection and recognition, the annotated objects (especially the annotated ROI) are very useful to train and validate concept detectors for search re-ranking.

On the other hand, some researchers also provide their own datasets for multimedia search re-ranking, although not very comprehensive. For example, Tian *et al.* collected 94,341 images from a commercial image search engine by issuing 105 queries [Tian et al. 2010]. Liu *et al.* collected 78,000 images by querying 29 representative queries (selected from the real-world query log) in three major commercial image search engines and Flickr [Liu et al. 2010], [Liu et al. 2009]. For each query, the top 1,000 returned images were collected. Jain *et al.* used 193,000 images crawled by 193 distinct queries in a commercial image search engine, with each query containing the top 1,000 searched images [Jain and Varma 2011]. However, most of these datasets do not include large-scale textual queries or enough metadata (such as surrounding text and relevance scores for each visual document), and therefore would limit the evaluation of multimedia search re-ranking.

Table II summarizes the datasets used for multimedia search re-ranking.

3.2 Performance Metrics

Since the goal of visual re-ranking is to improve the performance of multimedia search, it is natural to utilize the existing performance metrics of information retrieval (IR) and multimedia search to evaluate the performance of visual raranking. Additionally, there are also many other performance metrics that are more suitable for visual re-ranking. In this subsection, we will review these two categories of performance metrics.

3.2.1 *Performance Metrics for Multimedia Search*

—Precision-Recall (P-R curve). Precision and recall are the traditional metrics in the field of information retrieval [Baeza-Yates and Ribeiro-Neto 1999]. Success in the search task is measured through precision and recall as the central criteria to evaluate the performance of retrieval algorithms. Recall is defined as the fraction of retrieved relevant documents in the whole dataset, while precision is the fraction of the retrieved documents in the returned subset. Then, a P-R curve can be generated by plotting the curve of precision versus recall. The precision versus recall curve is usually based on 11 (instead of ten) standard recall levels: 0%, 10%, 20%, ..., 100% [Baeza-Yates and Ribeiro-Neto 1999].

—Mean Average Precision (MAP). Another widely adopted performance metric is the average precision over a set of retrieved visual documents. Precision is the number of relevant documents retrieved divided by the total number retrieved. Let $y_i \in \{0, 1\}$ denote if the i-th document d_i in the ranked list $\mathbf{r}$ is relevant ($y_i = 1$) or not ($y_i = 0$) [7]. The average precision (AP) is defined by $AP = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{i} \sum_{j=1}^{i} y_j$, where N is the number of retrieved documents, and $\sum_{j=1}^{i} \frac{y_j}{i}$ is the precision at given rank i. Note that AP corresponds to the area under a non-interpolated P-R curve. Then, MAP is computed by averaging the AP across all given query topics [Hauptmann and Christel 2004]. MAP is widely adopted

[7] Please note that here the definition of y_i is slightly different from that in Section 2.3.2 where $y_i \in \{-1, +1\}$.

as the major performance metric in many re-ranking methods that have conducted experiments on the TRECVID dataset.

—Normalized Discounted Cumulative Gain (NDCG). NDCG is a commonly adopted metric for evaluating a search engine's performance [Jarvelin and Kekalainen 2000], [Jarvelin and Kekalainen 2002]. NDCG measures the usefulness, or gain, of a ranked list of documents based on their positions in the ranked list. The gain is accumulated from the top of the ranked list to the bottom with the gain of each result discounted at lower ranks. NDCG is widely adopted when the ground-truth data has multiple levels of confidence scores (e.g., a scale of 5 ranging from 1 to 5 indicating the relevance degree) labeled by users.

—Response time. Due to its dynamic audiovisual nature, a multimedia search system could be evaluated more effectively than in a static performance metric. The VideoOlympics is a real-time evaluation showcase in which video search systems compete to answer specific video searches in front of a live audience [Snoek et al. 2008]. This is different than any other evaluation in which the evaluations are more focused on the effectiveness of collected retrieval results (e.g., TRECVID [TRECVID]). The influence of interaction mechanisms and the advanced visualizations are taken into account in the VideoOlympics: each participating search system forms a client that communicates independently to an evaluation server. The evaluation server instantly processes the incoming results, prioritizes them using a time stamp, compares them to the ground truth, and updates a score list to the audience in real time. This requires that response time (i.e., the time cost between when a user submits a query and the system returns the search results to the user) of a search system be as short as possible.

3.2.2 *Performance Metrics for Visual Re-ranking*

—Diversity. Another metric that attracts attention in the research community is "diversity"—whether the search results are both relevant and diverse in terms of visual appearance [Kennedy and Naaman 2008], [Popescu et al. 2009], [Song et al. 2006], [Wang et al. 2010], [Yang et al. 2010]. Because of the appearance-based nature of multimedia search results, a diverse presentation would enable users to have a quick glance and understanding of the research results. Since it is not easy to measuring the "diversity" in an objective way, researchers have proposed different metrics for "diversity" from subjective user studies. For example, in a landmark image retrieval system, Kennedy *et al.* ask users to answer four evaluation questions in different scales: *representativeness* (0–10 scale), *unique* (0–10 scale), *comprehensive* (1–5 scale), and *satisfying* (1–5 scale) [Kennedy and Naaman 2008]. Popescue *et al.* ask users to evaluate the search results in terms of *accuracy* and *diversity* in a user study [Popescu et al. 2009]. Wang and Yang *et al.* propose a new metric for measure "diversity"—named Average Diverse Precision (ADP) [Wang et al. 2010], [Yang et al. 2010]. The ADP is defined by extending the existing AP and taking the appearance dissimilarity into consideration. The ADP is given as:

$$ADP = \frac{1}{N}\sum_{i=1}^{N} y_i D(i) \frac{\sum_{j=1}^{i} y_j D(j)}{i}, \tag{10}$$

where the diversity score $D(i)$ for d_i indicates the minimal difference with the documents appearing before it, i.e., $D(j) = min_{1 \leqslant k < j}(1 - S(d_k, d_j))$. $S(\ )$ is the similarity between the document d_k and d_j.

