Contextual Image Search

Jingdong Wang, Wenhao Lu, Shengjin Wang, Xian-Sheng Hua and Shipeng Li

MSR-TR-2010-84

June 27, 2010

Abstract

In this paper, we propose a novel image search scheme, *contextual image search*. Different from conventional image search schemes that present a separate interface (e.g., text input box) to allow users to submit a query, the new search scheme enables users to search images by only masking a few words when they are reading through Web pages or other documents. Rather than merely making use of the explicit query input that is often not sufficient to express the search intent, our approach explores the context information to better understand the search intent, and expects to obtain better search results, through two key ways: query augmenting and search results reranking using context. To the best of our knowledge, this is the first attempt to conduct image search with both textual and visual context. Beyond contextual Web search, the context in our case is much richer and includes images besides texts. Experiments show that the proposed scheme makes image search more convenient and the search results are more relevant to user intention.

Index Terms

Image search, textual and visual context, contextual query augmentation, contextual reranking.

I. Introduction

Image search engines have been playing important roles for consumers to find desired images. Our investigation shows that search queries are often issued when people are browsing Web pages and working on emails, or other documents. In such cases, the context of the query from the associated document is useful to describe the user interest more clearly, e.g., disambiguate the query and hence help to capture the search intent. Better image search results are naturally expected with the help of the context. In this

Jingdong Wang, Xian-Sheng Hua and Shipeng Li are with the Media Computing Group, Microsoft Research Asia, Beijing, P.R. China. E-mail: {jingdw, xshua, spli}@microsoft.com

Wenhao Lu,and Shengjin Wang are with Department of Electronic Engineering, Tsinghua University, Beijing, P.R. China.

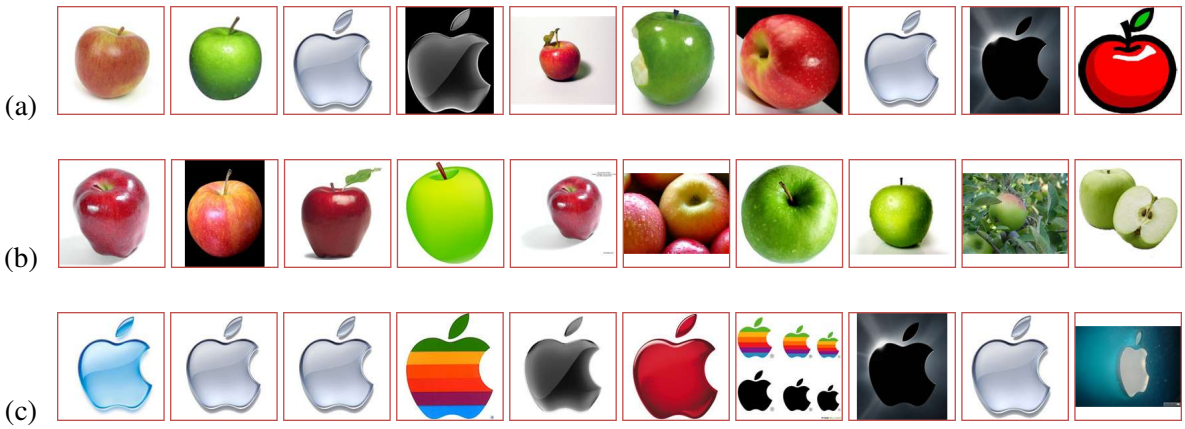


Fig. 1. Illustration of contextual image search to remove the ambiguity of the query "apple". (a) corresponds to the results from the raw query "apple", (b) corresponds to the contextual search results in the context of introducing fruits, and the textual context includes *fruit*, *stem*, *knobby*, and so on, and (c) corresponds to the contextual search results in the context of introducing the Apple Inc., and the textual context include *company*, *logo*, *iphone*, and so on.

paper, we propose a novel image search scheme, *contextual image search*, which enables users to issue a query by masking textual words in a document and then its context information is combined to find more relevant images. To the best of our knowledge, this is the first work on image search using context. Beyond contextual Web search, the context in our case is richer and contains visual components, and hence contextual image search is more challenging.

Let's look at several examples to illustrate how the context information facilitates image search. On the one hand, the context is capable of removing the possible ambiguity of a query. Given a textual query "apple" masked by a user, e.g., from http://www.mahalo.com/apple-fruit, merely from the word, it is not clear whether it refers to a fruit or a product logo. Borrowing its context information, it is natural that the essential intent of the user is a fruit. The contextual image search results using the proposed technique are shown in Fig. 1(b). In the context of the Web page, http://en.wikipedia.org/wiki/Apple_Inc., the query "apple" refers to a product logo, and the contextual image search results correspond to Fig. 1(c).

On the other hand, a user-masked query, even without ambiguities, is usually insufficient to express the search intent. For example, the words "George Bush", masked by the user when he is reading the Web page whose content is relevant to jokes of George Bush, http://www.gwjokes.com/, has large probability to (of course may not always) mean to search funny images as the *textual context* includes words such as *joke*, *fool*, *prank*, and so on. Then using the context to rerank search results, called contextual reranking, the corresponding reranking results will be more relevant to user intent, as shown Fig. 2(b). Compared the

Fig. 2. Illustration of reranking image search results using textual context to help find images that better match the user intent implied from the textual context. (a) corresponds to search results of "George Bush", (b) shows the reranking results using the context from the document describing jokes of George Bush, and the textual context includes *joke*, *fool*, *prank*, and so on, and (c) shows the results from one commercial image search engine with the query "Funny George Bush". The results in (b) show that reranking using textual context can indeed take effect and even make the results better than the results in (c) obtained from the explicit query "Funny George Bush".

