Contextual In-Image Advertising

Tao Mei, Xian-Sheng Hua, Shipeng Li
Microsoft Research Asia
5F Sigma Center, 49 Zhichun Road, Beijing 100190, P. R. China
{tmei, xshua, spli}@microsoft.com

ABSTRACT

The community-contributed media contents over the Internet have become one of the primary sources for online advertising. However, conventional ad-networks such as Google AdSense treat image and video advertising as general text advertising without considering the inherent characteristics of visual contents. In this work, we propose an innovative contextual advertising system driven by images, which automatically associates relevant ads with an image rather than the entire text in a Web page and seamlessly inserts the ads in the nonintrusive areas within each individual image. The proposed system, called *ImageSense*, represents the first attempt towards contextual in-image advertising. The relevant ads are selected based on not only textual relevance but also visual similarity so that the ads yield contextual relevance to both the text in the Web page and the image content. The ad insertion positions are detected based on image saliency to minimize intrusiveness to the user. We evaluate ImageSense on three photo-sharing sites with around one million images and 100 Web pages collected from several major sites, and demonstrate the effectiveness of ImageSense.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Selection process*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Web page segmentation, image saliency, image advertising.

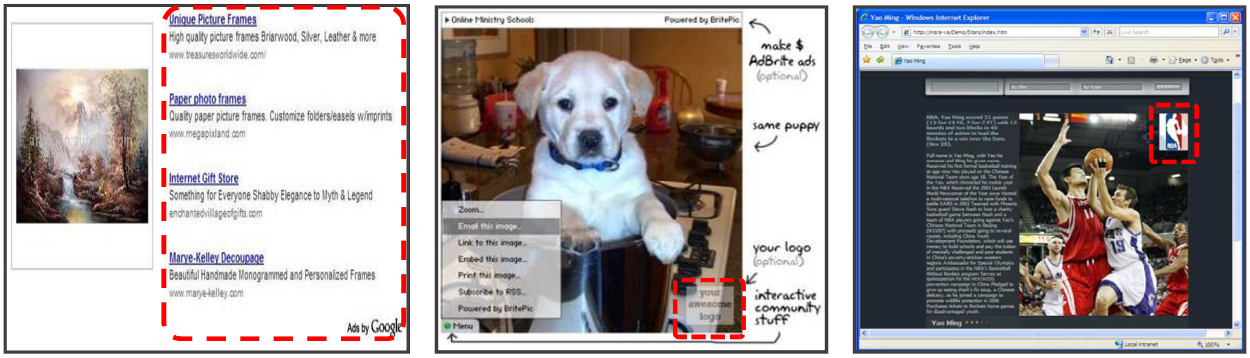
1. INTRODUCTION

The proliferation of digital capture devices and the explosive growth of online social media (especially along with the so called Web 2.0 wave) have led to the countless private image collections on local computing devices such as personal computers, cell phones, and personal digital assistants (PDAs), as well as the huge yet increasing public image collections on the Internet. Compared with text and video, image has some unique advantages: it is more attractive than plain text, and it has been found to be more salient than text [9], thus it can grab users' attention instantly; it carries more information that can be comprehended more quickly, just like an old saying, a picture is worth thousands of words; and it can be shown to the users faster than video. As a result, image has become one of the most pervasive media formats on the Internet [13]. On the other hand, we have witnessed a fast and consistently growing online advertising market in recent years. Motivated by the huge business opportunities in the online advertising market, people are now actively investigating new Internet advertising models. To take advantages of the image form of information representation, *image-driven contextual advertising*, which associates advertisements with an image or a Web page containing images, has become an emerging online monetization strategy.

Many existing ad-networks such as Google AdSense [1], Yahoo! [32], and BritePic [4], have provided contextual advertising services around images. However, conventional image advertising primarily uses text content rather than image content to match relevant ads. In other words, image advertising is usually treated as general text advertising without considering the potential advantages could be brought by images. There is no existing system to automatically monetize the opportunities brought by individual image. As a result, the ads are only generally relevant to the entire Web page rather than specific to images it contained. Moreover, the ads are embedded at a predefined position in a Web page adjacent to the image, which normally destroys the visually appealing appearance and the structure of original web page. It could not grab and monetize users' attention aroused by these compelling images. Figure 1 (a) and (b) show some exemplary image ad-networks, where Google AdSense [1] displays the ads side by side with the image, while BritePic [4] embeds the ads at a corner of the image. It is worth noting that although BritePic [4] also supports embedding the ads within an image, the ad positions are most likely predefined without considering whether these positions would be intrusive to the users.

We believe that it is not appropriate to treat image advertising as general text advertising. One point we want to make here is that you can view online image contents as an effective *information carrier* for commercials. The more

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-303-7/08/10 ...$5.00.



Unique Picture Frames
High quality picture frames Briarwood, Silver, Leather & more
www.treasuresworldwide.com/

Paper photo frames
Quality paper picture frames. Customize folders/easels w/imprints
www.megapixland.com

Internet Gift Store
Something for Everyone Shabby Elegance to Myth & Legend
enchantedvillageofgifts.com

Marye-Kelley Decoupage
Beautiful Handmade Monogrammed and Personalized Frames
www.marye-kelley.com

(a) AdSense [1]

(b) BritePic [4]

(c) ImageSense

Figure 1: The exemplary image advertising schemes. The highlighted rectangle areas indicate the associated ads. The existing image ad-networks like (a) and (b) match ads with Web pages only based on textual relevance and embed ads at a preserved position, while (c) associates ads not only with Web page but also with the image content and seamlessly embeds the ads at a non-intrusive position within the image.

compelling the image contents, the more audience will view them, then the more revenue will be generated. The following distinctions between image and text advertising motivate a new advertising model dedicated to images.

- We believe that beyond the traditional media of Web pages and videos, *images can be more powerful and effective carriers of online advertising.*

- We believe that the ads should be *dynamically embedded at the appropriate positions within each individual image* (i.e., *in-image*) rather than at a predefined position in the Web page. Based on this claim, it is reasonable to assume the ads should be mostly image logos, rather than just textual descriptions, which are more suited to be seamlessly embedded into images.

- We believe that the ads should be *locally relevant to image content and its surrounding text*, rather than globally relevant to the entire Web page. Therefore, it is reasonable to assume that the image content and its surrounding text should have much more contributions to the relevance matching than the whole Web page.

Motivated by the above observations, we propose in this paper an innovative in-image advertising model to automatically deliver contextually relevant ads by an online advertising system. The proposed system, named *ImageSense*, supports contextual image advertising by associating the most relevant ads to an online image and seamlessly embedding the ads at the most appropriate positions within this image. The ads are selected to be globally relevant to the Web page content, as well as locally relevant to the image content and its surrounding text. Meanwhile, the ads are embedded at the most non-salient positions within the image. By leveraging visual content analysis, we are on the positive side to better solve two challenges in online advertising, that is, ad *relevance* and ad *position*. Therefore, ImageSense is able to achieve effective advertising which is a powerful complement to conventional contextual advertising. To the best of our knowledge, this is one of the first attempts to investigate how visual content (both high-level concepts and low-level appearance) can benefit online image advertising. Figure 1 (c) gives an example of ImageSense. It is observed that a relevant ad (i.e., the NBA logo) has been inserted into the region with less information within this image. Please note that this ad logo can also progressively appear, and then disappear after a few seconds without affecting user viewing experience, which is less intrusive and more visually pleasant than conventional advertising.

