The Role of Attractiveness in Web Image Search *

Bo Geng [†‡†], Linjun Yang [‡], Chao Xu [†], Xian-Sheng Hua [‡], Shipeng Li [‡]
[†] Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, P.R.China
[‡] Microsoft Research Asia, Beijing, P.R.China
{gengbo, xuchao}@cis.pku.edu.cn, {linjuny, xshua, spli}@microsoft.com

ABSTRACT

Existing web image search engines are mainly designed to optimize *topical relevance*. However, according to our user study, *attractiveness* is becoming a more and more important factor for web image search engines to satisfy users' search intentions. Important as it can be, web image attractiveness from the search users' perspective has not been sufficiently recognized in both the industry and the academia. In this paper, we present a definition of web image attractiveness with three levels according to the end users' feedback, including *perceptual quality*, *aesthetic sensitivity* and *affective tune*. Corresponding to each level of the definition, various visual features are investigated on their applicability to attractiveness estimation of web images. To further deal with the unreliability of visual features induced by the large variations of web images, we propose a contextual approach to integrate the visual features with contextual cues mined from image EXIF information and the associated web pages. We explore the role of attractiveness by applying it to various stages of a web image search engine, including the online ranking and the interactive reranking, as well as the offline index selection. Experimental results on three large-scale web image search datasets demonstrate that the incorporation of attractiveness can bring more satisfaction to 80% of the users for ranking/reranking search results and 30.5% index coverage improvement for index selection, compared to the conventional relevance based approaches.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms, Experimentation, Human Factors, Performance.

[†]This work was performed when Bo Geng was visiting Microsoft Research Asia as a research intern.
*Area chair: Marcel Worring

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

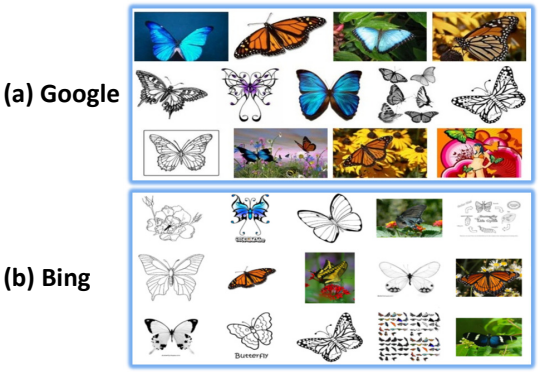


Figure 1: The first page image search results of the query "butterfly" for Google and Bing image search engines, recorded on March 10th, 2011.

Keywords

Web image search, Image attractiveness

1. INTRODUCTION

Existing image search engines are mainly designed to optimize the *topical relevance*, i.e., to what extent the topic of returned images matches that of the text query [13]. However, according to our user study in comparing image search results of Google and Bing, only 8% of the users preferred one to the other due to relevance, which means topical relevance may not be the leading discriminating factor of search engines at the current stage. This may be attributed to the continuous development of information retrieval technologies, with which a carefully designed image search engine can effectively return topically relevant images in most circumstances.

Besides topical relevance, we found that more than 60% of users selected their preferred image search results due to the factors related to the images' attractiveness, e.g. "high quality","sharp", and "beautiful". This indicates that users generally pose a strong demand on the attractiveness of images, even though they may not explicitly state such a need during the search. We call the users' need of getting attractive images the *attractiveness intent*, which is general, implicit in most times, along with the users' image search behavior. It is especially important for those queries for which a large number of relevant images can be returned. In this case, topical relevance is incapable of differentiating these images sufficiently, while attractiveness becomes a more favorable metric to rank the search results so as to satisfy users' attractiveness intent.

Although attractiveness intent is important to the users' satisfaction of a web image search engine, it is insufficiently recognized by both the industry and the academia. As an example shown in Fig. 1, both Google's and Bing's search results contain many unattractive images such as the simple line drawings. While attractiveness has been studied in several literatures, the existing works are more focused on the photographers' perspective [2, 6, 7]. In view of this, it is necessary to investigate what attractiveness implies in the context of web image search from the end users' perspective.

Based on a user study to discover what users expect for attractive images in web image search, we define the attractiveness in three levels, including *perceptual quality*, *aesthetic sensitivity* and *affective tune*. The perceptual quality is the users' basic requirement on the images' attractiveness, which is used to measure, for an image, whether users can perceive clearly the topics delivered by it. Beyond perceptual quality, aesthetic sensitivity further measures whether an image can be appreciated with an aesthetic experience, instead of a simple knowledge consumption. Finally, affective tune measures to what extent certain emotions can be sensed from the images.

To estimate the web image's attractiveness, we integrate visual features originally designed for photo quality assessment, aesthetics prediction, and affective classification. However, the techniques applied to extracting these visual features are still in its infancy and will usually achieve unstable performance. Moreover, the large variations in web images make the problem more challenging.

Fortunately, web images never appear in isolation but are accompanied with rich contextual information, such as EXIF (Exchangeable Image File Format) information and the associated web pages. Normally, EXIF reflects the camera settings and the environment conditions of the photo when it was taken. It can be used to infer the probability of a photo to be attractive. Besides EXIF, web image attractiveness can also be mined from its host web page. While the page descriptions may indicate the intention of using this image and therefore can be used to infer the image's quality, the page structure that reflects the design expertise or the dominance of the image in the page is also informative. Therefore, we propose a contextual approach to web image attractiveness estimation by integrating the content and the context. The experimental results demonstrate its effectiveness on a large-scale web image database.

Once the attractiveness estimation is finished, we further consider applying it into different stages of a web image search engine and extensively testing its effectiveness. A direct application is to incorporate attractiveness into the relevance based image search ranking model, so that images which are not only relevant but also visually attractive can be promoted in search results. However, attractive images are not always necessarily desirable, e.g., users may hope to browse the product designs instead of searching attractive images for the query "iphone". We consequently propose an interactive reranking approach, which allows users to choose whether to rerank by attractiveness or not. Besides the online ranking/reranking, attractiveness can also be incorporated into the offline process. While hundreds of billions of web images are available in the whole domain, we are only capable of indexing about several billions of them, due to the limitation of data storage and computational cost. The incorporation of attractiveness into index selection can help the search engine filter out the low quality, poorly composed or disgusting images.