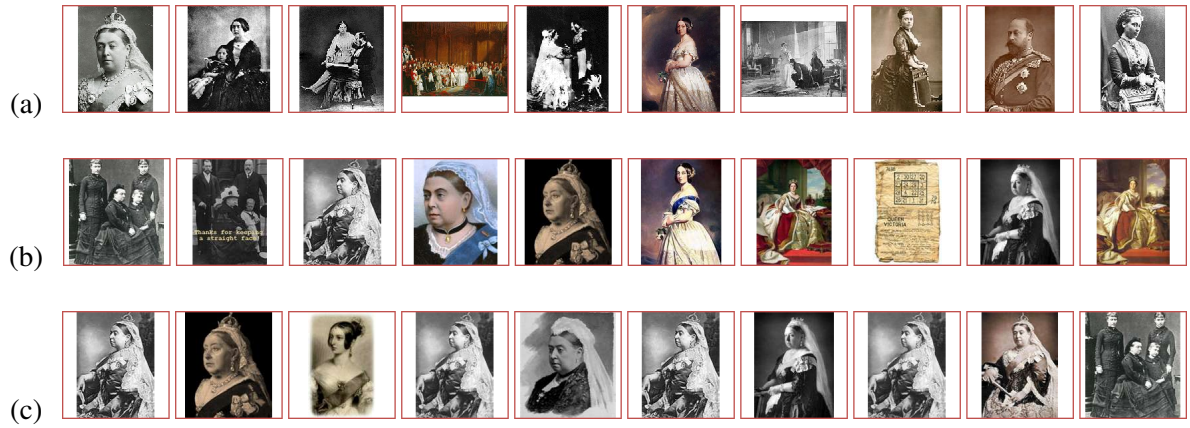


Fig. 3. Illustration of reranking image search results using visual context. (a) shows the visual context of "Queen Victoria", (b) shows the image search results of "Queen Victoria" without contextual reranking, and (c) shows the image search results of "Queen Victoria" with visual contextual reranking, which are more consistent with the visual context in (a).

results with the manually-created query "Funny George Bush" shown in Fig. 2(c), the textual contextual reranking results look more satisfactory.

As another illustration that context can help express the search intent more clearly, Figs. 3(a), 3(b) and 3(c) show an example in which *visual context* for reranking takes effect when masking a textual

query, “Queen Victoria” from http://en.wikipedia.org/wiki/Queen_Victoria. The visual context from this page is shown in Fig. 3(a), which reflects the user interest on classic images. Image search results without the help of the visual context and with its help are shown in Figs. 3(b) and 3(c), respectively. It can be observed that the results in Fig. 3(c) are more consistent with the content of the document and the style of the visual contexts.

To utilize context to make image search results more relevant to user intent, we propose a contextual image search framework. It consists of the following steps. First, we extract the context associated with the user input from the document. The context consists of two types: textual context and visual context. Second, we explore the textual context to remove the possible ambiguity for query augmentation. Third, the augmented query is used to perform first-round image search using the text-based image search technique to get a bunch of images. Finally, the textual and visual context information is used to rerank the images to make the results more relevant to user intent.

To summarize, this paper offers the following key contributions:

- To the best of our knowledge, our work is the first attempt to perform image search with the help of the context.
- We propose a contextual query augmentation manner to remove the query ambiguity by using the textual context. Particularly, we use the context to help select the most probable augmented query from the candidate augmented queries, instead of mining the augmented query only from the context.
- We present a contextual reranking way, which uses the context to rerank search results, to make search results more relevant to user intent. Beyond contextual Web search, the visual context is additionally explored for reranking.

A. Related Work

A lot of efforts have been conducted to improve image search by helping user more clearly indicate the search intent and making use of the search intent effectively [3], [9], [17]. Most of them focus on presenting interfaces to enable users to express the search intent more conveniently and more clearly, but they do not explore any context information for image search. This type of image schemes can be categorized as *image search without context*. In the following, we review the existing image search schemes without context and *image search in context* (but without using context), and then pose the promising scheme, *image search with context*.

Image search without context

The widely used commercial image search engines, including Google image search, Yahoo! image search

and Microsoft Bing image search, provide an explicit interface, a *text input box*, to enable users to issue textual queries and then rely only on the input to search the image database. However, a single textual query is frequently not sufficient to clearly indicate the search intent.

Some other image search engines provide features to allow users to upload or draw an image, i.e., issue a *visual query*, to illustrate the search intention visually. Tineye[1] enables users to update an image to trigger a content based image retrieval. Because of the gap between the image feature and the semantic content, such techniques succeed only in finding duplicate or near-duplicate images. An online similar image search engine[2] provides the feature of image search by sketch. However, merely using a visual example as the query for image search does not suffice because visually similar images do not guarantee to have similar semantic content.

To exploit the above two schemes for image search together, Google image search and Microsoft Bing image search provide "Find similar images" and "Show similar images", respectively. First, a user may issue a textual query to get the text-based image search results. Then, to help indicate the search intention more clearly, the user may select an image from the search results that is closer to what the user is interested in and use it to reorder the images according to the visual similarities. But it still suffers from the semantic gap as it is not clear what in an example image is indeed the search intention.

Besides, commercial image search engines provide *explicit filters* to help users to clarify the search intent into some scopes. For example, Google image search presents options to allow users to find images of different sizes, different types (e.g., face, photo, clip art, and line drawing), or different dominant colors. Microsoft Bing image search additionally provides options to find images with faces or head & shoulders.

There are many other efforts on *interactive query indication* to improve image search. Relevance feedback [15], [2], [11], [22] is one of the most traditional techniques, which allows users to clarify the search intent by selecting a few positive and/or negative images. A so-called CuZero system [23] is proposed to embrace the frontier of interactive visual search for informed users, and specifically an interactive interface is presented to enable users to navigate seamlessly in the concept space at-will and simultaneously while displaying the results corresponding to arbitrary permutations of multiple concepts in real time. The CueFlik system [6] allows end-users to provide examples of images to quickly create their own rules for reranking the images. The SkyFinder system [18] presents an attribute based image search system and attempts to search a desired sky image by building a sky graph based on the sky attributes

[1]http://www.tineye.com

[2]http://www.gazopa.com

to help users specify the interest, such as "a landscape with rich clouds at sunset". An interactive image search scheme [21] is proposed to help users find desired images with the requirement on how the concepts or colors are spatially distributed.