—Typicality. Other metrics include "typicality" [Liu et al. 2008], [Liu et al. 2010], [Tang et al. 2007], and "novelty" [Wu et al. 2007]. "Typicality" is defined as human perception of the degree of document relevance with respect to a given query or an object category, which can be derived from two components: the similarity between this document and other retrieved documents, and the dissimilarity with the documents not in the ranked list. For example, the Average Typicality (AT) is defined

by $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{i}\frac{t_j}{o_j}$ [Tang et al. 2007], where t_j is the ground truth of typicality score of the j-th document (labeled by human subjects) and o_j is the j-th highest typicality score in the ground truth.

4. CONCLUSIONS AND FUTURE CHALLENGES

This paper has attempted to provide a comprehensive survey of research on multimedia search re-ranking and to provide some structural categories for the methods described in over 150 papers. We have summarized the research into four paradigms, i.e., *self-re-ranking*, *example-based-re-ranking*, *crowd-re-ranking*, and *interactive-re-ranking*.

Although significant progress has been made in recent years, there are a lot of emerging topics that deserve further investigation and research. We would like to summarize the future challenges as follows.

—*Re-ranking difficulty: whether to re-rank the initial search results*. It has been found that re-ranking cannot always help improve initial search performance [Morioka and Wang 2011]. It depends on the query, the initial search results, and the knowledge. Note that successful re-ranking approaches mostly report their performance in terms of "average" precision or recall on the "popular" or "simple" queries (e.g., "animals," "Paris," "apple," and so on). We can observe that not all queries gain improvements by re-ranking. Few reports focus on the "nonpopular" or "complicated" queries—even for those "popular" or "simple" queries, the re-ranking performance is not always satisfying due to the ambiguity nature of the query terms. Then, the problem is whether we can know when to re-rank the initial search results. Specifically, given a query and an initial ranked list, how can we decide if re-ranking can truly improve search performance. We need to analyze the *re-ranking difficulty* of query, as well as the performance of the initial ranked list (i.e., how noisy or satisfying the initial list is) before we conduct re-ranking.

—*Large-scale re-ranking*. Most existing approaches to multimedia search re-ranking predominantly focus on the top (e.g., top 1,000) ranked documents, ignoring the low-ranked documents which might be also relevant (in different aspects). This is partially because: 1) the top ranked documents are regarded more relevant, and 2) the approach itself cannot be scaled up. Given the unsatisfying multimedia search results, there might be valuable and relevant information in the low-ranked documents. This requires that re-ranking is able to deal with large-scale of data. Another real-world application is (near-) duplicate or similar image search, where a user submits a query example or interactively indicates his/her object-of-interest within an image or video clip, the search system responds by returning a set of images from a million or billion scale of database. There have been attempts to address the large-scale problem. For example, the works in [Jegou et al. 2010], [Zhou et al. 2010] have proposed the spatial coding or geometric verification techniques to improve traditional vocabulary tree-based indexing scheme [Nister and Stewenius 2006]. The work in [Wang et al. 2010] represents one of the first attempts towards billion scale image search. Large-scale re-ranking brings a new challenge to the computer vision community.

—*User-centric re-ranking*. The goal of a search system is to bridge the gap between user's information needs and large-scale of available data [Rose and Levinson 2004]. If we take the initial ranked list as a generic search process, then re-ranking could be regarded as the second step for providing user-specific or personalized search results based on the initial general results. Then, the key of user-centric re-ranking is to understand user's search intent. Although there has been rich research on personalized search in the text domain [Koren et al. 2008], [Sontag et al. 2012], few attempts have been made in the visual domain. Possible investigations for personalized visual re-ranking include: 1) understanding query to guide re-ranking, e.g., query-dependent and feature-dependent (which type of visual features can best represent a user's search intent) re-ranking; 2) keeping the

user in the loop, e.g., providing a user-friendly search interface to enable formulation of user intent in a natural way; 3) understanding user preference from search logs and community-based behaviors [Trevisiol et al. 2012]; and 4) investigating the aesthetic and visualization aspects of human perception on visual re-ranking presentation [Geng et al. 2011].

—*Context-aware re-ranking*. When a user is conducting a search task, s/he actually provides rich context to the search system, e.g., the past behaviors in the same session, the browsed web pages if a search is triggered from browsing behavior, geographic location and time of the user, social networks if the user remains signed in, and so on. All these contexts provide valuable clues for contextual search. For example, we may leverage the surrounding text to provide a contextual image search if the user indicates his/her interested image in a web page [Lu et al. 2011]. We may also build a user model through the rich context on the mobile phone if s/he is conducting a mobile visual search [Zhuang et al. 2011], or if s/he contributes heterogenous data in a social media site [Zhuang et al. 2011].

—*Protocol for visual re-ranking*. We lack a standardized protocol for multimedia search re-ranking. While TRECVID [TRECVID] and Video Olympic [Snoek et al. 2008] are trying to build protocols (datasets, experimental and evaluation settings, performance metrics, etc.) for video search, there are no common and widely adopted protocols for multimedia search re-ranking. Unless we have a common evaluation protocol, we cannot quantatively compare the performance of the many re-ranking methods discussed in this paper.

Multimedia search re-ranking is a challenging and interesting problem in and of itself. However, it can also be seen as one of the few attempts at solving one of the grand challenges of multimedia - multimedia information retrieval. The development of multimedia understanding, indexing, visualization, and interaction technologies aimed at the distinct features of re-ranking is still in its early stage. Therefore, the future of multimedia search re-ranking depends a lot on the collective focus and overall progress in each aspect of multimedia retrieval, especially the crowdsourced knowledge contributed by a large amount of real users (e.g., search logs).