The contributions of the paper are summarized as follows:

- To the best of our knowledge, we are the first to give a comprehensive definition of the web image attractiveness based on the feedbacks from end users of a web image search engine, and to perform a comprehensive study on the role of attractiveness in web image search.
- By extensively studying the applicability of attractiveness to different stages of a web image search engine, including online ranking and interactive reranking, as well as offline index selection, we conclude that the attractiveness is an important factor in a user satisfactory image search engine.
- We propose a contextual approach to estimate web image attractiveness, which integrates the images' visual content and the contextual information including image EXIF information and the content and structure of its associated web pages.

The rest of the paper is organized as follows. Section 2 presents some related works. We present the overview of attractiveness in web image search in Section 3. To estimate attractiveness, we introduce the adopted visual and contextual features in Section 4 and Section 5 respectively. Section 6 presents the experimental results on attractiveness estimation. Section 7 reports the performance of applying attractiveness to various stages of the web image search. We finally conclude the paper in Section 8.

2. RELATED WORK

The development of automatic photo attractiveness estimation approaches can be categorized into three stages. In the early stage, researchers focused on photo quality assessment by analyzing the artifacts induced by compression, e.g., JPEG compression[14]. Since the compression artifact is not the only factor to influence photo quality, in the second stage, a large number of methods were proposed to analyze images' visual content according to the professional photography rules, such as *simplicity*, *rule of thirds*, and *visual weight*. Various visual features were designed based on these rules, and then integrated by machine learning algorithms to obtain an attractiveness score [2, 6, 7]. By leveraging the recent developments on social network and media sharing websites, Pedro and Siersdorfer [12] utilized the social information such as tags to improve the prediction of social images' attractiveness.

However, the above features are still not robust enough to precisely predict photo attractiveness, due to the limitation of the adopted techniques which are still not mature enough [2, 6, 7, 12]. In addition, most features are designed for as well as evaluated over photos, and may not work well for paintings, drawings, cliparts, etc. which are pervasive on the web. Different from these works, our work studies the components of attractiveness according to search engine end users' feedback, based on which we analyze the applicability of some existing visual features to the web images' attractiveness estimation, and propose to improve the estimation by using contextual information including image EXIF and the associated web pages.

The research on web image search in the multimedia community has been focusing on developing various content based

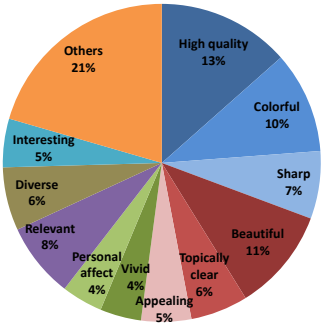


Figure 2: The top ten factors to evaluate the image search results.

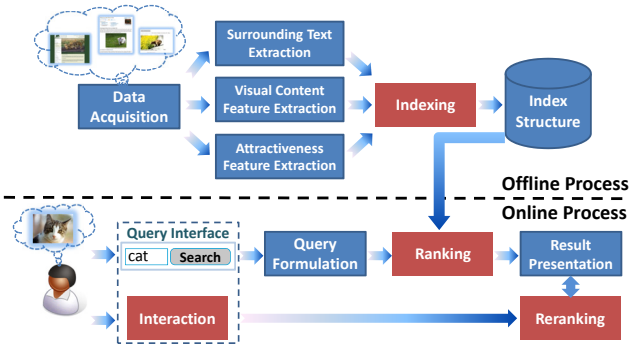


Figure 3: The flowchart of a general web image search engine. The components in which the attractiveness can be applied are highlighted in red.

methods to improve the relevance of image search results, e.g., concept detection [15] and visual reranking [16, 19]. However, visual attractiveness is seldom considered in these works. In [6, 12], the attractiveness models learned from photo collections were applied to ranking web image search results and social images respectively. However, these works only considered attractiveness as a possible metric to rank images, but did not systematically study the role of attractiveness in the web image search context. Zhang et al. [22] proposed an image search engine specific for high-quality photos. However, the engine only aggregated the photos from community based photo forums instead of the images in the whole web. The photo quality is estimated only based on human ratings which is generally unavailable for web images. As a distinguishing framework, our work treats attractiveness as a first class factor besides topical relevance to impact the general web image search performance. We perform extensive experiments to prove that the incorporation of attractiveness into the conventional online ranking and offline index selection problems can significantly improve user satisfaction.

3. OVERVIEW

In this section, we explore attractiveness in the web image search context, by investigating its components through a user study, and discussing its various applications to an image search engine.

3.1 User study on web image attractiveness

We performed a user study by asking 30 users to compare the search results of 12 queries collected from two well-known image search engines, which are denoted as Engine I and Engine II. The queries were sampled from a query log of Engine I, comprising various categories such as animal ("cat"), celebrity ("Michael Jackson"), scene ("beach"), activity ("hockey"), drawing object ("hearts"), and cartoon ("tinkerbell"). For each query, we downloaded the first-page search results of both engines and presented them to users. Users were asked to judge which search result is better and to give further reasons. They were encouraged to freely describe the reasons for their choice, without being given any suggestions or hints. Asking people to make a comparison is a common approach to evaluating search engines, as has been done in [20, 21]. We eliminated the presentation bias due to different page designs, the trust bias of search engines by hiding the search engine names, and the selection bias by randomly swapping the presentation orders of the two search engines for each query.

From the user provided answers, we summarized the most frequent factors related to users' satisfaction and showed them in Fig. 2. We found that "Relevance" only takes up 8% over all the reasons, which is much less frequent than others like "High quality" (13%), "Beautiful" (11%) and "Colorful" (10%). This is because *topical relevance* for most queries are comparably satisfactory for the results from both the search engines. The factors related to attractiveness can be categorized into three levels according to different users' search intention: (1) perceptual quality, whether users can perceive clearly the meaning conveyed by an image, including "High quality", "Colorful" and "Sharp"; (2) aesthetic sensitivity, whether an image can be appreciated with an aesthetic experience, instead of a simple knowledge consumption, including "Beautiful", "Topically clear" and "Appealing"; (3) affective tune, whether certain emotions can be sensed from an image, including "Vivid" and "Personal affect". Since in this paper, we focus on the estimation of attractiveness and its applications in web image search, the other factors including "Diverse" and "Interesting", will be studied in our future work.