Image search in context

One of the most common scenarios where users want to perform image search is reading or writing a document which make users be interested in some matters related to the document. Therefore, to facilitate image search for this scenario, image search engines, including Google image search, Microsoft Bing image search, and Tineye, provide browser (such as IE and Firefox) plugins to enable users to mask keywords or select animage in Web pages so an image search action can be performed without the necessity of issuing the query at the home page of an image search engine. Such schemes definitely accelerate image search. However, the very useful information besides the user input, called context, available from the document, is not explored for image search.

Image search with context

The concept "context" have different meanings in different application scenarios. For the scenario of object detection [4] in computer vision, the context is usually used to describe the interaction of different objects/concepts, for example, the co-occurrence of different objects in the similar scene. In the scenario of mobile search, the situation of performing image search, including location, time, action history, and search history, can be regarded as the context. In the scenario of personalized search, the personal information, including the personal interest, the search history and so on, are regarded as the search context. In this paper, we regard the surrounding textual and visual information of a query from the document as the context and focus on exploiting such context for image search, but it should be noted that the methodology in this paper can be applied to image search with other types of contexts.

Context information has other usages, e.g., user interests prediction [20] , media processing to bridge the sematic gap [7], in-image advertising [14], and image annotation using duplicate images [19]. Context has also been explored for Web search [5], [8]. But the context is only limited in textual components, and the rich visual context is not exploited. To the best of our knowledge, this paper is the first attempt to exploit contexts for image search.

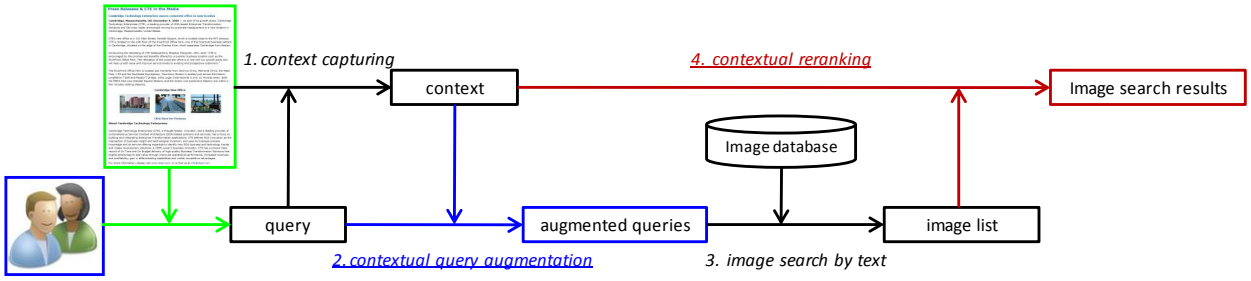


Fig. 4. System overview of contextual image search.

II. System Overview

The whole contextual image search system consists of two subsystems: database construction system and ranking system. In the database construction system, an image database, which may be a local database, or a global database that is crawled from the Internet, is built and organized as the search database. We extract the visual feature for each image, a bag-of-visual-words (BoW) representation in our implementation, and its text description from the document holding this image, which is obtained using the context capturing scheme that is described later. Then, we build an indexing system, dependently on the text description, which makes it efficient that images can be searched with a textual query given. Besides, each image is associated with a static rank, which is computed, for instance, from the static rank of the Web page holding this image.

The ranking system is illustrated in Fig. 4. The typical input of this system is a few textual keywords masked by the user from a document. The output is a list of ranked images that are from the image database. The remaining system consists of four modules: context capturing, contextual query augmentation, image search by text, and contextual reranking.

1) *Context capturing* is to find a set of textual keywords, called textual context, and a set of images, called visual context, from the document, according to the spatial position of the query in the document.
2) *Contextual query augmentation* is to use the context to refine the query. The textual context is used to augment the query by removing the ambiguity.
3) *Image search by text* is to search images using the augmented query based on the text-based image search technique.
4) *Contextual reranking* aims to make use of both the textual and visual contexts to promote images that have similar contexts with the context of the query.

The key novelty among the four modules lies in the second and fourth modules, which make use of context to improve the relevance of image search results.

Our system also supports the search scenario in which an image is selected as the query. In this case, our experiment shows that the search results are good by performing CBIR if the bag-of-visual-words representation is used to describe an image and the text-based search technique is used to search visual words, which is also demonstrated in the "Show more sizes" feature in Microsoft Bing image search. Moreover, the search-to-annotation technique [19] can also be used in this case to help image search by mining the annotation from the textual contexts of the duplicate images. In addition, our system supports the hybrid query, e.g., a pair of masked words and selected image, from a document. In this case, the search is completed by performing the first three step in our system and then doing visually similar image search, using the technique similar to "Show similar images" in Microsoft Bing image search or "Find similar images" in Google image search. All the above are not big contributions, and hence we do not describe them in detail.

III. Approach

A. Notation

The document that the user is reading is denoted as D, and it may contains texts, images, and even videos. The raw query, masked by the user from D, is denoted as q. An image in the document is denoted as I^c. The context is denoted as C. It should be noted that the context may contain different types of components and will be described in detail later.

The image in the database is denoted by I_k, and it is associated with a pair of features, a visual feature $\mathbf{h}_k^v$ and a textual feature $\mathbf{h}_k^t$. We represent the image using the popular BoW representation. To obtain a BoW representation, we extract a set of maximally stable extremal regions (MSERs) [13] for each image, represent each region by a scale-invariant transform feature (SIFT) descriptor [10], and then quantize each SIFT descriptor, with a vector quantization algorithm. The BoW representation for an image description will lead to the benefit that the fast indexing and retrieval algorithms used in text search can be directly adopted for image retrieval, which is shown in the video google technique [16]. The textual feature of an image is obtained by using a vector space model to describe its associated textual contexts, and such a textual feature is widely used in existing commercial image search engines.