5. ACKNOWLEDGEMENT

The authors would like to thank Dr. Yuan Liu, Dr. Jingdong Wang, Dr. Shiliang Zhang, Dr. Wengang Zhou, and Mr. Ting Yao, for their insightful discussions. The authors also would like to thank the anonymous reviewers for their valuable comments. This work was supported in part to Dr. Qi Tian by ARO grant W911NF-12-1-0057, NSF IIS 1052851, and 2012 UTSA START-R Research Award, respectively. This work was also supported in part by National Science Foundation of China (NSFC) 61128007.

REFERENCES

AMIR, A., ARGILLANDER, J., CAMPBELL, M., HAUBOLD, A., IYENGAR, G., EBADOLLAHI, S., AND KANG, F. 2005. IBM Research TRECVID-2005 video retrieval system. In *Proceedings of TRECVID workshop*.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. Addison Wesley.

BEN-HAIM, N., BABENKO, B., AND BELONGIE, S. 2006. Improving web-based image search via content based clustering. In *Proceedings of Computer Vision and Pattern Recognition Workshop on SLAM*.

BENDERSKY, M. AND KURLAND, O. 2008. Re-ranking search results using document-passage graphs. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*.

BENITEZ, A., BEIGI, M., AND CHANG, S.-F. 1998. Using relevance feedback in content-based image metasearch. *IEEE Internet Computing 2,* 4, 59–69.

BING. http://www.bing.com/.

BOGERS, T. AND BOSCH, A. 2009. Authoritative reranking in fusing authorship-based subcollection search results. In *Proceedings of Dutch-Belgian Information Retrieval Workshop*. 49–55.

BOLL, S. 2007. MultiTube—where multimedia and web 2.0 could meet. *IEEE Multimedia 14,* 1 (Jan.-March), 9–13.

BRIN, S. AND PAGE, L. 1998a. The anatomy of a large-scale hyper textual web search engine. *Computer Networks and ISDN Systems 30,* 1-7, 107–117.

BRIN, S. AND PAGE, L. 1998b. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the International Conference on World Wide Web*.

BRINKMEIER, M. 2006. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the International Conference on World Wide Web*. 282–301.

CAI, D., HE, X., LI, Z., MA, W.-Y., AND WEN, J.-R. 2004. Hierarchical clustering of WWW image search results using visual, textual and link information. In *Proceedings of ACM Multimedia*. 952–959.

CAO, J., ZHANG, Y. D., SONG, Y. C., CHEN, Z. N., ZHANG, X., AND LI, J. T. 2009. MCG-WEBV: A benchmark dataset for web video analysis. In *Technical Report, ICT-MCG-09-001*.

CAO, Y., WANG, C., ZHANG, L., AND ZHANG, L. 2011. Edgel index for large-scale sketch-based image search. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 761–768.

CAO, Y., WANG, H., WANG, C., LI, Z., ZHANG, L., AND ZHANG, L. 2010. Mindfinder: interactive sketch-based image search on millions of images. In *Proceedings of ACM Multimedia*.

CAO, Z., QIN, T., LIU, T.-Y., TSAI, M.-F., AND LI, H. 2007. Learning to rank: from pair-wise approach to list-wise approach. In *Proceedings of International Conference on Machine Learning*. 129–136.

CARBONELL, J. 1997. Translingual information retrieval: a comparative evaluation. In *Proceedings of International Joint Conference on Artificial Intelligence*.

CARBONELL, J. AND GOLDSTEIN, J. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*.

CHANG, S.-F., HSU, W., JIANG, W., KENNEDY, L., XU, D., YANAGAWA, A., AND ZAVESKY, E. 2006. Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. In *TREC Video Retrieval Evaluation Online Proceedings*.

CHANG, S.-F., MA, W.-Y., AND SMEULDERS, A. 2007. Recent advances and challenges of semantic image/video search. In *ICASSP*.

CHEN, S., WANG, F., SONG, Y., AND ZHANG, C. 2011. Semi-supervised ranking aggregation. *Information Processing & Management 47,* 3 (May), 415–425.

CHUA, T.-S., NEO, S.-Y., LI, K.-Y., WANG, G., SHI, R., ZHAO, M., AND XU, H. 2004. TRECVID 2004 Search and Feature Extraction Task by NUS PRIS. In *TREC Video Retrieval Evaluation Online Proceedings*.

CHUA, T.-S., TANG, J., HONG, R., LI, H., LUO, Z., AND ZHENG, Y.-T. 2009. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of ACM International Conference on Image and Video Retrieval*.

CHUM, O., MIKULIK, A., PERDOCH, M., AND MATAS, J. 2011. Total Recall II: Query Expansion Revisited. In *Proceedings of CVPR*.

CHUM, O., PHILBIN, J., SIVIC, J., ISARD, M., AND ZISSERMAN, A. 2007. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of ICCV*.

CROFT, W. B. 2000. Combining approaches to information retrieval. In *Advanced in Information Retrieval*. 1–36.

CUI, J., WEN, F., AND TANG, X. 2008a. IntentSearch: Interactive on-line image search re-ranking. In *Proceedings of ACM Multimedia*. 997–998.

CUI, J., WEN, F., AND TANG, X. 2008b. Real time Google and Live image search re-ranking. In *Proceedings of ACM Multimedia*. 729–732.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys 40,* 65.

DENG, H., LYU, M. R., AND KING, I. 2009. Effective latent space graph-based re-ranking model with global consistency. In *Proceedings of the Sencond ACM International Conference on Web Search and Data Mining*.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. 2009. A large-scale hierarchical image database. In *Proceedings of IEEE CVPR*.