3.2 Applications

To discuss the applications of attractiveness, we briefly introduce a general flowchart of a web image search engine, as illustrated in Fig. 3. In the offline phase, an image crawler is applied to acquire web images from the World Wide Web. For the crawled images, we extract various features including surrounding text and visual features to represent them. Based on these features, the images are indexed using techniques such as inverted file, in order to speed up the query response. In the online phase, when a user submits a textual query, the search engine applies its ranking model to rank the images according to their ranking features, and return the corresponding images to the user in order. During browsing the search results, users can perform interactions with the search engine to start a second round of search. From this flowchart, it is natural to apply the attractiveness in three components of a web image search engine, namely the ranking, indexing, and user interaction, as highlighted in Fig. 3.

- **Ranking**. For topically relevant images, attractive images are more favorable to users. Therefore, we can incorporate the attractiveness into the relevance based ranking model, by regarding the attractiveness features or scores as additional ranking features, so that not only topically relevant, but also visually appealing images can be promoted in search results.

- **Interactive reranking**. Attractive images may not

Table 1: Overview of the visual features for web image attractiveness estimation.

Category	Name	#Dim	Description
Perceptual quality	Lighting	2	The mean (brightness) and variance (contrast) of the pixel intensity in gray [12].
	Color	7	The mean and standard deviation of saturation and hue [2, 12]; Contrast of colors[12]; Colorfulness [12]; Naturalness color attributes[4].
	Sharpness & Blur	3	Sharpness: the mean and standard deviation of Laplacian image normalized by local average luminance [12]; Blur: the frequency distribution of FFT transformed image [6].
	Subject Quality	12	The Lighting, Color, Sharpness of the saliency map reweighted image; The blur detection of the subject region.
Aesthetic sensitivity	Rule of Thirds	1	The composition of the subject estimated by the nearest distance of the subject the region to one of stress points [1].
	Simplicity	4	The edge distribution of the original [6] and the saliency map reweighted image; The distinct hue count of the original image [6] and its subject region.
	Visual Weight	1	The clarity contrast between subject region and the whole image [7].
Affective tune	Dynamics	2	The relative number and length of the static vs. dynamic lines [9].
	Color Emotion	17	Histograms designed to express the emotional impact of image color [18].

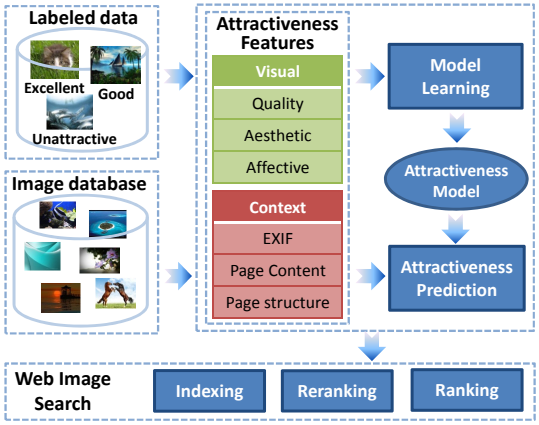


Figure 4: The flowchart of web image attractiveness estimation.

be heavily favored by users for some queries. For example, for the query "iphone", users may desire more the images about its visual designs instead of the attractive images containing an iphone. Therefore, it is more user-friendly to provide an option for users to rerank by attractiveness. With this explicit option, users can decide whether or not to rerank the returned images according to the attractiveness score.

- **Index selection**. There are tens of billions of images in the whole web, but we can only index about a small portion of them due to the storage and computation limitation. The existing approaches for index selection are for web pages, which selects the web pages that users will be interested to search for, based on a learned static rank [10]. We adoped the same framework for image index selection. Furthermore, beyond those features that are designed for web pages in [10], we apply the attractiveness as a new feature indicating whether the images are attractive enough to be searched for, and incorporate it into the learning framework. The incorporation of attractiveness to index selection can help to filter out the low-quality, poorly composed, and disgusting images from an index.

A web image attractiveness estimation system is the core of the above three application. As shown in Fig. 4, the system is based on a learning framework, to combine many carefully designed features. The features include the existing visual features designed mainly for estimating photo attractiveness, as well as the contextual features extracted from image EXIF and associated web pages. We adopted a linear model in this paper to accommodate the large-scale requirement of a practical web image search engine, since a linear model is more efficient to be learned and applied to billions of images. In the following we will mainly introduce the various features.

4. VISUAL FEATURES

In this section, we describe a selected subset of visual features originally designed for photo attractiveness estimation and study their effectiveness when applied to web images.

4.1 Feature Description

A summary of our adopted visual features are presented in Table 1. The visual features are categorized into three levels accordingly to the attractiveness definition: perceptual quality, aesthetic sensitivity, and affective tune.

The perceptual quality features describe the basic qualities of web images, including the brightness, contrast, colorfulness, and sharpness & blur [2, 6, 12]. Some of them are illustrated in Fig. 5 (a)(b)(c). Besides the original implementation which are performed on the whole image, we propose to additionally extract these features on subjects. In this paper, we didn't explicitly detect the subject, but use the saliency map [8] to compute a probability of each pixel to be on the subject. Then, we extract features over the whole image with reweighting by the probability value, or over the subject region detected by the minimal bounding box that contains 90% mass of all probabilities.

The aesthetic sensitivity describes whether an image is visually appealing from the aesthetics perspective. Existing aesthetics visual features are mostly designed based on well-known composition rules, including *Rule of Thirds* which is positioning the subject near the *golden ratio* points, *simplicity*, which is simplifying the subject topic, and *visual weight*, which is balancing the visual weight of the subject and the background. We apply the composition feature [1], the hue count and edge distribution features [6], and the clarity contrast feature [7] to estimate the conformance to each rule respectively. Illustrations of the rules are in Fig. 5(d)(e)(f) respectively.