B. Problem Formulation

In this subsection, we formulate our problem and show that the mathematical decomposition will be equivalent to four modules described in Sec. II. The contextual image search problem can be formally formulated as follows. Given a query q and the associated document D, the goal is to order the images $\mathcal{I} = \{I_k\}_{k=1}^N$ by computing relevance scores $\mathcal{R} = \{r_k\}_{k=1}^N$ of their visual and contextual information and with the query q and the document D. Mathematically, we define the relevance score as the probability of I_k conditioned on the query, its associated document and the other images in the search database,

$$r_k = P(I_k|q, D, \mathcal{I}_{/\{I_k\}}). \tag{1}$$

This conditional probability can be computed from the joint probability,

$$P(I_1, \cdots, I_k, \cdots, I_N|q, D). \tag{2}$$

To formulate this joint probability, we introduce the intermediate variables, an augmented query, $\bar{q}^*$, and a context, C^*, which are obtained from the user input and the document. Then Eqn. (2) can be written as follows,

$$P(I_1, \cdots, I_N|q, D) \approx P(I_1, \cdots, I_N|\bar{q}^*, C^*). \tag{3}$$

This transform is essentially based on the Bayesian estimation. The following equation will hold

$$P(I_1, \cdots, I_N|q, D) \tag{4}$$

$$= \sum\nolimits_{\bar{q},C} P(I_1, \cdots, I_N|\bar{q}, C)P(\bar{q}, C|q, D) \tag{5}$$

$$\approx P(I_1, \cdots, I_N|\bar{q}^*, C^*), \tag{6}$$

if $P(\bar{q}^*, C^*|q, D)$ is large enough. Here,

$$(\bar{q}^*, C^*) = \arg\max\nolimits_{\bar{q},C} P(\bar{q}, C|q, D). \tag{7}$$

For convenience, we may drop * in the following description. With this decomposition, our problem can be solved in two steps: (1) processing the document to discover the context and the augmented query, and (2) ranking the images with the discovered context and augmented query.

To obtain the context and the augmented query, we transform the probability $P(\bar{q}, C|q, D)$ as follows,

$$P(\bar{q}, C|q, D) = P(\bar{q}|C, q)P(C|q, D). \tag{8}$$

This factorization is reasonable because (1) the context is only dependent on the input query and the document and (2) the augmented query can be approximately determined by the query and its context.

Therefore, the process can be finished in two steps, context extraction, i.e., discovering the context C^* according to q and D, and contextual query augmentation, i.e., finding $\bar{q}^*$, so that $P(\bar{q}^*|C^*, q)$ is maximized.

To evaluate $P(I_1, \cdots, I_N|\bar{q}, C)$, we propose a two-step scheme. The first step is to perform text-based image search using the augmented query $\bar{q}$. The second step is to perform a reranking step by exploiting the context information. Essentially, the two-step scheme is equivalent to the following decomposition,

$$P(I_1, \cdots, I_N|\bar{q}, C) = \prod\nolimits_{I_k} P(I_k|\bar{q}, C) \tag{9}$$

$$\propto \prod\nolimits_{I_k} P(\bar{q}, C|I_k)P(I_k) \tag{10}$$

$$= \prod\nolimits_{I_k} P(C|I_k)P(\bar{q}|I_k)P(I_k). \tag{11}$$

Here, $\propto$ holds with $\bar{q}$, C given. The terms, $P(\bar{q}|I_k)P(I_k)$, corresponds to the first step, and the rank model of existing text-based image search engines is essentially equivalent to this model, with $P(I_k)$ being the static rank. The term, $P(C|I_k)$, corresponds to the second step, contextual reranking.

In summary, the implementation of our system consists of the following modules: context capturing ($P(C|D, q)$), contextual query augmentation ($P(\bar{q}|C, q)$), image search by text ($P(\bar{q}|I_k)P(I_k)$), and contextual reranking ($P(C|I_k)$). In the following, we describe the four modules in detail.

C. Context Capturing

Context capturing aims to discover the visual and textual context for a query to rank images, as well as the context of an image for database construction. The textual and visual contexts, are denoted by C_t and C_v. The visual context of a query consists of a set of images from the same document and their local textual contexts $C_v = \{C_v^v, C_v^t\}$.

To extract the textual context of an image, we propose to use the vision-based page segmentation (VIPS) algorithm [1]. VIPS can extract the sematic structure of a Web page based on its visual representation. The VIPS algorithm first extracts all the suitable blocks from the Document Object Model tree in html, and then finds the separators between these blocks, where the separators denote the horizontal or vertical lines in a page visually crossing without blocks. Based on these separators, a Web page can be represented by a semantic tree in which each leaf node corresponds to a block. In this way, contents with different topics are distinguished as separate blocks in a Web page.

The VIPS algorithm can be naturally used for surrounding text extraction. For an image, we find the surrounding text from its corresponding block as a part of its textual context. Specifically, the textual

context of an image includes $C_t = \{C_t^1, C_t^2, C_t^3\}$, with C_t^1, C_t^2, C_t^3 corresponding to image name and its description with a high weight, page title and document title with a middle weight, and other surrounding text with low weight, respectively. The images in the database are actually processed using the above scheme to extract their associated textual contexts.

The extraction of textual context for masked keywords is relatively easy. Besides the page title and document title, the neighboring words of the masked keywords are viewed as surrounding texts, and used as a part of the textual context, called local context. In this case, the textual context includes $C_t = \{C_t^2, C_t^3\}$.

D. Contextual Query Augmentation

We propose to make use of the textual context to augment the textual query to remove possible ambiguities. The augmented queries for a textual query q can be formed by combining q and the keyword from the textual context C_t. This is an optional solution to use them as queries to search the images. However, obviously. augmented queries obtained in this way may not always be meaningful, and then the returned images may not be satisfactory. Therefore, rather than exploiting the textual context to directly augment queries, we use them as a support to disambiguate the query. To this goal, we first build a set of candidate augmented queries, $Q = \{\bar{q}_1, \bar{q}_2, \cdots, \bar{q}_n\}$, which can remove the ambiguities of the query q. Then, we vote each augmented query in Q using the context C_t.