The rest of the paper is organized as follows. Section 2 reviews closely related research on advertising. Section 3 provides a system overview of ImageSense. The details of ImageSense are described in Section 4. Section 5 gives the experiments, followed by conclusions in Section 6.

2. RELATED WORK

Driven by the huge online business opportunities, online advertising has become emerging areas of research. One of the fundamental problems in online advertising is ad relevance which in studies detracts from user experience and increases the probability of reaction [19]. The other one is how to pick suitable keywords or Web pages for advertising. The literature review will focus on two key problems: 1) *ad keyword selection*, and 2) *ad relevance matching*.

Typical advertising systems analyze a Web page or query to find prominent keywords or categories, and then match these keywords or categories against the words for which advertisers bid. If there is a match, the corresponding ads will be displayed to the user through the web page. Yih *et al.* has studied a learning-based approach to automatically extracting appropriate keywords from Web pages for advertisement targeting [35]. Instead of dealing with general Web pages, Li *et al.* proposed a sequential pattern mining-based method to discover keywords from a specific broadcasting content domain [16]. In addition to Web pages, queries also play an important role in paid search advertising. In [29], the queries are classified into an intermediate taxonomy so that the selected ads are more targeted to the query.

Research on ad relevance has proceeded along three directions from the perspective of what the ads are matched against: (1) keyword-targeted advertising (also called "paid search advertising" or "sponsored search") in which the ads are matched against the originating query [12] [20], (2) content-targeted advertising (also called "contextual advertising") in

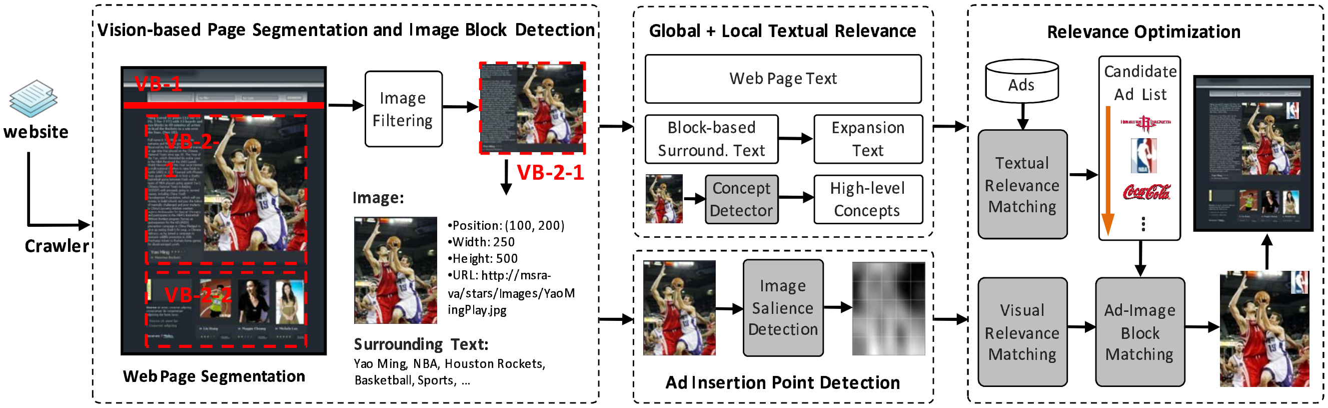

website
Crawler
Vision-based Page Segmentation and Image Block Detection
VB-1
VB-2-1
VB-2-2
Web Page Segmentation
Image Filtering
VB-2-1
Image:
•Position: (100, 200)
•Width: 250
•Height: 500
•URL: http://msra-va/stars/Images/YaoMingPlay.jpg
Surrounding Text:
Yao Ming, NBA, Houston Rockets, Basketball, Sports, …
Global + Local Textual Relevance
Web Page Text
Block-based Surround. Text
Expansion Text
Concept Detector
High-level Concepts
Image Salience Detection
Ad Insertion Point Detection
Relevance Optimization
Ads
Textual Relevance Matching
Candidate Ad List
Visual Relevance Matching
Ad-Image Block Matching

Figure 2: System framework of ImageSense.

which the ads are associated with the Web page content rather than the keywords [5] [26], and (3) user-targeted advertising (also called "audience intelligence") in which the ads are driven based on user profile and demography [11], or behavior [27]. Although the paid search market develops quicker than contextual advertising market, and most textual ads are still characterized by "bid phrases," there has been a drift to contextual advertising as it supports a long-tail business model [15]. For example, a recent work [26] examines a number of strategies to match ads to Web pages based on extracted keywords. To alleviate the problem of exact keyword match in conventional contextual advertising, Broder *et al.* propose to integrate semantic phrase into traditional keyword matching [5]. Specifically, both the pages and ads were classified into a common large taxonomy, which was then used to narrow down the search of keywords to concepts. Most recently, Hu *et al.* propose to predict user demographics from browsing behavior [11]. The intuition is that while user demographics are not easy to obtain, browsing behaviors indicate a user's interest and profile.

Recently, researchers have invented context-aware video ads that can be inserted into video clips using intelligent video content analysis techniques [22] [30]. However, due to current network bandwidth limitations, video is still not as popular in Web pages as text and image.

3. SYSTEM OVERVIEW

Figure 2 illustrates the system framework of ImageSense, which consists of four components, i.e., vision-based page segmentation and image block detection, textual relevance matching, ad insertion point detection, and relevance optimization. Given a Web page which contains some images, a vision-based page segmentation algorithm is applied to partition the page into several blocks with consistent content. The images suitable for advertising are then picked and represented by visual content and surrounding text (such as the titles and descriptions) within the corresponding blocks. A textual relevance module ranks the ads according to the global and local textual relevance between the ads and the content of Web page and images. The textual information of each image comes from the text of the Web page, the surrounding text in the corresponding block, the expansion words of the surrounding text, and the automatically recognized concepts based on visual content analysis. Meanwhile, candidate ad insertion positions are detected in each image through visual saliency analysis. The ads will be reranked by simultaneously considering local content relevance which is derived from the visual similarity between each ad and ad insertion point, as well as the intrusiveness of ad insertion point. Finally, the most contextually relevant ads will be embedded at the most non-intrusive position within each image by a relevance optimization module.

4. IMAGESENSE

In this section, we describe the implementation of ImageSense. We will show how the tasks of associating ads with the triggering page and image are formulated as an optimization problem and how this problem is practically solved using a heuristic searching approach.