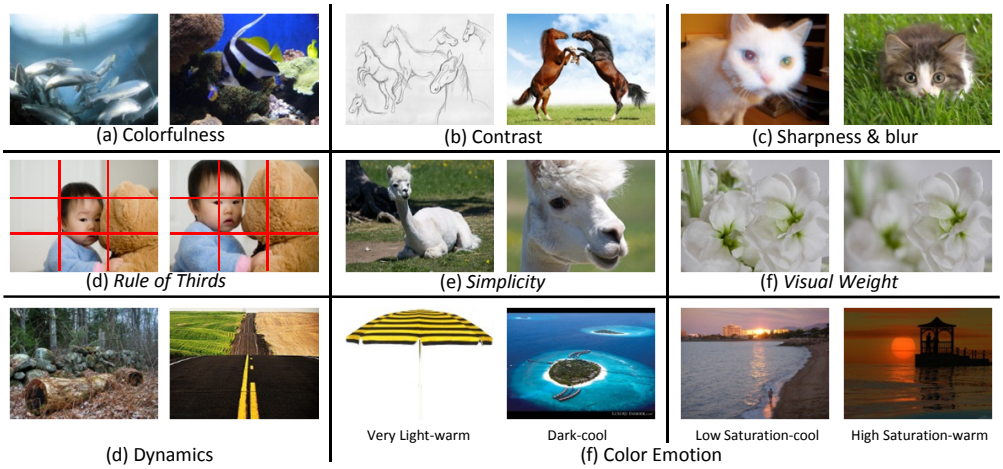


Figure 5: Example images to illustrate the visual features. Left (right) image in each group is unattractive (attractive).

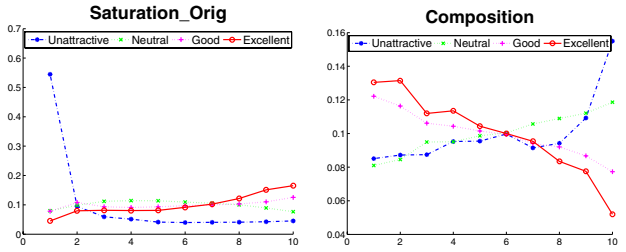
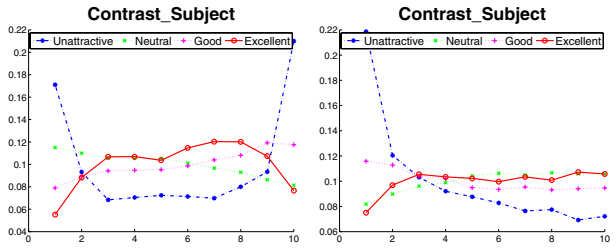


Figure 6: The image distribution of each attractiveness degree over different visual features.



(a) Before Transformation (b) After Transformation

Figure 7: The distribution of images before and after the feature transformation.

The affective tune features model the emotional impact of an image's visual elements to humans. We applied two categories of features to estimate the affective tune, based on the lines and colors respectively. The line-based features are calculated according to the distributions of static and dynamic lines which reflect different emotions [9], as illustrated in Fig. 5(g). The color-based features are represented as a histogram which quantizes the impact of color to emotions, as illustrated in Fig. 5(h)(i).

4.2 Feature Analysis

We analyzed the effectiveness of all above 49-dimensional visual features over the 100Q dataset comprising 95,193 images, which are manually labeled as "Excellent", "Good", "Neutral", or "Unattractive" according to the perceived attractiveness. More details can be found in Section 6.1.

To study the effectiveness of a certain feature, we illustrate the feature distribution, approximated by a 10-bin histogram, for each of the four attractiveness degrees, as shown in Fig. 6. We can observe that for some features, such as "Saturation_Orig"[1] and "Composition", the feature is linearly correlated with the attractiveness degree. For example, an image with a smaller "Composition" value corresponds to an attractive image with a higher probability. However, the correlation for others may be nonlinear. For example, attractive images more probably have a moderate contrast value, as shown in Fig. 7(a). Furthermore, we observe that for these features, attractive (as well as unattractive) images are usually governed by a unimodal distribution. In order to obtain linear features that may perform well for a linear model, we propose to transform these features by the following equation.

$$g(f_i) = \begin{cases} \exp(-\sigma_l(r_i - f_i)) & \text{if } f_i \le r_i \\ \exp(-\sigma_r(f_i - r_i)) & \text{otherwise} \end{cases} \quad (1)$$

where the parameter r_i is a threshold, σ_l (σ_r) is a parameter to scale the features smaller (larger) than r_i. The transformed feature values of "Contrast_Subject" are shown in Fig. 7(b) as an example, which indicates a linear correlation of the feature value to the attractiveness degree can be obtained. In our implementation, 18 of the 49 features are transformed according to Eqn. (1), including "clarity", "dynamics"; subject "sharpness"; the original and subject-reweighted "lighting", "color contrast", the standard deviation of"sharpness"; the original and subject region's "edge distribution", "blur", and "hue count".

However, the above visual features are still not sufficient for predicting web image attractiveness, due to three-fold reasons. First, the visual features are still not robust enough to predict the rules precisely, since the underlying techniques such as blur detection and saliency extraction are still open research problems that are not well solved. Second, photographers often utilize professional techniques, such as adding special glass filters and post editing, to violate the rules in order to generate special yet attractive effects. Fig. 8(a) shows an attractive image created with the polaroid "old day atmosphere" which is predicted to be unattractive using our visual features. Third, existing visual features are mostly designed for photos, while the source of web images are quite diverse, additionally including paintings, drawings, cliparts, graphics, etc. The attractiveness of these non-photo images is weakly correlated with the photography rules. For

[1]Here, "Orig" denotes the feature extracted over the original whole image, while "Subject" is over the saliency reweighted image or the subject region.

Table 2: The text words with top 10 information gain values.

Words	wallpaper	wallpapers	desktop	gif	coloring	jpg	download	backgrounds	background	printable
IG	0.0213	0.0203	0.0179	0.0114	0.0091	0.0084	0.0069	0.0056	0.0053	0.0050

(a) Photo with Polaroid effect (b) Clipart (c) Painting

Figure 8: Example images to illustrate the limitation of visual features.

example, Fig. 8(b) is an ordinary clipart image predicted as very attractive, while Fig. 8(c) is a beautiful wallpaper painting predicted as unattractive, by using the visual features. Fortunately, a web image never exists in isolation but is accompanied with contextual information other than visual content, which can be utilized to improve the attractiveness estimation.