The following presents a mathematical derivation and its implementation algorithm. We find an optimal augmented query, by checking the posterior of each candidate augmented query, given the context obtained from the document. Mathematically, the posterior can be computed as

$$P(\bar{q}|C_t, q) = \frac{P(C_t|\bar{q}, q)P(\bar{q}|q)}{P(C_t|q)} \tag{12}$$

$$\propto P(C_t|\bar{q}, q)P(\bar{q}|q), \tag{13}$$

where $P(C_t|\bar{q}, q)$ is the likelihood of the augmented query with respect to the context C_t, and $P(\bar{q}|q)$ is the prior of the augmented query $\bar{q}$. The denominator $P(C_t|q)$ can be ignored because it is independent on $\bar{q}$ and hence is viewed as a constant.

In our implementation, the prior is computed as

$$P(\bar{q}|q) = \frac{1}{|Q|}, \tag{14}$$

if $\bar{q} \in Q$, and $P(\bar{q}|q) = 0$ if $\bar{q} \notin Q$. This is reasonable because we have no any bias on the candidate augmented queries $\bar{q}$ if without any context support.

The likelihood $P(C_t|\bar{q}, q)$ is essentially to evaluate the relevance between the context C_t and $\bar{q}$. To this end, we borrow the idea of evaluating the relevance between queries and documents. First, we extend the context to get $\bar{C}_t$, which is obtained by expanding the words in C, e.g., using synonyms, stemming, and so on [12]. Then, we adopt the Okapi BM25 algorithm that is used by search engines to rank documents with a search query. Given a candidate augmented query $\bar{q}$ containing n terms $\{q, q_1, q_2, \cdots, q_n\}$, with q being the raw query and q_i being the expanded words, the BM25 score between the extended context and this query is computed as

$$\text{score}(\bar{q}, C_t) = \sum_{i=1}^{n} \frac{idf(q_i) \times tf(q_i, \bar{C}_t) \times (k+1)}{tf(q_i, \bar{C}_t) + k(1 - b + b \times ndl(C_t))}, \tag{15}$$

where $tf(q_i, \bar{C}_t)$ is the term frequency of q_i in $\bar{C}_t$, $idf(q_i)$ is the inverse document frequency, $ndl(C_t) = \frac{|C_t|}{|\tilde{C}_t|}$ is normalized textual context length, $|C_t|$ and $|\tilde{C}_t|$ indicate the context length of C_t and the average context length in the database, and k and b are two parameters and chosen as $k = 2.0$ and $b = 0.75$ in our implementation. Then the likelihood is calculated so that $P(C_t|\bar{q}, q) \propto \text{score}(\bar{q}, C_t)$.

Then, the optimal augmented query is selected as

$$\bar{q}* = \arg\max_{\bar{q}} P(C_t|\bar{q}, q) P(\bar{q}|q) \tag{16}$$

$$= \arg\max_{\bar{q}} P(C_t|\bar{q}, q). \tag{17}$$

In the cases that $P(C_t|\bar{q}, q)$ for all $\bar{q}$ has the similar values or that $P(C_t|\bar{q}*, q)$ is very small, i.e., the context is not enough to disambiguate the query, we keep the original raw query, $\bar{q} = q$.

E. Image Search by Text

Given the augmented query, we perform text-based image search, by matching the augmented query with the textual context of each image in the database. The text-based image search returns a list of images, ranked by the static score $P(I_k)$ and the relevances $P(\bar{q}_t|I_k)$ between the textual contexts of images and the query.

F. Contextual Reranking

Contextual query augmentation explores only the texts in the textual context that are related to the candidate augmented queries and useful to disambiguate the query. The remaining texts, e.g., describing the situation, and visual contextual are also very important and useful for image ranking. In the example of Fig. 2, the words, *joke*, in the textual context can help promote funny images. Therefore, this step, contextual reranking, aims to exploit the context information that presents hints on the search intent to

reorder the search results from text-based search so that the top images are more consistent to the search intent.

Contextual reranking is different from the previous work on visual reranking. Visual reranking explores the visual similarities, reorders the visually similar images together and at the same time the original order is kept as far as possible. Instead, contextual reranking aims to promote images that match the context better, and also wants to keep the original order as much as possible. Of course, we may also explore the visual similarities for the contextual reranking. But we find that the visual reranking does not get the search results improved in our case, especially when the visual context takes effect, and moreover visual reranking is a little time-consuming due to costly pairwise similarity computation. Therefore, in our implementation, we investigate only context for reranking.

Textual contextual reranking

Contextual reranking aims to evaluate the probability, $P(C|I)$. We decompose it into two terms, $P(C|I) = P(C_t|I)P(C_v|I)$, to compute the reranking scores from the textual and visual contexts, respectively. Textual contextual reranking, to evaluate $P(C_t|I)$, is conducted as follows. The textual context except the query related context is helpful to describe the situation that may be of user interest and is related to the search intent (as illustrated in Fig. 2). In our implementation, the score from the textual context is computed as the document similarity between the textual context and the text description of images in the search results, and specifically evaluated by the BM25 algorithm. The computation formula is similar to Eqn. (15). But differently, the score is computed between two reduced contexts, which is obtained by discarding the augmented query related words in the original textual contexts, because the augmented query related textual words have been explored in contextual query augmentation for text-based search and the remaining words are useful for reranking. Suppose the similarity from textual context is denoted by $\mathrm{sim}_t(C_t, I_k)$, The probability $P(C_t|I_k)$ is computed as $P(C_t|I_k) \propto \exp(\mathrm{sim}_t(C_t, I_k))$

Visual contextual reranking

Visual contextual reranking, corresponding to evaluate $P(C_v|I)$, aims to promote the images that are similar to the images from the document that are relevant to the textual query. The similarity from the visual context can be evaluated from the bag-of-visual-words representation. Furthermore, to obtain better visual-words presentation, we perform an augment step, which is based on the following observations. 1) The images in a document usually have similar semantic content if their local textual contexts are very relevant to the query, and 2) if each local feature in the bag-of-visual-word representation is homogeneously regarded and not differentiated, some local features that may be irrelevant to the semantic

content, e.g., coming from the background, will influence the performance.