4.1 Problem Definition

Without loss of generality, we define the task of ImageSense as the association of ads with a Web page. We will show that it is easy to extend to the tasks of image- and website- based advertising in a single formulation. Let $\mathbf{I}$ denote an image in the Web page $\mathbf{P}$, which consists of N_b ad insertion points, represented by $\mathcal{B} = \{b_i\}_{i=1}^{N_b}$. These insertion points are usually blocks or regions in the image. Let $\mathcal{A}$ denote the ad database containing N_a ads, represented by $\mathcal{A} = \{a_j\}_{j=1}^{N_a}$. The task of ImageSense can be described as given an image $\mathbf{I}$ with a set of ad insertion points $\mathcal{B}$ and an ad database $\mathcal{A}$, select N elements from $\mathcal{B}$ and $\mathcal{A}$, and associate each $a_j \in \mathcal{A}$ with an appropriate $b_i \in \mathcal{B}$. N is the number of expected ads to be inserted into each image, which can be given by the publisher. Note that in conventional contextual advertising (e.g, Google AdWords and AdSense), the position of selected ads is usually fixed to a preserved block within the Web page. Thus, the main focus is only the selection of the most relevant ads to the Web page. However, a set of appropriate ad positions is automatically detected within each individual image in ImageSense. Therefore, not only the relevant ads to the image but also the association of the ads with the detected insertion points should be decided.

As we have mentioned in Section 3 that the ads are relevant not only to the Web page content but also to the local surrounding text, as well as the neighboring image blocks around the insertion points, we introduce the following three relevance items for each ad.

- **Global textual relevance** $R_g(\mathbf{P}, a_j)$—the textual relevance between the Web page $\mathbf{P}$ and the ad a_j, which is given by the cosine similarity in the vector space model, as in most of conventional advertising systems [1] [14] [26]. Since $\mathbf{P}$ is always given, we can drop $\mathbf{P}$ and rewrite such relevance as $R_g(a_j)$.

- **Local textual relevance** $R_\ell(\mathbf{I}, a_j)$—the textual relevance between the image $\mathbf{I}$ and the ad a_j, where the text associated with $\mathbf{I}$ comes from not only the surrounding text in the corresponding page block, but also the hidden concepts recognized by visual content analysis. Since $\mathbf{I}$ is always given, we can also drop $\mathbf{I}$ and rewrite such relevance as $R_\ell(a_j)$.

- **Local content relevance** $R_c(b_i, a_j)$—the combination of content-based visual similarity between the ad insertion position b_i and the ad a_j, as well as the non-intrusiveness of b_i for inserting the ads. To minimize intrusiveness to the user, the ads are to be inserted into the most non-intrusive positions, and to be similar to the corresponding neighboring image blocks.

To support effective image advertising, the three kinds of relevances should be simultaneously maximized. We integrate the above properties into an optimization problem. Suppose we introduce the following design variables $\mathbf{x} \in \mathbb{R}^{N_b}$, $\mathbf{y} \in \mathbb{R}^{N_a}$, $\mathbf{x} = [x_1, \ldots, x_{N_b}]^T$, $x_i \in \{0, 1\}$, and $\mathbf{y} = [y_1, \ldots, y_{N_a}]^T$, $y_j \in \{0, 1\}$, where x_i and y_j indicate whether b_i and a_j are selected ($x_i = 1, y_j = 1$) or not ($x_i = 0, y_j = 0$). The above expectation can be formulated as the following nonlinear 0-1 integer programming problem (NIP) [3],

$$
\begin{aligned}
\max_{(\mathbf{x},\mathbf{y})} f(\mathbf{x}, \mathbf{y}) &= w_g \sum_{j=1}^{N_a} y_j R_g(a_j) + w_\ell \sum_{j=1}^{N_a} y_j R_\ell(a_j) \qquad (1)\\
&\quad + w_c \sum_{i=1}^{N_b} \sum_{j=1}^{N_a} x_i y_j R_c(b_i, a_j)\\
&= w_g \mathbf{y}^T \mathbf{R}_g + w_\ell \mathbf{y}^T \mathbf{R}_\ell + w_c \mathbf{x}^T \mathbf{R}_c \mathbf{y}\\
s.t. \quad & \sum_{i=1}^{N_b} x_i = N, \ \sum_{j=1}^{N_a} y_j = N, \ x_i, y_j \in \{0, 1\}
\end{aligned}
$$

where $\mathbf{R}_g = [R_g(a_1), R_g(a_2), \ldots, R_g(a_{N_a})]^T$, $\mathbf{R}_\ell = [R_\ell(a_1), R_\ell(a_2), \ldots, R_\ell(a_{N_a})]^T$, $\mathbf{R}_c \in \mathbb{R}^{N_b \times N_a}$. The parameters (w_g, w_ℓ, w_c) control the emphasis on global and local textual relevance, as well as local content relevance, and satisfy the constraints: $0 \leqslant w_g, w_\ell, w_c \leqslant 1$ and $w_g + w_\ell + w_c = 1$. The parameters of (w_g, w_ℓ, w_c) can be empirically set to $w_c < w_\ell < w_g$, as it is reasonable to assume that textual relevance is more important than image content relevance [10] [21]. Alternatively, the parameters can be trained by cross-validation experiments.

Although the above equation is based on the assumption of one pair ($\mathbf{I}$, $\mathbf{P}$), it is reasonable to treat each pair independently using the same equation if $\mathbf{P}$ contains more than one image, since different images belong to different page blocks. Furthermore, the above formulation can be easily extended to support both web site- and image- based advertising. For example, if we drop $w_g \mathbf{y}^T \mathbf{R}_g$ in equation (1), then it can support image-based advertising. To support web sites, the triggering pages (or ad landing pages) are first crawled, and then each Web page can be treated using equation (1).

4.2 Preprocess: Vision-based Web Page Segmentation and Image Block Detection

Given a Web page, it is desirable to first segment it into several blocks with coherent topic, detect the blocks with suitable images for advertising, and extract the semantic structure such as the images and their surrounding texts from these blocks. Therefore, the image content and surrounding text obtained in this step can be further used for computing the local textual and content relevance.

The Vision-based Page Segmentation (VIPS) algorithm [6] is adopted to extract the semantic structure of a Web page based on its visual presentation. The VIPS algorithm first extracts all the suitable blocks from the Document Object Model (DOM) 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. Figure 3 illustrates the vision-based structured of a sample page. It is observed that this page has two main blocks and the block "VB-1-2-1-2" is detected as an image block.