DING, H., LIU, J., AND LU, H. 2008. Hierarchical clustering-based navigation of image search results. In *Proceedings of ACM Multimedia*. 741–744.

DONALD, K. M. AND SMEATON, A. F. 2005. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Proceedings of ACM International Conference on Image and Video Retrieval*.

DWORK, C., KUMAR, S. R., NAOR, M., AND SIVAKUMAR, D. 2001. Rank aggregation methods for the web. In *Proceedings of International World Wide Web Conference*. 613–622.

EBADOLLAHI, S., JOSHI, D., NAPHADE, M., NATSEV, A., SEIDL, J., SMITH, J. R., SCHEINBERG, K., TESIC, J., AND XIE, L. 2006. IBM Research TRECVID-2006 Video Retrieval System. In *TREC Video Retrieval Evaluation Online Proceedings*.

EITZ, M., HILDEBRAND, K., BOUBEKEUR, T., AND ALEXA, M. 2011. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Trans. on Visualization and Computer Graphics 17,* 11 (Nov.), 1624–1636.

EVERINGHAM, M., GOOL, L. V., WILLIAMS, C. K. I., WINN, J., AND ZISSERMAN, A. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision 88,* 2, 303–338.

FACEBOOK. http://www.facebook.com/.

FERGUS, R., FEI-FEI, L., PERONA, P., AND ZISSERMAN, A. 2005. Learning object categories from google's image search. In *Proceedings of International Conference on Computer Vision*. 1816–1823.

FERGUS, R., PERONA, P., AND ZISSERMAN, A. 2004. A visual category filter for Google images. In *Proceedings of European Conference on Computer Vision*.

FLICKR. http://www.flickr.com/.

FOGARTY, J., D. TAN, A. K., AND WINDER, S. 2007. Cueflik: Interactive concept learning in image search. In *Proceeding of SIGCHI Conference on Human Factors in Computing Systems*. 29–38.

GENG, B., YANG, L., XU, C., HUA, X.-S., AND LI, S. 2011. The role of attractiveness in web image search. In *Proceedings of ACM Multimedia*. 63–72.

GOOGLE. http://www.google.com/.

HAUBOLD, A., NATSEV, A., AND NAPHADE, M. R. 2006. Semantic multimedia retrieval using lexical query expansion and model-based reranking. In *IEEE International Conference on Multimedia & Expo*. NY,USA.

HAUPTMANN, A., CHEN, M.-Y., CHRISTEL, M., HUANG, C., LIN, W.-H., NG, T., PAPERNICK, N., VELIVELLI, A., YANG, J., YAN, R., YANG, H., AND WACTLAR, H. D. 2004. Confounded Expectations: Informedia at TRECVID 2004. In *TREC Video Retrieval Evaluation Online Proceedings*.

HAUPTMANN, A. G., CHRISTEL, M., AND YAN, R. 2008a. Video retrieval based on semantic concepts. In *Proceedings of the IEEE*. 602–622.

HAUPTMANN, A. G. AND CHRISTEL, M. G. 2004. Successful approaches in the TREC video retrieval evaluations. In *Proceedings of ACM Multimedia*. 668–675.

HAUPTMANN, A. G., CHRISTEL, M. G., AND YAN, R. 2008b. Video retrieval based on semantic concepts. *Proceedings of the IEEE 96,* 4 (April), 602–622.

HAUPTMANN, A. G., LIN, W. H., YAN, R., YANG, J., AND CHEN, M. Y. 2006. Extreme video retrieval: Joint maximization of human and computer performance. In *Proceedings of the ACM International Conference on Multimedia*. Santa Barbara, USA.

HAUPTMANN, A. G., YAN, R., LIN, W.-H., CHRISTEL, M., AND WACTLAR, H. 2007. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transaction on Multimedia 9,* 5, 958–966.

HAVELIWALA, T. H. 2002. Topic-sensitive pagerank. In *Proceedings of the International Conference on World Wide Web*.

HE, J., ZHANG, C., ZHAO, N., AND TONG, H. 2005. Boosting web image search by co-ranking. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. 409–412.

HERBRICH, R., GRAEPEL, T., AND OBERMAYER, K. 2000. Large margin rank boundaries for ordinal regression. *In Advances in Large Margin Classifiers*, 115–132.

HOI, S. C. AND LYU, M. R. 2007. A multimodal and multilevel ranking framework for content-based video retrieval. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1520–6149.

HONG, R., LI, G., NIE, L., TANG, J., AND CHUA, T.-S. 2010. Exploring large scale data for multimedia QA: an initial study. In *Proceedings of ACM International Conference on Image and Video Retrieval*. 74–81.

HSU, W. AND CHANG, S.-F. 2007. Video search reranking through random walk over document-level context graph. In *Proceedings of ACM Multimedia*. Augsburg, Germany, 971–980.

HSU, W., KENNEDY, L., AND CHANG, S.-F. 2007. Reranking methods for visual search. *IEEE Multimedia 14,* 3 (July-Sept.), 14–22.

HSU, W. H., KENNEDY, L. S., AND CHANG, S.-F. 2006. Video search reranking via information bottleneck principle. In *Proceedings of the ACM International Conference on Multimedia*. Santa Barbara, USA.

HU, Y., YU, N., LI, Z., AND LI, M. 2007. Image search result clustering and re-ranking via partial grouping. In *IEEE International Conference on Multimedia & Expo*. 603–606.

HUA, G. AND TIAN, Q. 2009. What can visual content analysis do for text based image search? In *IEEE International Conference on Multimedia & Expo*. 1480–1483.

HUISKES, M. J. AND LEW, M. S. 2008. The MIR Flickr retrieval evaluation. In *Proceedings of ACM International Conference on Multimedia Information Retrieval*.