In our implementation, we view each image as a document, and borrow the inverse document frequency (idf) technique to weight different visual words, which is often used in information retrieval and text mining to learn the weight for each word. First, we adopt a textual context based filter scheme to filter out images whose semantic contents may not be relevant to the visual query. To this end, we compute the similarity of local textual context of each image in the document with the textual query. The similarity is also evaluated based on the BM 25 algorithm, similar to Eqn. (15). Then, if the similarity is larger than a threshold, the corresponding image will be counted for the idf computing. Specifically, the weight of a visual word in each image is set as $w_i = tf(f_i)/idf(f_i)$ with f_i corresponding to a visual word, which is different from the conventional tf-idf weighting that aims to remove the meaningless words, while we aim to find the common pattern in the images, which is important for visual similarity computation in our case.

After the augmentation step, we turn to compute the similarity score between visual contexts and each image I_k in the search results from text-based image search. Suppose I_i^c is an image in the filtered visual context and its bag-of-words representation is written as a histogram vector $\mathbf{h}_i^c$, then the similarity between I_i^c and I_k is computed as the weighted histogram intersection,

$$\mathrm{sim}_v(I_i^c, I_k) = \sum\nolimits_j \min(h_{ij}^c, h_{kj}) w_j. \tag{18}$$

Then, we compute the similarity between the visual context and an image in the search results by finding the largest one among the similarities between images in the filtered visual context and the image,

$$\mathrm{sim}_v(C_v, I_k) = \max\nolimits_{I_i \in C_v} \mathrm{sim}_v(I_i^c, I_k)\delta_{[I_i]}. \tag{19}$$

Here $\delta_{[I_i]}$ is an indicator to show if I_i lies in the filtered visual context. The probability $P(C_v|I_k)$ is set as $P(C_v|I_k) \propto \exp(\mathrm{sim}_v(C_v, I_k))$.

G. Overall Ranking

For one image I_k, its probability conditioned on the context and the augmented query can be computed by

$$\begin{aligned} P(I_k|C, \bar{q}) \propto &\quad P(C|I_k)P(\bar{q}|I_k)P(I_k) &(20)\\ = &\quad P(C_t|I_k)P(C_v|I_k)P(\bar{q}|I_k)P(I_k) &(21)\\ \propto &\quad \exp\left[\lambda_1 \mathrm{sim}_t + \lambda_2 \mathrm{sim}_v + \lambda_3 \mathrm{score}_i\right], &(22) \end{aligned}$$

where sim_t, sim_v and score_i, corresponds to the similarities from the textual context, the visual context, and text-based image search, respectively, and λ_1, λ_2, and λ_3 are their associated weights, to adjust the degrees that we trust the three factors. $\lambda_1 = 0.2$, $\lambda_2 = 0.2$, and $\lambda_3 = 1$ in our implementation.

IV. EXPERIMENT

In our experiment, we implement one instance of contextual image search, specifically for the Web page. To make our system useable, we crawled image search results with textual queries from one existing commercial image search engine. There are 5000 top search queries and their candidate augmented queries (about 10000). For each query, we crawl about 1000 images, and totally we got about 15,000,000 images. For each image, we analyze its Web page and get its textual context using the aforementioned context capturing scheme.

Data set

We collect the data set to evaluate contextual image search from the search logs that were recorded when users tried the proposed search system. Specifically, we present the contextual system to the users, and show how to use it to perform image search using the three examples shown in Figs. 1, 2 and 3. Then we allow the users to play the system for about two hours. During the search process, we record each search session, including the Web page URL, and the selected query, into search logs. After all trials, we process the search logs and arrange search sessions together by merging the search sessions with the same raw query as a group of search queries. Totally, we got about 100 groups of search queries. On average, there are about 4 individual search sessions for each group. We randomly sampled 50 groups among them to build the ground truth for quantitative evaluation. These 50 groups of queries include different types, e.g., famous person, site, and products.

Ground truth

The ground truth of the search results are built as follows. For each search session, we ask annotators to label the results: contextual image search result and image search result only with the raw query. Besides the raw query, we also show the document to the annotators to allow annotators to get familiar with the context. To differentiate different relevance degrees, we adopted a graded relevance scale, and use four levels from level 0 (the least relevant) to level 3 (the most relevant). We asked 5 users to judge each search session, and then selected the most frequent level as the final level for each image. To avoid any bias on the labeling, those users were selected such that they have no special knowledge on image search and are initially unknown about the proposed technique.

Evaluation criteria

To evaluate the performance, we use the normalized discounted cumulative gain (nDCG) measure. DCG measures the usefulness, or gain, of a document based on its position in the result list. The gain is accumulated cumulatively from the top of the result list to the bottom with the gain of each result discounted at lower ranks. Two assumptions of DCG measure are that highly relevant documents are more useful when appearing earlier in a search engine result list (have higher ranks) and that highly relevant documents are more useful than marginally relevant documents, which are in turn more useful than irrelevant documents. Comparing a search engine performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position (e.g., p) should be normalized across queries. This is done by sorting documents of a result list by the ground truth, producing an ideal DCG at position p. Mathematically, nDCG at a rank position p is calculated as

$$\mathrm{nDCG}(p) = \frac{\mathrm{DCG}(p)}{\mathrm{iDCG}(p)}, \tag{23}$$

$$\mathrm{DCG}(p) = \sum\nolimits_{i=1}^{p} \frac{r_i}{\log_2(i+1)}, \tag{24}$$

where r_i is the graded relevance of the result at position i, and calculated as $r_i = 2^{c_i} - 1$, with c_i the groundtruth level of the image at position i, $\mathrm{iDCG}(p)$ is an ideal DCG at position p. The nDCG values for all queries can be averaged to obtain a measure of the average performance for many queries.

A. Quantitative Evaluation

Given the raw query, the context information can be used to disambiguate the raw query to make search intention clearer or present more hints to make search intention more specific, by contextual query augmentation and contextual reranking, respectively. In the following, we compare the schemes, only using contextual query augmentation, only using contextual reranking, and using both the two schemes, with the baseline scheme that directly performs image search with the raw query.