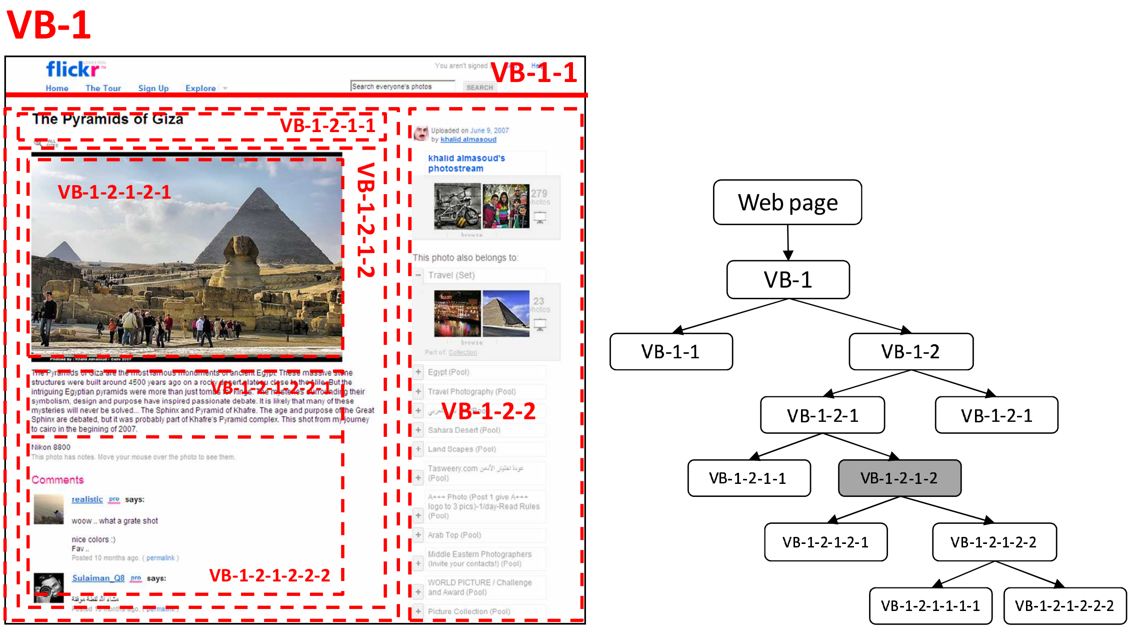
The VIPS algorithm can be naturally used for surrounding texts extraction. Specifically, after obtaining all the blocks via VIPS, we elaborately select the images which are suitable for advertising in the Web page. Intuitively, the images with poor visual qualities, or belonging to the advertisements (usually placed in certain positions of a page) or decorations (usually are too small), are first filtered out. Then, the corresponding blocks with the remaining images are selected as the advertising page blocks. The surrounding texts (e.g., *title* and *description*) are used to describe each image.

4.3 Global and Local Textual Relevance

After page segmentation, each suitable image $\mathbf{I}$ can be represented by $\mathbf{I} = \{\mathbf{I}^{(V)}, \mathbf{I}^{(T)}\}$, where $\mathbf{I}^{(V)}$ and $\mathbf{I}^{(T)}$ denote the visual and textual information associated with $\mathbf{I}$, respectively. In general, $\mathbf{I}^{(T)}$ is composed of an image *title* and a textual *description* in the corresponding page block. Since expansion terms of the triggering page can improve textual relevance [26], we expand the *title* and the *description* by classifying the textual content into a category of a predefined hierarchy. The predefined taxonomy consists of more than 1,000 words [33] which are trained by text categorization techniques based on Support Vector Machine (SVM) [34]. Furthermore, to deal with the case that there are very few words associated with image $\mathbf{I}$, we leverage concept detection to classify the visual content of each image into a set of predefined concepts [21] [1]. These concepts are elaborately selected from a light ontology used in TRECVID community [31] in terms of frequent presence in real-world [2]. We have achieved 11.92% in terms of the mean average precision (MAP) on these concepts in TRECVID 2007 [21]. As

[1] As mentioned, photo-sharing sites like Flickr [8] and photo blogs like MySpace [24] usually have very poor or few tags. Therefore, it is desirable to obtain the hidden text from the visual content to discover further ad matches.

[2] These concepts include *Airplane*, *Animal*, *Boat_Ship*, *Building*, *Bus*, *Car*, *Charts*, *Computer_TV-screen*, *Court*, *Crowd*, *Desert*, *Face*, *Maps*, *Meeting*, *Mountain*, *Office*, *Outdoor*, *Person*, *Road*, *Sky*, *Snow*, *Sports*, *Studio*, *Truck*, *Urban*, *Vegetation*, *Waterscape_Waterfront*, and *Weather*.



(a) Segmentation result (b) "VB" structural tree

Figure 3: Vision-based structure of a sample Web page. "VB" indicates the visual block. The surrounding texts are extracted within the image blocks, i.e., "VB-1-2-1-2."

a result, the textual content $\mathbf{I}^{(T)}$ can be decomposed into three parts, i.e., $\mathbf{I}^{(T)} = \mathbf{I}^{(T_1 \cup T_2 \cup T_3)} = \{\mathbf{I}^{(T_1)}, \mathbf{I}^{(T_2)}, \mathbf{I}^{(T_3)}\}$, where $\mathbf{I}^{(T_1)} = \{\textit{title}, \textit{description}\}$, $\mathbf{I}^{(T_2)} = \{\textit{expansion words of } \mathbf{I}^{(T_1)}\}$, and $\mathbf{I}^{(T_3)} = \{\textit{concepts detected based on } \mathbf{I}^{(V)}\}$. Figure 4 shows an example of the original surrounding text, expansion text, and concept text. The corresponding image is extracted from "VB-1-2-1-2" in Figure 3.

Correspondingly, each ad a_j can be represented by $a_j = \{a_j^{(V)}, a_j^{(T)}\}$, where $a_j^{(V)}$ and $a_j^{(T)}$ denote the visual and textual information associated with a_j, respectively. Note that $a_j^{(V)}$ corresponds to the visual content of ad logo and $a_j^{(T)}$ the original text provided by the advertiser without expansion words and concepts. The textual information $a_j^{(T)}$ consists of four structural parts: a *title*, *keywords*, a textual *description*, and a *hyperlink*.

We use Okapi BM25 algorithm [28] which is widely used in information retrieval to compute the relevance between an ad landing page and an ad [21]. Given a query q with n terms $\{q_1, q_2, \ldots, q_n\}$, the BM25 score of a document d is

$$sim(q,d) = \sum_{i=1}^{n} \frac{idf(q_i) \times tf(q_i,d) \times (k+1)}{tf(q_i,d) + k(1 - b + b \times ndl(d))} \tag{2}$$

where $tf(q_i,d)$ is the term frequency of q_i in d, $ndl(d) = |d|/|\overline{d}|$ is normalized document length, $|d|$ and $|\overline{d}|$ indicate the document length of d and the average document length in document collection, k and b are two parameters and usually chosen as $k = 2.0$ and $b = 0.75$, $idf(q_i)$ is the inverse document frequency of the query q_i, give by

$$idf(q_i) = \log \frac{M - m(q_i) + 0.5}{m(q_i) + 0.5}$$

where M is the total number of documents in the collection, and $m(q_i)$ is the number of documents containing q_i.