5. CONTEXTUAL FEATURES

Our contextual features are extracted from three sources, including the image EXIF, as well as the content and structure of the associated web page. We will analyze these information sources and design corresponding features as follows.

5.1 EXIF

EXIF contains rich information on camera settings, which can reflect the photographer's expertise and environment conditions when the photo was taken. Professional photographers often utilize advanced equipments to take photos, such as telephoto and large aperture lens. By manually setting the modes special to their equipments, such as large focal length, they can produce many attractive effects such as shallow depth of field effect (sharpening the foreground and blurring the background). In addition, a nice environment with enough lighting can increase the possibility of producing attractive photos, which often corresponds to a short exposure time and a low ISO. Therefore, we choose the following EXIF metadata to design attractiveness features, including Exposure Program, Focal Length, ISOspeed, Exposure Time, and F-Number. We verified each of these feature by analyzing the feature distributions. Fig. 9 shows two examples, from which we can observe that attractive images are usually taken with "Manual" Exposure Program (value=1) mode instead of "Normal Program" (value=2), as well as telephoto lens with large focal length settings.

It is natural to adopt the values of different EXIF information directly as features. However, in the 100Q dataset, only 11.2% images are accompanied with EXIF. Thus, we need to tackle those images without EXIF. This is achieved by quantizing each feature into four discrete bins, and putting the images without EXIF into a separate bin. Because Exposure Program is categorical information with 10 options, we directly transform it into a binary 10-dimensional vector with proper consideration of missing values. Finally, 30-dimensional EXIF features are extracted for each image.

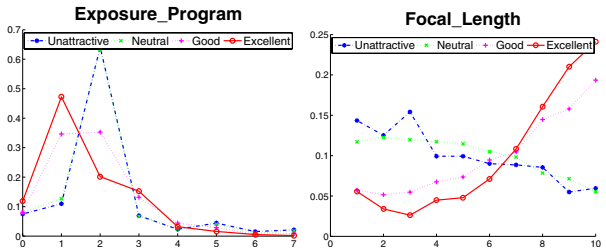


Figure 9: The image distribution of each attractiveness degree over different EXIF features.

5.2 Page Content

Every web image is embedded in at least one web page. The web page authors usually leave rich information in the page content to describe an image, which can be utilized to infer the image's attractiveness. To discover useful information from the pages, we apply information gain (IG) [17] to estimate the correlation between the image attractiveness and the text surrounding it. We select the 10 words with top IG in our 100Q dataset, and show the results in Tabel. 2. It can be observed that the discriminative text words for attractiveness can be categorized into two groups, i.e., image intention ("wallpaper", "desktop", "background" and "download") and image quality ("printable", "coloring", "jpg", and "gif"[2]).

Based on the above analysis, we define a binary feature vector to represent the presence or absence of the eight[3] selected words in seven text sources of a web page, including the text associated with the image such as anchor text, name, surrounding text, and url, as well as the text associated with the page such as title, meta description, and meta keywords. The texts associated with the image and with the page are also aggregated respectively to generate two additional sources.We consider the nine text sources separately since different sources have different confidences to predict attractiveness. Finally, 72-dimensional binary page content features are extracted for each image.

5.3 Page Structure

The page structure can be used to infer the authors' expertise and the the importance of the image w.r.t. the page. Fig. 10 shows three examples, with (a) an attractive image on a professionally designed web page; (b) an amateur image in a personal blog; and (c) an icon in a corner of the page for decoration. We can observe that attractive images are more probably embedded in a professionally/carefully designed web page, e.g., with an appropriate image display size, a structurally long file name to facilitate its identification, and an appropriate length of surrounding text to describe the image content. In addition, attractive images are usually positioned near the center of a page to capture users' attention, while unattractive ones are in the peripheral areas.

[2]Here, "gif" is negatively correlated, while the others are positively correlated.

[3]"wallpaper" and "wallpapers", as well as "background" and "backgrounds", are merged as one word in our implementation.

Figure 10: Example images to illustrate the page features related to image attractiveness. The target image is highlighted with a red bounding box.

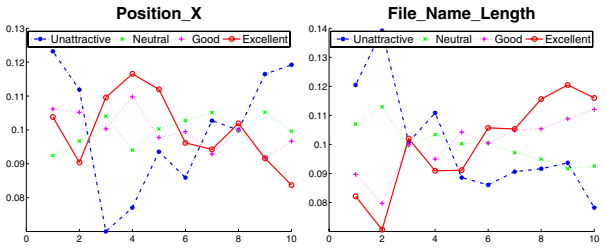


Figure 11: The image distribution of each attractiveness degree for different page structure features.

Based on these properties, we design five features related to page structure, including the image's display size in the page, the length of the image file name, the number of words surrounding the image, and the position of the image in horizontal and vertical dimensions respectively. Fig. 11 presents the distribution of two example features, which illustrates the effectiveness of the proposed features. We utilize the feature transformation in Eqn. (1) to transform the features (except "File_Name_Length"), resulting into a 5-dimensional page structure feature.

6. EXPERIMENTS ON ATTRACTIVENESS ESTIMATION

In this section, we report several experiments on a large-scale web image dataset 100Q to evaluate the effectiveness of the proposed attractiveness estimation approach.

6.1 Dataset and Experiment Setting

The 100Q dataset is collected from a commercial web image search engine (Engine I). We sampled 100 queries according to the query log and crawled the returned images by using them to query Engine I. The selected queries cover many topics that users commonly searched, including celebrity ("michael jackson"), animal ("cat"), cartoon ("tinkerbell"), product ("iphone"), event ("hockey"), and scene ("beach"). For each query, no more than the top 1000 returned images were downloaded, resulting into totally 95,193 images in our dataset. Three labelers were employed to label the attractiveness of each image into four degrees, i.e., "Excellent", "Good", "Neutral", and "Unattractive". The ground truth is determined as the median of the results from the three labelers.