HUISKES, M. J., THOMEE, B., AND LEW, M. S. 2010. New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative. In *Proceedings of ACM International Conference on Multimedia Information Retrieval*.

IMAGECLEF. Imageclef - the clef cross language image retrieval track.

IMAGENET. Imagenet.

JAIN, V. AND VARMA, M. 2011. Learning to re-rank: Query-dependent image re-ranking using click data. In *Proceedings of International World Wide Web Conference*.

JARVELIN, K. AND KEKALAINEN, J. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*.

JARVELIN, K. AND KEKALAINEN, J. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans on Information Systems 20,* 4, 422–446.

JEGOU, H., DOUZE, M., AND SCHMID, C. 2010. Improving bag-of-features for large scale image search. *International Journal of Computer Vision 87,* 3 (May), 316–336.

JING, F., WANG, C., YAO, Y., DENG, K., ZHANG, L., AND MA, W.-Y. 2006. Igroup: web image search results clustering. In *Proceedings of ACM Multimedia*. 377–384.

JING, Y. AND BALUJA, S. 2008a. PageRank for product image search. In *Proceedings of International World Wide Web Conference*. Beijing, China, 307–315.

JING, Y. AND BALUJA, S. 2008b. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. on Pattern Analysis and Machine Intelligence 30,* 11 (Nov.), 1877–1890.

KENNEDY, L. AND CHANG, S.-F. 2007. A reranking approach for context-based concept fusion in video indexing and retrieval. In *Proceedings of ACM International Conference on Image and Video Retrieval*.

KENNEDY, L., CHANG, S.-F., AND NATSEV, A. 2008a. Query-adaptive fusion for multimodal search. *Proceedings of IEEE 96,* 4, 567–588.

KENNEDY, L., CHANG, S.-F., AND NATSEV, A. 2008b. Query-adaptive fusion for multimodal search. *Proceedings of the IEEE 96,* 4 (April), 567–588.

KENNEDY, L. AND NAAMAN, M. 2008. Generating diverse and representative image search results for landmarks. In *Proceedings of International World Wide Web Conference*. Beijing, China, 297–306.

KENNEDY, L., NAAMAN, M., AHERN, S., NAIR, R., AND RATTENBURY, T. 2007. How Flickr helps us make sense of the world: Context and content in community-contributed media collections. In *Proceedings of ACM Multimedia*. Augsburg, Germany.

KOREN, J., ZHANG, Y., AND LIU, X. 2008. Personalized interactive faceted search. In *Proceedings of WWW*. Beijing, China.

KRAPAC, J., ALLAN, M., VERBEEK, J., AND JURIE, F. 2010. Improving web-image search results using query-relative classifiers. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 1094–1101.

KURLAND, O. AND LEE, L. 2005. Pagerank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*. 306–313.

LEE, K.-S., PARK, Y.-C., AND CHOI, K.-S. 2001. Re-ranking model based on document clusters. *Information Processing & Management 37,* 1 (Jan.), 1–14.

LEW, M. S. 2000. Next-generation web searches for visual content. *IEEE Computer Society 3,* 11, 46–53.

LEW, M. S., SEBE, N., DJERABA, C., AND JAIN, R. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications 2,* 1 (February), 1–19.

LI, H., WANG, M., AND HUA, X.-S. 2009. MSRA-MM 2.0: A large-scale web multimedia dataset. In *Proceedings of ICDM Workshop on Internet Multimedia Mining*.

LI, J., CHANG, S.-F., LESK, M., LIENHART, R., LUO, J., AND SMEULDERS, A. W. M. 2007. New challenges in multimedia research for the increasingly connected and fast growing digital society. In *Proceedings of ACM SIGMM International Workshop on Multimedia Information Retrieval*. PA, USA, 3–10.

LI, L., SU, H., LIM, Y., AND FEI-FEI, L. 2010. Objects as attributes for scene classification. In *Proceedings of European Conference of Computer Vision, International Workshop on Parts and Attributes*.

LI, L.-J., SU, H., XING, E. P., AND FEI-FEI, L. 2010. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Proceedings of the Neural Information Processing Systems*.

LI, X., SNOEK, C. G. M., AND WORRING, M. 2010. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. 10–17.

LI, X., WANG, D., LI, J., AND ZHANG, B. 2007. Video search in concept subspace: A text-like paradigm. In *Proceedings of ACM International Conference on Image and Video Retrieval*. 603–610.

LIN, J. 2008. Pagerank without hyperlinks: Reranking with related document networks. In *Technical Report LAMP-TR-146*.

LIU, J., LAI, W., HUA, X.-S., HUANG, Y., AND LI, S. 2007. Video search re-ranking via multi-graph propagation. In *Proceedings of ACM Multimedia*. 208–217.

LIU, T.-Y. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval 3,* 3, 225–331.

LIU, X., LI, Z., SHI, Z., AND SHI, Z. 2009. Filter object categories: employing visual consistency and semi-supervised approach. In *IEEE International Conference on Multimedia & Expo*. 678–681.

LIU, Y. AND MEI, T. 2011. Optimizing visual search reranking via pairwise learning. *IEEE Trans. on Multimedia 13,* 2 (April), 280–291.

LIU, Y., MEI, T., AND HUA, X.-S. 2009. CrowdReranking: exploring multiple search engines for visual search reranking. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*. 500–507.

LIU, Y., MEI, T., HUA, X.-S., TANG, J., WU, X., AND LI, S. 2008. Learning to video search rerank via pseudo preference feedback. In *IEEE International Conference on Multimedia & Expo*.

LIU, Y., MEI, T., WANG, M., WU, X., AND HUA, X.-S. 2010. Typicality-based image search reranking. *IEEE Trans. on Circuits System and Video Technology 20,* 5 (May), 749–755.