Let $\mathbf{P}^{(T)}$ denote the textual information of Web page $\mathbf{P}$, the global textual relevance $R_g(\mathbf{P}, a_j)$ is given by

$$R_g(\mathbf{P}, a_j) = sim(\mathbf{P}^{(T)}, a_j^{(T)}) \tag{3}$$

By considering $\mathbf{P}^{(T)}$ as the query and $a_j^{(T)}$ as the document, we can rank the ads with regard to the Web page $\mathbf{P}$, which is widely applied in most existing contextual advertising systems [1] [5] [14] [23] [26]. We adopt the global textual relevance for ad ranking as the baseline in ImageSense. Accordingly, the local textual relevance $R_\ell(\mathbf{I}, a_j)$ is

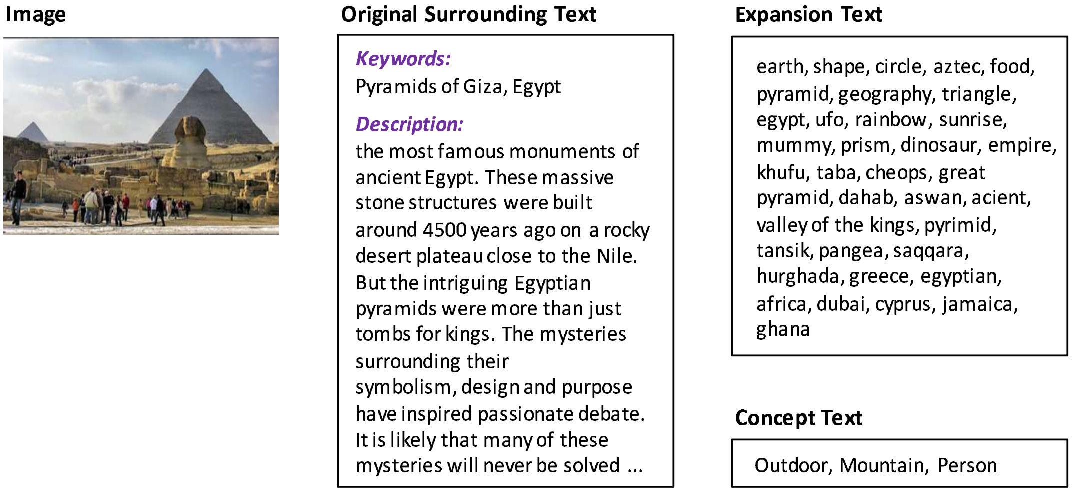
$$R_\ell(\mathbf{I}, a_j) = sim(\mathbf{I}^{(T)}, a_j^{(T)}) = sim(\mathbf{I}^{(T_1 \cup T_2 \cup T_3)}, a_j^{(T)}) \tag{4}$$

The fusion of global and local textual relevance can be regarded as ad reranking. We will investigate how local relevance can benefit ad selection in the experiments.

4.4 Local Content Relevance

Note that ImageSense will embed the contextually relevant ads at certain spatial positions within an image. With analysis on local content relevance, we aim to find the non-intrusive ad positions within an image and select the ads whose logos are visually similar or have similar visual style to the image, in order to minimize intrusiveness and improve user experience. Specifically, the candidate ad insertion positions (usually the blocks in $\mathcal{B}$) are detected based on a salience measure which is derived from an image saliency map, while visual similarity is measured based on the HSV color feature. These two measures are combined for computing the local content relevance $R_c(b_i, a_j)$.

Given an Image $\mathbf{I}$ which is represented by a set of blocks $\mathcal{B} = \{b_i\}_{i=1}^{N_b}$, a saliency map $\mathbf{S}$ representing the importance of the visual content is extracted for image $\mathbf{I}$, by investigating the effects of contrast in human perception [18]. Figure 5(b) shows an example of the saliency map of an image. The



Image

Original Surrounding Text

Keywords:
Pyramids of Giza, Egypt

Description:
the most famous monuments of ancient Egypt. These massive stone structures were built around 4500 years ago on a rocky desert plateau close to the Nile. But the intriguing Egyptian pyramids were more than just tombs for kings. The mysteries surrounding their symbolism, design and purpose have inspired passionate debate. It is likely that many of these mysteries will never be solved ...

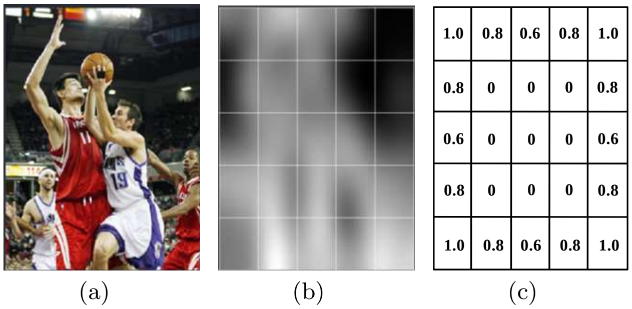
Expansion Text

earth, shape, circle, aztec, food, pyramid, geography, triangle, egypt, ufo, rainbow, sunrise, mummy, prism, dinosaur, empire, khufu, taba, cheops, great pyramid, dahab, aswan, acient, valley of the kings, pyrimid, tansik, pangea, saqqara, hurghada, greece, egyptian, africa, dubai, cyprus, jamaica, ghana

Concept Text

Outdoor, Mountain, Person

Figure 4: An example of the original surrounding text, expansion text, and concept text. The image is extracted from "VB-1-2-1-2" in Figure 3.



1.0	0.8	0.6	0.8	1.0
0.8	0	0	0	0.8
0.6	0	0	0	0.6
0.8	0	0	0	0.8
1.0	0.8	0.6	0.8	1.0

(a) (b) (c)

Figure 5: An example of saliency map and weight map of an image. (a) I: original image, (b) S: saliency map with weight grids overlaid, (c) W: weight map. The brighter the pixel in the saliency map, the more important or salient it is.

brighter the pixel in the salience map, the more salient or important it is. Intuitively, the ads should be embedded at the most non-salient regions, so that the informative content of the image will not be occluded and the users may not feel intrusive. Meanwhile, the image block set $\mathcal{B}$ is obtained by partitioning image $\mathbf{I}$ into $M \times M$ grids. Each grid corresponds to a block b_i ($i = 1, 2, \ldots, N_b$) where $N_b = M^2$, and also a candidate ad insertion point. Consequently, there are N_b candidate ad insertion points in total. For each block b_i, a saliency energy s_i ($0 \leqslant s_i \leqslant 1$) is computed by averaging all the normalized energies of the pixels within b_i. As the saliency map does not consider the spatial importance for ad insertion, a weight map $\mathbf{W} = \{w_i\}_{i=1}^{N_b}$ is designed to weight the energy s_i, so that the ads will be inserted into the corners or sides rather than the center blocks. Figure 5(c) shows an example of the weight map. Therefore, $w_i \times (1 - s_i)$ indicates the suitability of block b_i for embedding an ad.

In addition to the non-salient ad insertion position, the ads are assumed to have similar appearance with the neighboring blocks around the insertion position. To measure the visual similarity between a_j and b_i, we compute the L_1 distance between $a_j^{(V)}$ and the neighboring block set B_i of b_i. The distance is defined in a HSV color space which has been widely adopted in visual content retrieval and search systems [21], given by

$$d(B_i, a_j^{(V)}) \;=\; \frac{1}{|B_i|} \sum_{b_i \in B_i} \sum_{k=1}^{K} \left| f^{(b_i)}(k) - f^{(a_j)}(k) \right| \qquad (5)$$

where $|B_i|$ denotes the number of blocks adjacent with b_i, $f^{(b_i)}(k)$ and $f^{(a_j)}(k)$ denote the k-th color feature of block b_i and ad a_j, respectively. K is the feature dimension. For example, if we use a 64-dimensional HSV color histogram to describe the visual content, then $K = 64$. As a result, the local content relevance $R_c(b_i, a_j)$ between a candidate ad insertion point b_i and an ad a_j is given by

$$R_c(b_i, a_j) = w_i \times (1 - s_i) \times \left(1 - d(B_i, a_j^{(V)})\right) \qquad (6)$$

4.5 Problem Solution

It is observed that there are $C_{N_a}^{N} C_{N_b}^{N} N!$ solutions in total to equation (1). When the number of elements in $\mathcal{A}$ and $\mathcal{B}$ is large, the searching space for optimization increases dramatically. For instance, if there are 10,000 ads in $\mathcal{A}$, 16 blocks in $\mathcal{B}$ (see Figure 5), and three ads embedded in each image, then the number of solutions will be nearly $5.6 \times 10^{+14}$. When a Web page $\mathbf{P}$ contains more than one image for advertising, the searching space will be extremely large. However, we can use the Genetic Algorithm (GA) to find solutions approaching the global optimum. Practically, we can adopt the following heuristic searching algorithm similar to that used in [22] to solve the above problem. Algorithm 1 give the details of the practical solution. As a result, the number of possible solutions is significantly reduced to $O(N_b + N_a + N \times N_a')$.