The dataset was evenly split into ten folds, each of which contains the images for ten queries. We chose one fold to tune the parameters of feature transformation in Eqn. 1. Among the remaining nine folds, we selected seven to train the model, one to validate the model parameters, and the rest to evaluate the performance. We switched the fold configuration in a round-robin fashion so that each fold is ensured to be taken as test and validation once. The results

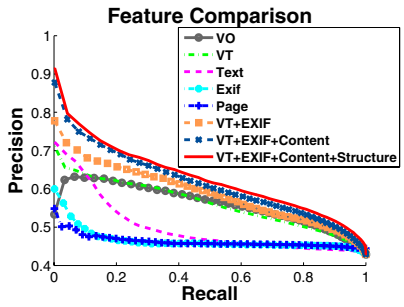


Figure 12: Precision-recall curve of attractiveness prediction using different features.

reported in the paper are the average of the nine results under a specific evaluation metric. RankSVM [5] was applied to train the attractiveness model.

6.2 Results

To evaluate the attractiveness estimation, we present the precision-recall curve in Fig. 12. For computing the precision and recall, we treat "Excellent" and "Good" as positive and the others as negative. It can be observed that due to the limitation of visual features and the diversity of web images, the performance of the original visual features (VO) is quite low, with only 0.62 precision at 0.2 recall. The transformation (VT) can improve the performance a little bit, especially in reducing the false alarm by VO. Among the three categories of contextual features, page content (Content) is comparatively more robust than EXIF and page structure (Structure), even better than VT at low recall. However, Content performs worse at high recall because the feature vector is very sparse and many web pages do not contain the selected eight words. EXIF and Structure are only effective for a limited number of images at low recall, because many images miss EXIF while Structure only is weakly correlated to attractiveness. Even though each independent contextual feature is unable to achieve a high precision, they can complement the visual features. By gradually combining VT with each contextual feature, the performance steadily increases, and finally achieves precision 0.7 at recall 0.2, with totally 12.9% improvements over VO.

7. APPLICATIONS

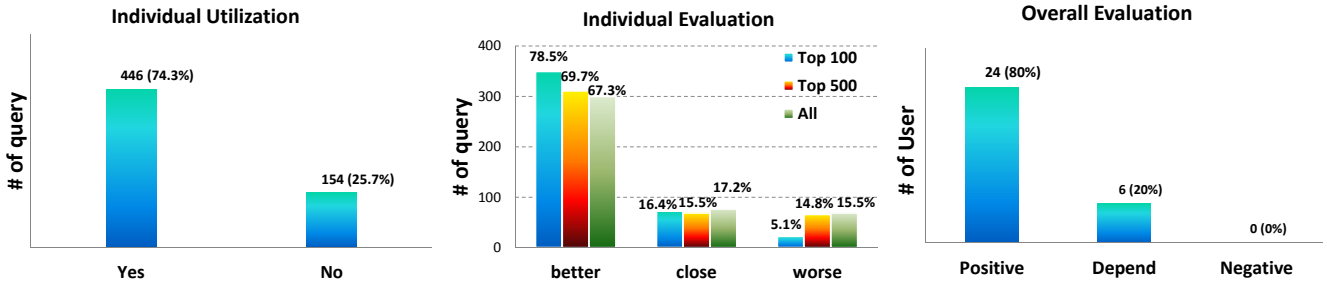
We applied the proposed attractiveness estimation to three components of web image search, including interactive reranking, ranking and index selection respectively.

7.1 Interactive Reranking

We firstly quantitatively evaluated the performance of reranking by attractiveness over the 100Q dataset. Here, all the returned images are reranked for each query according to their attractiveness scores predicted in Section 6.2. Since images in the first result page (usually containing around 20 images) is more critical to a user's satisfaction to a search result, we evaluated the performance in terms of various measures on the top 20 ranked images, including Precision (Precision@20), Mean Average Precision (MAP@20), and Normalized Discounted Cumulative Gain (NDCG@20). Since the occurrence of an unattractive image in the top ranked results may greatly degrade the user's experience, we propose a new measure called Unattractive Rejection (UR), which is

Table 3: The performance of attractiveness reranking for different features.

	Engine I	VO	VT	EXIF	Content	Structure	VT+EXIF	VT+EXIF +Content	VT+EXIF +Content+Structure
Precision@20	0.559	0.654	0.692	0.563	0.692	0.520	0.721	0.771	**0.782**
MAP@20	0.408	0.485	0.545	0.395	0.544	0.351	0.575	0.652	**0.662**
NDCG@20	0.429	0.481	0.510	0.423	0.503	0.415	0.520	0.559	**0.568**
UR	0.196	0.103	0.072	0.144	0.060	0.119	0.076	0.050	**0.044**



(a) (b) (c)

Figure 13: User study on reranking by attractiveness for web image search.

defined as

$$\text{UR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

where $|Q|$ denotes the number of queries in test set Q, and rank_i is the rank of the first "unattractive" image in the search results of query i.

The performance comparisons are shown in Table. 3. Similar to the results for attractiveness prediction, utilizing both the transformed visual features and three kinds of contextual features obtains the best performance for all the metrics. In terms of Precision@20, we can conclude that 78% of the top 20 images are above "Neutral", compared to 56% for the search result of Engine I. The 0.044 UR of our final result denotes that the top ranked "Unattractive" image is on average positioned below the 20th, which is much more desirable than the result of Engine I, for which the first unattractive images appear at the 5th position. This result implies that our reranking method can on average eliminate unattractive images in the first page well.

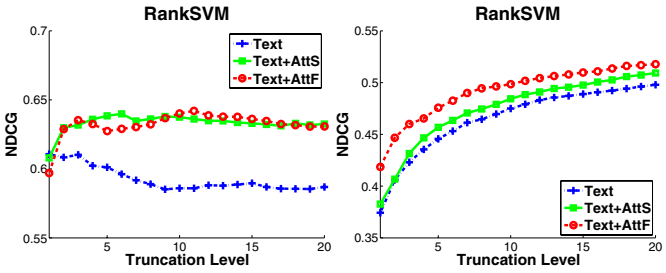
Besides the quantitative results, we also performed a user study of interactive reranking to evaluate whether it is useful for average users. 30 users composed of 13 males and 17 females, with ages ranged from 18 to 29, were invited to perform the study. Among them, 25 users used image search engines frequently, while 5 users seldomly did. We randomly selected 20 queries for each user, and ensured that each query was performed by exactly 6 different users.