We report nDCG scores at different positions of the schemes, only contextual query augmentation, only contextual reranking, and the whole scheme, for image search. Here we present the scores for the first 40 images because the investigation shows that users often only check the first 40 images (i.e., the first two pages of results). In addition, we also report the result of the baseline algorithm, using the raw query for image search. The comparison results are shown in Fig. 5. From this figure, it can be observed that both contextual query augmentation and contextual reranking can individually make search results improved. To have a deeper view of contextual reranking, we also present the nDCG curves of reranking

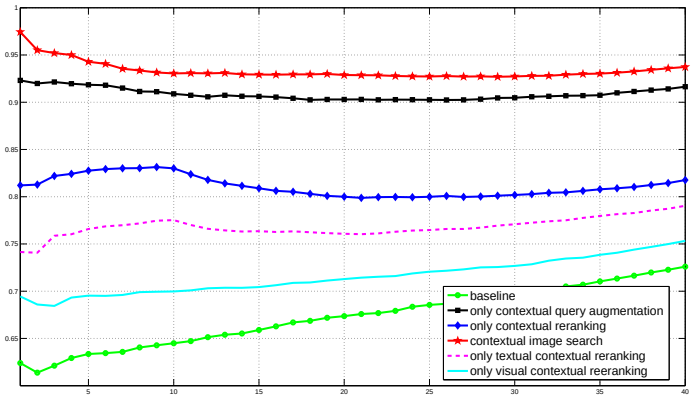


Fig. 5. The quantitative evaluation of contextual image search. The average nDCG curves at different positions of the schemes, only using contextual query augmentation, only using contextual reranking, using both the two schemes, and the baseline algorithm without using the context, are presented. We can observe that context can help improve the performance. Particularly, the nDCG curves of reranking, with textual context and visual context, respectively, are also reported to illustrate the effects of the two contexts.

	A	B	C	D	E
1	56.16	47.94	30.13	18.83	11.30
5	48.82	44.97	30.62	20.87	9.746
10	44.28	40.93	28.70	20.20	8.497
20	37.88	34.04	18.76	12.94	5.823
40	29.11	26.22	12.62	8.873	3.752

TABLE I

RELATIVE IMPROVEMENTS AT POSITIONS $\{1, 5, 10, 20, 40\}$ OVER THE BASELINE SCHEME FOR FIVE SCHEMES: A - CONTEXTUAL IMAGE SEARCH, B - ONLY CONTEXTUAL QUERY AUGMENTATION, C - ONLY CONTEXTUAL RERANKING D - ONLY TEXTUAL CONTEXTUAL RERANKING, AND E - ONLY VISUAL CONTEXTUAL RERANKING. THE UNIT IS %.

only with textual context or visual context, respectively. We can see that both visual and textual contexts can help improve the relevance.

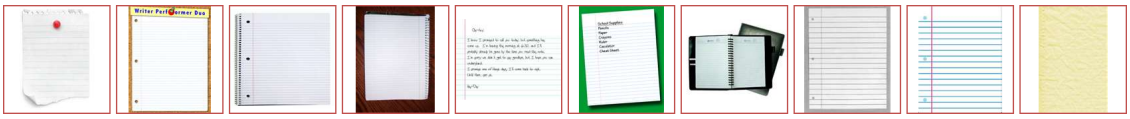
We also report the relative improvements of these schemes over the baseline algorithm at positions, $\{1, 5, 10, 20, 40\}$, which is shown in Tab. I. It can be observed that the proposed contextual image search scheme even gets about 50% improvement over the baseline scheme.

(a) Results with the raw query "Ronaldo".

(b) Results with the augmented query "Ronaldo Brazil". The textual context includes *Brazil* and so on.

(c) Results with the augmented query "Cristiano Ronaldo". The textual context includes *Cristiano*, *Manchester* and so on.

(d) Results with the raw query "notebook".



(e) Results with the augmented query "notebook paper". The textual context includes *paper*, *notepad*, *writing* and so on.

(f) Results with the augmented query "laptop". The textual context includes *laptop*, *computer*, *battery* and so on.

Fig. 6. Visual illustration of contextual query augmentation.

B. Visual Results

This subsection presents visual results to illustrate the contextual image search performance. To show the effects of contextual query augmentation, textual contextual reranking, and visual contextual reranking, we categorize the results and report the visual results.

Contextual query augmentation

We present two visual comparison results shown in Fig. 6. The first example is about famous soccer stars, "Ronaldo". The documents introducing these stars usually only use the name for convenience.

(a) Results with “Cambridge England” without contextual reranking.

(b) Results of “Cambridge England” with textual contextual reranking. The textual context contains *river*, *boat* and *floating*.

(c) Results of “Michael Jordan” without contextual reranking.

(d) Results of “Michael Jordan” with textual contextual reranking. The textual context contains *dunk* and *slam*.

Fig. 7. Visual illustration of textual contextual reranking.

We present two contextual search results, when masking the query *Ronaldo* from two Web pages, http://news.bbc.co.uk/sport2/hi/football/8529228.stm and [3], respectively. In the textual context from the former Web page includes *Brazil*, which suggests the search intent be “Ronaldo Brazil” [4] and the words, *Cristiano* and *Manchester*, in the textual context from the latter Web page, suggests the search intent be “Cristiano Ronaldo”. In another example, users masked the word “notebook” when reading the Web pages, http://en.wikipedia.org/wiki/Notebook, http://www.consumeraffairs.com/news04/2006/08/dell_fire.html, and the queries are augmented as “notebook paper” and “laptop”, respectively, according to the context. The corresponding results are shown in Fig. 6.

Textual contextual reranking

The visual illustration of textual contextual reranking is presented in Fig. 7. Figs. 7(a) and 7(b) show the

[3]http://www.telegraph.co.uk/sport/football/cristianoronaldo/7234785/Cristiano-Ronaldo-Manchester-United-return-possible.html

[4]Actually, his full name is Ronaldo Luś Nazário de Lima, but less used due to the complexity. Therefore, we use “Ronaldo Brazil” as the augmented query.