5. EXPERIMENTS

We evaluated the effectiveness of ImageSense from the perspectives of ad relevance and user experience.

5.1 Data and Methodology

In ImageSense, we set $M = 5$ so that there are 16 candidate ad insertion points in Section 4.4. Note that the

Algorithm 1 The heuristic searching algorithm for Eq. (1)

```
1: Initialize: set the labels of all the elements in $\mathbf{x}$ and $\mathbf{y}$ as "0" (i.e., "unselected").
2: Rank all the elements in $\mathbf{x}$ according to $(w_i \times (1 - s_i))$ in a descendent order, select the top $N$ elements, and set their labels as "1" (i.e., "selected").
3: Rank all the elements in $\mathbf{y}$ according to $w_g R_g + w_\ell R_\ell$ in a descendent order, and select the top $N'_a$ $(N < N'_a \ll N_a)$ elements.
4: For each $x_i$ in the top $N$ elements in $\mathbf{x}$, select the unselected $y_j$ from the top $N'_a$ elements in $\mathbf{y}$ with the $\max\{w_c R_c + w_g R_g + w_\ell R_\ell\}$, and set the label of $y_j$ as "1".
5: Output all the pairs with $(x_i = 1, y_j = 1)$.
```

blocks with the weight "0" are neglected. In conventional evaluations on textual relevance [5] [14] [26], the number of relevant ads is usually fixed to three (i.e., $N = 3$), while we have set $N = 5$ for evaluating ad relevance and $N = 1$ for evaluating user experience in ImageSense, as well as $N'_a = 100$. For concept text, we only selected the top five concept words according to the confidence scores as current performance of concept detection is still not satisfying. A large-scale and general concept ontology rather than the one used in TRECVID will help improve relevance matching within the same framework. As there are currently very few ad-networks supporting real collection of ads with product logos, we have collected around $7,285$ unique ad product logos from innovativeclipart.com. We invited 20 subjects to annotate each product logo (i.e., *title*, *keywords*, *description*, and *hyperlink*). As a result, these ads were associated with $32,480$ unique words.

The implementation of ImageSense consists of two part: 1) offline process including crawler, page segmentation, image block detection, image saliency extraction, and inverted index building of ad database; 2) online matching including global and local relevance computation, as well as optimization. According to the above parameters, the computational time for online matching is around 0.2 second. Note that it typically takes less than 0.1 second for ranking all the ad logos in the database based on the inverted index.

We prepared three photo-sharing sites with around one million high-quality photos in total. Specifically, we collected $443,895$ images from Corbis.com, $382,371$ images from Tango (an internal photo-sharing site within Microsoft Research Asia), and around $200,000$ images from Flickr.com. We created two websites based on the images and their surrounding texts for Corbis and Flickr. To evaluate ad relevance, the 100 representative top image queries from Live image search [17] were put to the three sites as the queries and the top 10 images were returned for each query. Table 1 lists some representative image queries. As a result, there are $1,000$ images and the corresponding surround texts.

Table 1: The representative top image queries.

party, pics, pyramids, dragon, girl, baby, hearts, zebra, spring, wallpaper, car, cat, rose, money, dog, flower, chat, butterfly, fish, graffiti, cute kittens, angel, nature, football, puppies, wedding cakes, horse

In addition, similar to text advertising [5] [14] [26], we quantified the effect of ImageSense supporting Web page-based advertising using a set of 100 Web pages. These pages were selected from the most recent news in several major news web sites, such as Yahoo! [32], CNN [7], MySpace [24], as well as the popular pages in several major photo-sharing web sites, such as Flickr [8], photoSIG [25]. Consequently, there were 1,100 triggering pages including 1,000 images and 100 pages used as the test bed. For each triggering page, we returned the top five ad logos by the following advertising strategies for evaluation, which is similar to the experimental settings for general Web page advertising [14] [26]. Note that the following results were produced according to different settings of the parameters in equation (1).

- **I: Global textual relevance** ($w_g = 1$, $w_\ell = 0$, $w_c = 0$). This is similar to conventional contextual advertising in which the ads are selected according to the generic content of the Web page [26]. This is used as the baseline for comparison.
- **II: Global and local textual relevance without concept** ($w_g = 0.3, w_\ell = 0.5, w_c = 0, \mathbf{I}^{(T)} = \mathbf{I}^{(T_1 \cup T_2)}$). In addition to global textual relevance, local textual relevance based on Web page segmentation and image block detection is also taken into account, while concept text is neglected.
- **III: Global and local textual relevance** ($w_g = 0.3$, $w_\ell = 0.5$, $w_c = 0$, $\mathbf{I}^{(T)} = \mathbf{I}^{(T_1 \cup T_2 \cup T_3)}$). Both global and local textual relevance is linearly combined for ad ranking, with concept text included in the computation of local textual relevance.
- **IV: Global and local textual relevance, as well as local content relevance** ($w_g = 0.33$, $w_\ell = 0.33$, $w_c = 0.33$). All the relevance including global and local textual relevance, as well as local content relevance, is linearly combined. The weights are empirically set to be equal for comparison, as we aim to investigate the different contributions of multimodal relevance for image advertising.
- **V: Global and local textual relevance, as well as local content relevance, with different weights to IV** ($w_g = 0.3$, $w_\ell = 0.5$, $w_c = 0.2$). All the relevance including global and local textual relevance, as well as local content relevance, is linearly combined. Note that the weights are set different according to different contributions from the multimodal relevance. This strategy is the optimal setting of ImageSense.

As a result, there were $27,500$ $(1100 \times 5 \times 5)$ page-ad pairs. For each triggering page, we selected no more than 25 ads. The ads, which consisted of *logo*, *title*, *description*, and *hyperlink*, appeared besides the images in the Web pages. There were 15 subjects with different background invited to manually judge the page-ad relevance. These subjects did not know by which among the above five strategies the current results were produced. Each page-ad pair has been judged by three or more subjects on a 1-3 scale [5] as follows.

- **Irrelevant (1)**. The ad is definitely unrelated to the image or the corresponding Web page.
- **Somewhat relevant (2)**. The ad is related to the secondary subject of the image and the Web page, or related to the image or the Web page in a general way.
- **Relevant (3)**. The ad is directly related to the main subject of the image and the Web page content.