For a specific query, a user was firstly presented with the search result from Engine I, and asked whether or not to apply reranking by attractiveness. From the results shown in Fig. 13(a), we found that for 446 out of 600 times users chose to utilize reranking by attractiveness, taking up 74.3% of all. In addition, we observed that for queries such as animal (e.g., cat, dog, butterfly), scene (e.g., beach, sunset, paris), and other queries with attractiveness requirements on images (e.g., wedding dress, rainbow, love), more than five (out of six) users applied reranking. While for queries such as products (e.g., iphone, nike), objects (e.g., shoes, guns, world map), no more than two users chose to rerank. In the latter case, the relevance based results already provide sufficient information related to the queries or attractive counterparts are not their search targets.

For the situations that users chose to apply reranking by attractiveness, we provided them three results in a random order, namely reranking the top 100, top 500 and all images. For each result, users were asked to compare whether the reranked result is better than, close to, or worse than the original result from Engine I, and leave free-style reasons for each comparison. The results are shown in Fig. 13(b). We can observe that for reranking the top 100 search results, it can satisfy the user's search intention for at most 78.5% of the cases, with a degradation for 5.1%. However, as the number of images to be reranked increases, the performance of reranking decreases a little. This is caused by the degradation of topical relevance, since the involvement of more images for reranking may promote the attractive but topically irrelevant images from the bottom of original search results.

We further analyzed the user provided reasons, and found that "Colorfulness", "Visual appeal", "Content richness", "Relevance", and "High quality" were the top five reasons for preferring reranking results, taking up 11.6%, 9.3%, 7.9%, 5.7% and 5.7% of all the factors respectively. Based on this feedback, we can derive that reranking is able to significantly improve the users satisfaction by returning more attractive images. It's also interesting to find that users sometimes thought the reranking result was better because it is more "Relevant", which mostly occurred for queries that favor attractiveness, e.g., butterfly, beach, and wedding dress. Here, the relevance in users' mind, so-called *user relevance*, stands for how "good" a retrieved result is with regard to the user's information need. It is broader than the *topical relevance* by encompassing many other concerns of the user such as attractiveness, interestingness and novelty [13]. On the other hand, the top reasons for reranking being worse is "Irrelevance", "Unwanted color" and "Unusual effects", where "Irrelevance" takes up 46.2% of all reasons. Therefore, a better strategy should be to balance the attractiveness with topical relevance, and treat user relevance as the objective to design web image search engines.

Finally, after all the above studies, users were asked whether or not they would prefer to being provided with the option



(a) 100Q (b) 594Q

Figure 14: The ranking performance of different feature configurations.

of attractiveness based interactive reranking for a general web image search. Fig. 13(c) shows that 80% of them supported adding such an option, while 20% thought it would be better to guarantee the relevance, as well as the accuracy of attractiveness estimation before the adoption, and none of them objected the proposal.

7.2 Ranking

We evaluate the ranking with attractiveness through three feature configurations, by comparing the performance of the ranking model trained over text features (Text), text plus the attractiveness score (Text+AttS), and text plus all the proposed attractiveness features (Text+AttF). 154-dimensional textual ranking features were extracted for each query-image pair from the text associated with the images, according to a common text search approach[4]. RankSVM [5] was applied here to learn the ranking model. Note that due to the data split issue, the scores of images from different queries may be predicted by different models while images from the same query are ensured by the same one. This doesn't influence the ranking model learning here, because RankSVM only cares about the pairwise feature differences of images returned for the same query.

We labeled the relevance of each image w.r.t. its query into four degrees including "Excellent", "Good", "Fair" and "Irrelevant". The labeling is not only based on relevance, but also on attractiveness of images. Specifically, images that are topically irrelevant to their associated queries are labeled as "Irrelevant". The remaining ones are labeled as one of "Excellent", "Good", and "Fair" based on whether they are attractive, natural or unattractive respectively. The data split and evaluation setting is the same as in Section 6.1.

From the results in Fig. 14(a), it can be observed that by combining Text with either attractiveness score or features, the performance is significantly improved, by 7.8% and 7.5% w.r.t. Text respectively. This proves that the incorporation of attractiveness features can help the ranking models differentiate topically relevant images with different attractiveness degrees and return more user desired images. We found that Text+AttS and Text+AttF obtained comparable performance. Furthermore, since the 100Q dataset is crawled from the top search results of Engine I, the ranking learning and evaluation may be biased to the search engine's ranking model. To eliminate the bias, we use another large-scale dataset 594Q which contains not only top ranked images but also middle and tail ranked images for each query. It is composed of 594 queries and 137,348 query-image pairs

[4]http://research.microsoft.com/en-us/projects/mslr/feature.aspx

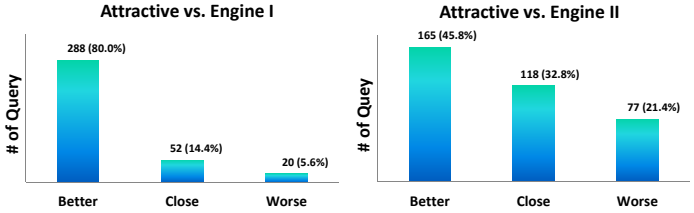


Figure 15: The user study on the ranking performance comparisons.

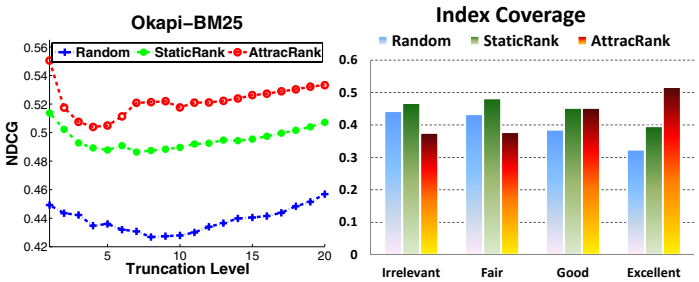
and designed to evaluate the relevance based performance of ranking models [3]. The dataset was split into ten folds and we applied the round-robin setting same as that in Section 6.1 to evaluate the performance. The results are shown in Fig. 14(b). We can derive that Text+AttF and Text+AttS improve 4% and 2.4% over Text for NDCG@20 respectively. Although the improvement is less than that for 100Q, it is statistically significant and the p-values of the t-test of the two approaches over Text in terms of NDCG@20 are 4.24e-9 (Text+AttF) and 3.60e-006 (Text+AttS) respectively.