(a) Irrelevant (b) Irrelevant (c) Relevant

Fig. 8. Image context for "Tower bridge". With the filtering scheme based on the relevance of the local textual context with the query, images in (a) and (b) are filtered out, and the image in (c) is left for visual contextual reranking.

(a) Results of "Tower bridge" without contextual reranking.

(b) Results of "Tower bridge" with visual contextual reranking, which are more consistent with the visual context in Fig. 8.

Fig. 9. Visual illustration of visual contextual reranking.

image search results without contextual reranking and with contextual reranking, when masking a raw query "Cambridge England" from http://www.travelpod.com/travel-photo/flyin_bayman/castles_beer-06/1147366740/s-cambridge-punting.jpg/tpod.html. The textual context of this query from the document contains the keywords *river*, *boat* and *floating*. Therefore, with textual contextual reranking, the images with rivers are promoted as shown in Fig. 7(b). As another example shown in Figs. 7(c) and 7(d), a user may issue a raw text query "Michael Jordan" when reading the Web page, http://www.nba.com/jordan/mjslamdunk.html. The textual context of this query contains keywords like *dunk* and *slam*. As a result, images with these keywords (or expanded words) in its textual contexts, which have large probability to illustrate slam dunks of Michael Jordan, are promoted. The two examples show that the textual context can help find images that are more consistent to user search intention.

Visual contextual reranking

We also show visual examples to illustrate visual context can also help clarify user search intention from the visual contextual reranking. The example in Fig. 3 has shown that visual context is helpful to find images that are more relevant to users. Here, we present another example with the query "Tower bridge" from http://www.the-pass.co.uk/ArticleDetails.asp?ArticleID=123. Its visual context is shown in Fig. 8, and with the filtering scheme based on the relevance of the local text context with the query, only the image in Fig. 8(c) is used for visual contextual reranking. The image search results without contextual reranking and with visual contextual reranking are shown in Figs. 9(a) and 9(b). From the two figures, we can see that the images with tower bridge under the night are ranked on the top after visual contextual reranking, which is more reasonable as the user, reading the document about tower bridge describing the scene under the night, may be more interested in such images.

C. User Study

We conducted user studies to show that the proposed contextual image search is very convenient and helpful for users to perform image search. We recruit 30 volunteers, students from university campus and our research lab, to take part in the user study. Their grades vary from freshman to graduate grade 3. Their ages range from 19 to 24. All participants are Web image search engine users.

We first present a question to them. The question is about the situations where an image search action is triggered from their experiences. The answers show that there are three major situations to trigger users to perform image search: famous sites and people when reading documents, interesting objects heard or seen from other way, and meaningful things demanded in their work. The answers show that the proposed contextual image search scheme will make image search very convenient for users as image search actions often take place when reading documents.

Then, we allow them to use three image search engines: existing commercial image search engine with a text query input box, reduced contextual image search without using contexts, and our contextual image search. After using them about three hours, they give us the feedbacks on using them. The feedbacks show that 1) the latter two schemes for image search make them search image more efficient and convenient and 2) search results of contextual image search are more satisfactory and most of them match their search intention very well although the intention is not indicated in the issued raw query.

V. CONCLUSION

In this paper, we present a contextual image search scheme that uses the context to better understand the search intent and expect better image search results. The context information, including the surrounding text, other main text information and the images, is first extracted from the document where the query is generated. Then we present two key ways to make good use of the context, i.e., remove query ambiguities and promote images that are more consistent with the textual and visual context. The experimental results and user studies justify that the proposed contextual image search scheme is very helpful and effective.

In the future, we will develop more general contextual image search, including mobile image search with wider contexts (e.g., position, time, and history). Moreover, we will extend contextual image search to contextual video search by applying the proposed methodology and investigating extra video contexts.

REFERENCES

[1] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Vips: a vision-based page segmentation algorithm. Technical Report MSR-TR-2003-79, Microsoft, 2003.

[2] J. Cui, F. Wen, and X. Tang. Intentsearch: interactive on-line image search re-ranking. In *ACM Multimedia*, pages 997–998, 2008.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008.

[4] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, pages 1271–1278, 2009.

[5] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, 2002.

[6] J. Fogarty, D. S. Tan, A. Kapoor, and S. A. J. Winder. Cueflik: interactive concept learning in image search. In *CHI*, pages 29–38, 2008.

[7] R. Jain. Multimedia information retrieval: watershed events. In *Multimedia Information Retrieval*, pages 229–236, 2008.

[8] R. Kraft, C.-C. Chang, F. Maghoul, and R. Kumar. Searching with context. In *WWW*, pages 477–486, 2006.

[9] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, 2(1):1–19, 2006.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[11] Y. Luo, W. Liu, J. Liu, and X. Tang. Mqsearch: image search by multi-class query. In *CHI*, pages 49–52, 2008.

[12] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[14] T. Mei, X.-S. Hua, and S. Li. Contextual in-image advertising. In *ACM Multimedia*, pages 439–448, 2008.

[15] Y. Rui and T. S. Huang. A novel relevance feedback technique in image retrieval. In *ACM Multimedia (2)*, pages 67–70, 1999.

[16] J. Sivic and A. Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):591–606, 2009.

[17] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[18] L. Tao, L. Yuan, and J. Sun. Skyfinder: attribute-based sky image search. *ACM Trans. Graph.*, 28(3), 2009.

[19] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1919–1932, 2008.

[20] R. W. White, P. Bailey, and L. Chen. Predicting user interests from contextual information. In *SIGIR*, pages 363–370, 2009.

[21] H. Xu, J. Wang, X.-S. Hua, and S. Li. Interactive image search by 2d semantic map. In *WWW*, 2010.

[22] R. Yan, A. Natsev, and M. Campbell. Multi-query interactive image and video retrieval: theory and practice. In *CIVR*, pages 475–484, 2008.

[23] E. Zavesky and S.-F. Chang. Cuzero: embracing the frontier of interactive visual search for informed users. In *Multimedia Information Retrieval*, pages 237–244, 2008.