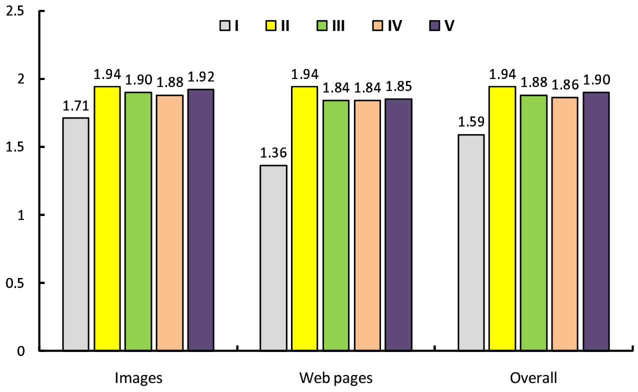


Figure 6: The evaluations on ad relevance.

Table 2: Evaluations on ad position and satisfaction.

Satisfaction score	Conventional ads	ImageSense
Ad position	3.00	3.75
Overall	3.68	3.85

5.2 Evaluation on Ad Relevance

After page segmentation, there were averagely 86.74 words per page in the 1100 triggering pages. In the following, we assumed that the page-ad pairs judged with “2” or “3” are positive and those judged with “1” are negative. We have also examined counting only the pairs judged with “3” as positive, but did not find a significant change in the relative performance. For each triggering page, there were 15.05 ads returned among which 9.13 ads were judged relevant.

Figure 6 shows the average evaluations on ad relevance using the five advertising strategies described in Section 5.1. It is reasonable to observe that combining surrounding text can achieve better relevance than using only the text in the whole page. We can see that local content relevance sometimes degraded relevance, which indicates the visual similarity between ads and ad insertion points cannot improve relevance. We believe local content relevance is a tradeoff between ad relevance and user experience. Overall, we can see that ImageSense can improve the contextual relevance for image advertising over conventional text advertising.

Figure 7 shows the precision-recall curves of the five strategies. We used 11-point average figures to quantify the precision [2] [26]. As we are not able to evaluate the entire ad collection, the recall values are relative to the set of evaluated ads. It is observed that the local surrounding text obtained by Web page segmentation significantly improved ad relevance over the general text in the Web pages. Based on local textual information by strategy II, strategy III, IV, and V achieved steady improvement over strategy I. We can also see that strategy III and V achieved the best performance in terms of relevance, while strategy IV underperformed V. This has proved the assumption on different contributions from different kinds of textual information—the local surround texts of the images and their expansion words are more important than the general Web page content and visual content from the perspective of ad relevance. We believe an optimal set of parameters of (w_g, w_ℓ, w_c) can further improve the advertising performance.

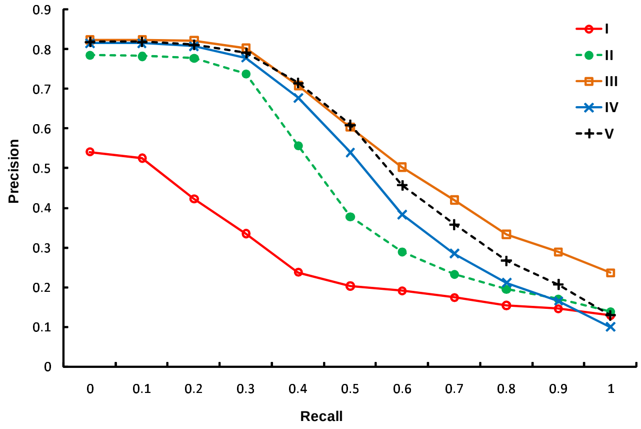


Figure 7: Precision-recall curves.

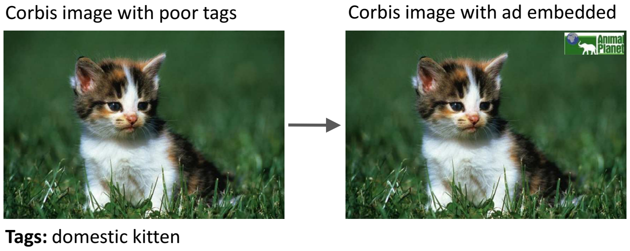


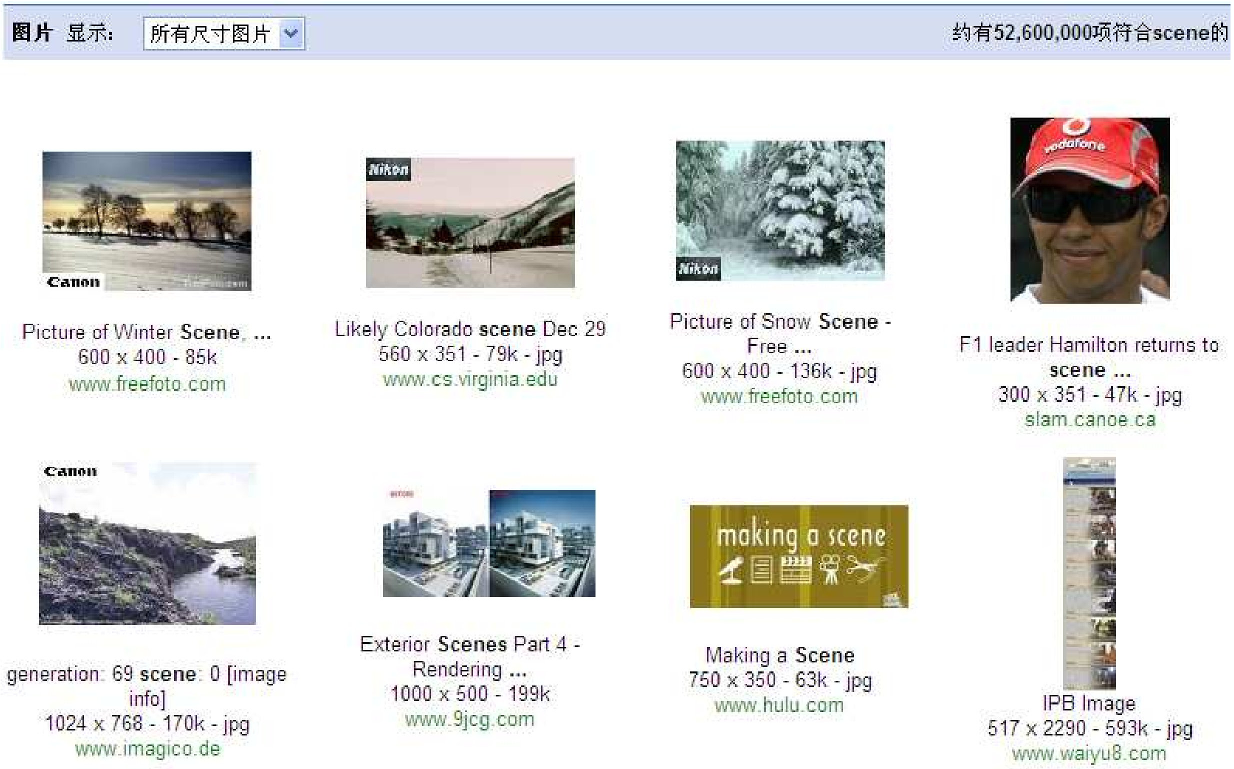
Figure 8: An example of the image before (left) and after (right) advertising.