We also performed a user study to test the whole-page user satisfaction on the search results. The study was performed for the 12 queries over the 100Q dataset under the same setting as in Section 1. Here, the previously employed 30 users were again invited to compare the first page search results of Text+AttF and that of Engine I and Engine II respectively. From the results shown in 15, we can observe that by incorporating the attractiveness into the ranking model, the performance is significantly better than Engine I with 80% voting for "better" while only 5.6% for "worse", and much better than Engine II with 45.5% for "better" while 21.4% for "worse". Actually, as the 100Q dataset was crawled from Engine I, which was shown to perform worse than Engine II for 83% of the cases according to the user study on the same 12 queries, the fact that the incorporation of attractiveness made Engine I outperformed Engine II's results further demonstrates the effectiveness of the proposed method in satisfying end user's search intentions.

7.3 Index selection

To demonstrate the effectiveness of incorporating attractiveness into image index selection, we collected one million images from the whole web, and targeted to select 100,000 among them for indexing. Two baseline approaches including random selection (Random) and static rank based selection (StaticRank) are used to compare with the proposed static rank plus attractiveness based selection (AttracRank). We implemented a modified version of static rank features according to [10], including site-level PageRank, in degree and out degree of each page, the click count of each image, the length and word count of image anchor text, the length of page and image urls, which accounts for totally 8-dimensional features. We utilized RankSVM to train the index selection models for StaticRank and AttracRank approaches over the 100Q dataset respectively, with the relevance label as the ground truth and the model parameters determined by 10-fold cross validation. The selection models were used to predict the scores of one million images, and the top scored 100,000 images were collected for each method.

We evaluated the index selection methods in terms of the ranking performance over the selected subset based on a carefully selected retrieval model. To reduce the uncertainty brought by the setting of different retrieval model parameters, we applied the Okapi-BM25 implemented in



(a) Ranking (b) Index Coverage

Figure 16: The performance of three index selection methods in terms of ranking and index coverage.

Lemur[5] with its parameters set to the default value, which was proven as a stable ad-hoc retrieval model [11]. For a specific query, the retrieval model was applied to each index, and all the top 40 ranked images were collected for human labeling. Each query-image pair is labeled into four-degree relevance as for 100Q. We tested the ranking performance of 100 queries over each index, and showed the average NDCG in Fig. 16(a). It can be observed that the ranking performance of Random is rather low, while it is significantly improved for StaticRank by using the features from pages. With the incorporation of attractiveness, the NDCG@20 on AttracRank is improved by 5.14% compared to StaticRank.

Besides the ranking performance, we also evaluated the index coverage of each index, by computing the coverage of all labeled data for each relevance degree. Here, to convert the data from query-dependent to query-independent, the maximum relevance degree is taken as the ground truth of an image if it is returned for more than one query, because the image relevant to at least one query is potentially of interest to search engine users. The results in Fig. 16(b) demonstrate that AttracRank obtained the most "Excellent" images while and the least "Irrelevant" ones. The improvement of AttracRank over StaticRank is 30.5% for indexing "Excellent" and 19.6% for reducing "Irrelevant".

8. CONCLUSIONS

We have witnessed the importance of web image attractiveness to impact users' search experience, especially when the topical relevance has already been effectively satisfied. However, this has not been sufficiently recognized by both the industry and the academia. Therefore, as the first work studying the web image attractiveness from the search engine end users' perspective, we define it according to the users' feedback from three levels, including perceptual quality, aesthetic sensitivity and affective tune. We utilized the existing approaches to extract visual features according to the three levels and investigate their applicability to web image search. To accommodate the unreliability of visual features induced by large variations of web images, we propose a contextual approach to integrate the contextual cues mined from the image EXIF and its associated web pages to complement the visual features, and obtain 12.9% precision improvement on attractiveness estimation. With such a success, we further investigate the role of attractiveness by applying it to various stages of web image search, including online ranking and interactive reranking, as well as offline index selection. The results over three large scale datasets demonstrate that, compared with the conventional relevance based approaches, the incorporation of attractiveness can improve the user experience of online ranking significantly, with nearly 80% of users' preference. Furthermore, it can achieve over 30.5% index coverage improvement for the offline image index selection.

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

9. ACKNOWLEDGMENTS

This work is partially supported by NBRPC 2011CB302400, NSFC 60975014 and NSFB 4102024.

10. REFERENCES

[1] S. Bhattacahrya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *SIGMM*, 2010.
[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.
[3] B. Geng, L. Yang, C. Xu, and X.-S. Hua. Content-aware ranking for visual search. In *CVPR*, 2010.
[4] K.-Q. Huang, Q. Wang, and Z.-Y. Wu. Natural color image enhancement and evaluation algorithm based on human visual system. *CVIU*, 2006.
[5] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.
[6] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.
[7] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008.
[8] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE TMM*, 2005.
[9] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *SIGMM*, 2010.
[10] M. Richardson, A. Prakash, and E. Brill. Beyond pagerank: machine learning for static ranking. In *WWW*, 2006.
[11] S. Robertson and D. A. Hull. The trec-9 filtering track final report. In *TREC*, 2000.
[12] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *WWW*, 2009.
[13] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. In *JASIS*, 1976.
[14] H. Sheikh, A. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: Jpeg2000. *IEEE TIP*, 2005.
[15] C. G. M. Snoek and A. W. M. Smeulders. Visual-concept search solved? *IEEE Computer*, 2010.
[16] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *SIGMM*, 2008.
[17] G. Wang and F. H. Lochovsky. Feature selection with conditional mutual information maximin in text categorization. In *CIKM*, 2004.
[18] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *ICSMC*, 2006.
[19] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *SIGMM*, 2010.
[20] Y. Yue, R. Patel, and H. Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In *WWW*, 2010.
[21] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *SIGMM*, 2009.
[22] L. Zhang, L. Chen, F. Jing, K. Deng, and W.-Y. Ma. Enjoyphoto: a vertical image search engine for enjoying high-quality photos. In *SIGMM*, 2006.