LIU, Y., MEI, T., WU, X., AND HUA, X.-S. 2008. An optimization-based framework for video search re-ranking. In *Proceedings of ACM SIGMM Workshop on Multimedia Information Retrieval*.

LIU, Y.-T., LIU, T.-Y., QIN, T., MA, Z.-M., AND LI, H. 2007. Supervised rank aggregation. In *Proceedings of International World Wide Web Conference*. 481–489.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60,* 2, 91–110.

LU, W., WANG, J., HUA, X.-S., WANG, S., AND LI, S. 2011. Robust visual reranking via sparsity and ranking constraints. In *Proceedings of ACM Multimedia*. 513–522.

MEI, T., HUA, X.-S., LAI, W., YANG, L., AND ET AL. 2007. MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search. In *TREC Video Retrieval Evaluation Online Proceedings*.

MEI, T. AND RUI, Y. 2009. *Image Similarity*. Chapter in: Ling Liu, M. Tamer Ozsu (Eds.), Encyclopedia of Database Systems, Springer.

MORIOKA, N. AND WANG, J. 2011. Robust visual reranking via sparsity and ranking constraints. In *Proceedings of ACM Multimedia*. 533–542.

MYOUPO, D., POPESCU, A., BORGNE, H. L., AND MOELLIC, P.-A. 2009. Visual reranking for image retrieval over the wikipedia corpus. In *CLEF*.

NA, S.-H., KANG, I.-S., AND LEE, J.-H. 2008. Structural re-ranking with cluster-based retrieval. In *Advanced in Information Retrieval*. 658–662.

NAPHADE, M., SMITH, J. R., TESIC, J., CHANG, S.-F., HSU, W., KENNEDY, L., HAUPTMANN, A., AND CURTIS, J. 2006. Large-scale concept ontology for multimedia. *IEEE MultiMedia 13,* 3, 86–91.

NATSEV, A., HAUBOLD, A., TESIC, J., XIE, L., AND R.YAN. 2007. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of ACM Multimedia*. 991–1000.

NGO, C. W. 2009. VIREO/DVMM at TRECVID 2009: High-level feature extraction, automatic video search, and content-based copy detection. In *TREC Video Retrieval Evaluation Online Proceedings*.

NGO, C.-W., ZHU, S. A., ZHANG, W., TAN, C.-C., YAO, T., PANG, L., AND TAN, H.-K. 2011. VIREOTRECVID 2011: Instance search, semantic indexing, multimedia event detection and known-item search. In *TREC Video Retrieval Evaluation Online Proceedings*.

NISTER, D. AND STEWENIUS, H. 2006. Scalable recognition with a vocabulary tree. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 2161–2168.

NUS-WIDE. http://lms.comp.nus.edu.sg/research/nus-wide.htm.

OLIVARES, X., CIARAMITA, M., AND ZWOL, R. 2008. Boosting image retrieval through aggregating search results based on visual annotations. In *Proceedings of ACM Multimedia*. 189–198.

PARK, G., BAEK, Y., AND LEE, H.-K. 2005. Re-ranking algorithm using post-retrieval clustering for content-based image retrieval. *Information Processing & Management 41,* 2 (March), 177–194.

PASCAL. http://pascallin.ecs.soton.ac.uk/challenges/voc/.

PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J., AND ZISSERMAN, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of CVPR*.

POPESCU, A., MOELLIC, P.-A., KANELLOS, I., AND LANDAIS, R. 2009. Lightweight web image reranking. In *Proceedings of ACM Multimedia*. 657–660.

RENDA, M. E. AND STRACCIA, U. 2003. Web metasearch: Rank vs. score based rank aggregation methods. In *Proceedings of ACM symposium on Applied computing*. 841–846.

ROHINI, U. AND VARMA, V. 2007. A novel approach for re-ranking of search results using collaborative filtering. In *Proceedings of the International Conference on Computing: Theory and Applications*.

ROSE, D. E. AND LEVINSON, D. 2004. Understanding user goals in web search. In *Proceedingos of WWW*. 13–19.

RUDINAC, S., LARSON, M., AND HANJALIC, A. 2009. Exploiting visual reranking to improve pseudo-relevance feedback for spoken-content-based video retrieval. In *Proceedings of International Workshop on Image Analysis for Multimedia Interactive Services*. 17–20.

RUI, Y., HUANG, T. S., AND CHANG, S.-F. 1999. Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation 13,* 10, 39–62.

RUI, Y., HUANG, T. S., ORTEGA, M., AND MEHROTRA, S. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. on Circuits and Video Technology 8,* 5 (September), 644–655.

R.YAN AND HAUPTMANN, A. G. 2003. The combination limit of video retrieval. In *Proceedings of ACM Multimedia*. 339–342.

SALTON, G., WONG, A., AND YANG, C. S. 1975. A vector space model for automatic indexing. *Communications of the ACM 18,* 11, 611–620.

SAVVYSEARCH. Search.com.

SIVIC, J. AND ZISSERMAN, A. 2003. Video Google: a tex retrieval approach to object matching in videos. In *Proceedings of ICCV*.

SMEATON, A. F., OVER, P., COSTELLO, C. J., VRIES, A. P. D., DOERMANN, D., HAUPTMANN, A., HAUPTMANN, E., RORVIG, M. E., SMITH, J. R., AND WU., L. 2002. The TREC2001 video track: Information retrieval on digital video information. In *Proceedings of European Conference on Research and Advanced Technology for Digital Libraries*.

SMEATON, A. F., OVER, P., AND KRAAIJ., W. 2006. Evaluation campaigns and trecvid. In *Proceedings of ACM Workshop on Multimedia Information Retrieval*.

SMEULDERS, A. W. M., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence 22,* 12, 1349–1380.

SMITH, J., JAIMES, A., LIN, C.-Y., NAPHADE, M., NATSEV, A., AND TSENG, B. 2003. Interactive search fusion methods for video database retrieval. In *Proceedings of IEEE International Conference on Image Processing*.

SNOEK, C. G. M., VAN GEMERT, J. C., GEVERS, T., HUURNINK, B., KOELMA, D. C., LIEMPT, M. V., ROOIJ, O. D., VAN DE SANDE, K. E. A., SEINSTRA, F. J., SMEULDERS, A. W. M., THEAN, A. H., VEENMAN, C. J., AND WORRING, M. 2006. The MediaMill TRECVID 2006 Semantic Video Search Engine. In *TREC Video Retrieval Evaluation Online Proceedings*.

SNOEK, C. G. M. AND WORRING, M. 2009. Concept-based video retrieval. *Foundations and Trends in Information Retrieval 4,* 2, 215–322.

SNOEK, C. G. M., WORRING, M., DE ROOIJ, O., VAN DE SANDE, K. E. A., YAN, R., AND HAUPTMANN, A. G. 2008. Videolympics: Real-time evaluation of multimedia retrieval systems. *IEEE MultiMedia 15,* 1, 86–91.

SNOEK, C. G. M., WORRING, M., AND SMEULDERS, A. W. M. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of ACM Multimedia*. Singapore, 399–402.

SNOEK, C. G. M., WORRING, M., VAN GEMERT, J. C., GEUSEBROEK, J. M., AND SMEULDERS, A. W. M. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*. Santa Barbara, USA.

SONG, K., TIAN, Y., GAO, W., AND HUANG, T. 2006. Diversifying the image retrieval results. In *Proceedings of ACM Multimedia*. 707–710.

SONTAG, D., COLLINS-THOMPSON, K., BENNETT, P. N., WHITE, R. W., DUMAIS, S., AND BILLERBECK, B. 2012. Probabilistis models for personalizing web search. In *Proceedings of ACM International Conference on Web Search and Data Mining*. Seattle, WA, USA.

T. L. BERG, D. A. F. 2006. Animals on the web. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1463–1470.

TANG, J., HUA, X.-S., QI, G.-J., AND WU, X. 2007. Typicality ranking via semi-supervised multiple-instance learning. In *Proceedings of ACM Multimedia*. 297–300.

TEEVAN, J., ADAR, E., JONES, R., AND POTTS, M. A. S. 2007. Information re-retrieval: repeat queries in yahoos logs. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*. 151–158.

TIAN, X., TAO, D., HUA, X.-S., AND WU, X. 2010. Active reranking for web image search. *IEEE Trans. on Image Processing 19,* 3 (March), 805–820.

TIAN, X., YANG, L., WANG, J., YANG, Y., WU, X., AND HUA, X.-S. 2008. Bayesian video search reranking. In *Proceedings of ACM Multimedia*. 131–140.

TORRALBA, A. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence 30,* 11, 1958–1970.

TRECVID. http://www-nlpir.nist.gov/projects/trecvid/.

TREVISIOL, M., CHIARANDINI, L., AIELLO, L. M., AND JAIMES, A. 2012. Image ranking based on user browsing behavior. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*. 445–454.

TSENG, Y.-H., TSAI, C.-Y., AND CHUANG, Z.-J. 2008. On the robustness of document re-ranking technique, a comparison of label propagation, knn, and relevance feedback. In *Proceedings of NTCIR-6 Workshop Meeting*.

TURCOT, P. AND LOWE, D. G. 2009. Better matching with fewer features: The selection of useful features in large database recognition problems. In *Proceedings of ICCV Workshop on Emergent Issues in Large Amounts of Visual Data*.

VIDEOSURF. Videosurf.

WANG, G. AND FORSYTH, D. 2008. Object image retrieval by exploiting online knowledge resources. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

WANG, J. AND HUA., X.-S. 2011. Interactive image search by color map. *ACM Transactions on Intelligent Systems and Technology 3,* 1.

WANG, L., YANG, L., AND TIAN, X. 2009. Query aware visual similarity propagation for image search reranking. In *Proceedings of ACM Multimedia*. 725–728.

WANG, M., LI, H., TAO, D., AND LU, K. 2012. Multimodal graph-based reranking for web image search. *IEEE Trans. on Image Processing 21,* 11 (Nov.), 4649–4661.

WANG, M., YANG, K., HUA, X.-S., AND ZHANG, H.-J. 2010. Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia 12,* 8 (Dec.), 829–842.

WANG, M., YANG, L., AND HUA, X.-S. 2009. MSRA-MM: Bridging research and industrial societies for multimedia information retrieval. In *Microsoft Technical Report*.

WANG, S., JING, F., HE, J., Q.DU, AND ZHANG, L. 2007. IGroup: presenting web image search results in semantic clusters. In *Proceeding of SIGCHI Conference on Human Factors in Computing Systems*. 587–596.

WANG, X., LI, Z., AND TAO, D. 2011. Subspaces indexing model on grassmann manifold for image search. *IEEE Trans. on Image Processing 20,* 9 (Sept.), 2627–2635.

WANG, X.-J., ZHANG, L., LIU, M., LI, Y., AND MA, W.-Y. 2010. ARISTA–image search to annotation on billions of web photos. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 2987–2994.

WANG, Y., MEI, T., WANG, J., LI, H., AND LI, S. 2011. JIGSAW: interactive mobile visual search with multimodal queries. In *Proceedings of ACM Multimedia*. 73–82.

WEI, S., ZHAO, Y., ZHU, Z., AND LIU, N. 2009. Multimodal fusion for video search reranking. *IEEE Trans. on Knowledge and Data Engineering*.

WHITE, R. W., RICHARDSON, M., BILENKO, M., AND HEATH, A. P. 2008. Enhancing web search by promoting multiple search engine use. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*. 43–50.

WIKIPEDIA. http://en.wikipedia.org/wiki/click-through_rate.

WILKINS, P., FERGUSON, P., AND SMEATON, A. F. 2006. Using score distribution for query-time fusion in multimedia retrieval. In *Proceedings of ACM Workshop on Multimedia Information Retrieval*. 51–60.

WILKINS, P., SMEATON, A. F., AND FERGUSON, P. 2010. Properties of optimally weighted data fusion in cbmir. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*. 643–650.

WNUK, K. AND SOATTOH, S. 2008. Filtering internet image search results towards keyword based category recognition. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*.

WORRING, M., SNOEK, C. G. M., DE ROOIJ, O., NGUYEN, G. P., AND SMEULDERS, A. W. M. 2007. Recent advances and challenges of semantic image/video search. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1213–1216.

WU, X., HAUPTMANN, A. G., AND NGO, C.-W. 2007. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In *Proceedings of ACM Multimedia*. 168–177.

XU, H., WANG, J., HUA, X.-S., AND LI, S. 2010. Image search by concept map. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*. 275–282.

YAHOO! http://www.yahoo.com/.

YAMAMOTO, T., NAKAMURA, S., AND TANAKA, K. 2007. Rerank-by-example: Efficient browsing of web search results. *Database and Expert Systems Applications*, 801–810.

YAN, R. AND HAUPTMANN, A. G. 2004. Co-retrieval: a boosted reranking approach for video retrieval. In *Proceedings of ACM International Conference on Image and Video Retrieval*.

YAN, R. AND HAUPTMANN, A. G. 2006. Probabilistic latent query analysis for combining multiple retrieval sources. In *Proceedings of ACM SIGIR conference on Research and Development in Information Retrieval*. 324–331.

YAN, R. AND HAUPTMANN, A. G. 2007a. Query expansion using probabilistic local feedback with application to multimedia retrieval. In *Proceedings of ACM conference on Information and Knowledge Management*. 361–370.

YAN, R. AND HAUPTMANN, A. G. 2007b. A review of text and image retrieval approaches for broadcast news video. *Information Retrieval 10,* 4-5 (October), 445–484.

YAN, R., HAUPTMANN, A. G., AND JIN, R. 2003. Multimedia search with pseudo-relevance feedback. In *Proceedings of ACM International Conference on Image and Video Retrieval*.

YAN, R., YANG, J., AND HAUPTMANN, A. 2004. Learning query-class dependent weights in automatic video retrieval. In *Proceedings of ACM Multimedia*. 548–555.

YANAGAWA, A., CHANG, S. F., KENNEDY, L., AND HSU, W. 2007. Columbia university's baseline detectors for 374 lscom semantic visual columbia university's baseline detectors for 374 LSCOM semantic visual concepts. In *Columbia University ADVENT technical report*.

YANG, K., WANG, M., HUA, X.-S., AND ZHANG, H.-J. 2010. Social image search with diverse relevance ranking. In *Proceedings of International MultiMedia Modeling Conference*. Chongqing, China, 174–184.

YANG, Y., NIE, F., XU, D., LUO, J., ZHUANG, Y., AND PAN, Y. 2012. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans. on Pattern Analysis and Machine Intelligence 34,* 4 (April), 723–742.

YOUTUBE. http://www.youtube.com/.

ZAVESKY, E. AND CHANG:, S.-F. 2008. CuZero: embracing the frontier of interactive visual search for informed users. In *Proceedings of ACM Multimedia Information Retrieval*. 237–244.

ZHA, Z.-J., YANG, L., MEI, T., WANG, M., AND WANG, Z. 2009. Visual query suggestion. In *Proceedings of ACM Multimedia*. Beijing, China, 15–24.

ZHA, Z.-J., YANG, L., MEI, T., WANG, M., WANG, Z., CHUA, T.-S., AND HUA, X.-S. 2010. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Transactions on Multimedia Computing, Communications, and Applications 6,* 3.

ZHOU, W., LU, Y., LI, H., SONG, Y., AND TIAN, Q. 2010. Spatial coding for large scale partial-duplicate web image search. In *Proceedings of ACM Multimedia*. 511–520.

ZHOU, X. S. AND HUANG, T. S. 2002. Relevance feedback in content-based image retrieval: Some recent advances. *Information Sciences 148,* 1-4 (Dec.), 129–137.

ZHOU, X. S. AND HUANG, T. S. 2003. Relevance feedback for image retrieval: a comprehensive review. *Multimedia Systems 8,* 6 (April), 536–544.

ZHU, G., YAN, S., AND MA, Y. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the ACM International Conference on Multimedia*. 461–470.

ZHUANG, J., MEI, T., HOI, S. C. H., HUA, X.-S., AND LI, S. 2011. Modeling social strength in social media community via kernel-based learning. In *Proceedings of ACM Multimedia*. 113–122.

ZHUANG, J., MEI, T., HOI, S. C. H., XU, Y.-Q., AND LI, S. 2011. When recommendation meets mobile: contextual and personalized recommendation on the go. In *Proceedings of ACM International Conference on Ubiquitous Computing*. Beijing, China, 153–162.

ZHUANG, Z. AND CUCERZAN, S. 2006. Re-ranking search results using query logs. In *Proceedings of ACM conference on Information and Knowledge Management*. 860–861.

ZITOUNI, H., SEVIL, S., OZKAN, D., AND DUYGULU, P. 2008. Re-ranking of web image search results using a graph algorithm. In *Proceedings of International Conference on Pattern Recognition*. 1–4.

ZLOOF, M. M. 1975a. Query by example. In *Proceedings of AFIPS National Compute Conference*. 431–438.

ZLOOF, M. M. 1975b. Query-by-example: the invocation and definition of tables and forms. In *Proc. of VLDB*. 1–24.