5.3 Evaluation on User Experience

We evaluated user experience in terms of ad insertion positions and overall satisfaction from the perspective of viewers. The 15 subjects involved in our experiments were invited to give a satisfaction score (1-5) of the ad position and overall satisfaction. The higher score indicates the higher satisfaction. We compared ImageSense to the setting in conventional advertising like Google AdSense [1] where a textual ad unit (i.e., *title*, *description*, and *hyperlink*) is displayed adjacent to the image on the right. For each triggering page, two results including conventional advertising and ImageSense were provided to the user. Each page was judged by three different subjects. Note that the ads were identical in the two results except for different presentations: the ads in conventional advertising are textual while the ads in ImageSense are product logos. The satisfaction score for ad position in conventional advertising was fixed to 3.0 for comparison. Table 2 lists the evaluation results. It is observed that ImageSense has provided better ad positions and overall satisfaction, and thus achieved better user experience than conventional advertising. Table 2 corresponds to the fact that there were 78% and 55% judgments in which the score of ImageSense is higher than conventional advertising in terms of ad position and overall satisfaction, respectively.

5.4 Advertising Examples

Figure 8 shows an example of image advertising by the five strategies. We can see that the image was tagged only with two words (“domestic” and “kitten”). The final advertising result by ImageSense (based on strategy V) is shown in Figure 8. Figure 9 shows an example of ImageSense for a



图片 显示： 所有尺寸图片

约有52,600,000项符合scene的

Picture of Winter Scene, ...
600 x 400 - 85k
www.freefoto.com

Likely Colorado scene Dec 29
560 x 351 - 79k - jpg
www.cs.virginia.edu

Picture of Snow Scene - Free ...
600 x 400 - 136k - jpg
www.freefoto.com

F1 leader Hamilton returns to scene ...
300 x 351 - 47k - jpg
slam.canoe.ca

generation: 69 scene: 0 [image info]
1024 x 768 - 170k - jpg
www.imagico.de

Exterior Scenes Part 4 - Rendering ...
1000 x 500 - 199k
www.9jcg.com

Making a Scene
750 x 350 - 63k - jpg
www.hulu.com

IPB Image
517 x 2290 - 593k - jpg
www.waiyu8.com

Figure 9: An example of advertising for image search result page. The query is "scene." The first three images in the top row and the first two in the bottom row are selected for advertising.

page corresponding to the image search results of "scene." It is observed that after image block detection, there are five images selected for advertising.

6. CONCLUSIONS

We have presented an innovative system for contextual in-image advertising. The system, called ImageSense, is able to automatically decompose the Web page into several coherent blocks, select the suitable images from these blocks for advertising, detect the nonintrusive ad insertion positions within the images, and associate the relevant advertisements with these positions. In contrast to conventional ad-networks which treat image advertising as general text advertising, ImageSense is fully dedicated to image and aims to monetize the visual content rather than surrounding text. We formalize the association of ads with images as an optimization problem which maximizes the overall contextual relevance. We have conducted extensive experiments over a large-scale of real-world image data and achieved promising results in terms of ad relevance and user experience. To the best of our knowledge, this is one of the first works showing how visual content analysis at different levels (i.e., low-level appearance and high-level semantic) can yield more effective image advertising, especially from the perspectives of ad *position* detection and contextual *relevance*.

7. REFERENCES

[1] AdSense. http://www.google.com/adsense/.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] BritePic. http://www.britepic.com/.

[5] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *Proceedings of SIGIR*, 2007.

[6] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. VIPS: a vision-based page segmentation algorithm. In *Microsoft Technical Report (MSR-TR-2003-79)*, 2003.

[7] CNN. http://www.cnn.com/.

[8] Flickr. http://www.flickr.com/.

[9] R. J. Gerrig and P. G. Zimbardo. *Psychology and Life (16 Edition)*. Allyn & Bacon, 2001.

[10] X. He, D. Cai, J.-R. Wen, W.-Y. Ma, and H.-J. Zhang. Clustering and searching www images using link and page layout analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(2), 2007.

[11] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of WWW*, 2007.

[12] A. Joshi and R. Motwani. Keyword generation for search engine advertising. In *Proceedings of the Workshops of IEEE International Conference on Data Mining*, 2006.

[13] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How Flickr helps us make sense of the world: Context and content in community-contributed media collections. In *Proceedings of ACM Multimedia*, Augsburg, Germany, 2007.

[14] A. Lacerda, M. Cristo, M. A. Goncalves, *et al.* Learning to advertise. In *Proceedings of SIGIR*, 2006.

[15] H. Li, S. M. Edwards, and J.-H. Lee. Measuring the intrusiveness of advertisements: scale development and validation. *Journal of Advertising*, 31(2):37–47, 2002.

[16] H. Li, D. Zhang, J. Hu, H.-J. Zeng, and Z. Chen. Finding keyword from online broadcasting content for targeted advertising. In *International Workshop on Data Mining and Audience Intelligence for Advertising*, 2007.

[17] Live Image Search. http://www.live.com/?&scope=images.

[18] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of ACM Multimedia*, pages 374–381, Nov 2003.

[19] S. Mccoy, A. Everard, P. Polak, and D. F. Galletta. The effects of online advertising. *Communications of The ACM*, 50(3):84–88, 2007.

[20] A. Mehta, A. Saberi, U. Vazirani, and V. Vazirani. Adwords and generalized on-line matching. *Journal of the ACM*, 54(5), Oct. 2007.

[21] T. Mei, X.-S. Hua, W. Lai, L. Yang, *et al.* MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search. In *TREC Video Retrieval Evaluation Online Proceedings*, 2007.

[22] T. Mei, X.-S. Hua, L. Yang, and S. Li. VideoSense: Towards effective online video advertising. In *Proceedings of ACM Multimedia*, pages 1075–1084, Augsburg, Germany, 2007.

[23] V. Murdock, M. Ciaramita, and V. Plachouras. A noisy channel approach to contextual advertising. In *International Workshop on Data Mining and Audience Intelligence for Advertising*, 2007.

[24] MySpace. http://www.myspace.com/.

[25] photoSIG. http://www.photosig.com/.

[26] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. Moura. Impedance coupling in content-targeted advertising. In *Proceedings of SIGIR*, 2005.

[27] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating click-through rate for new ads. In *Proceedings of WWW*, 2007.

[28] S. E. Robertson, S. Walker, and M. Hancock-Beaulieu. Okapi at TREC-7. In *the Seventh Text REtrieval Conference*, Gaithersburg, USA, Nov. 1998.

[29] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Building bridges for web query classification. In *Proceedings of SIGIR*, 2006.

[30] S. H. Srinivasan, N. Sawant, and S. Wadhwa. vADeo - video advertising system. In *Proceedings of ACM Multimedia*, 2007.

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

[32] Yahoo! Image. http://images.search.yahoo.com/images.

[33] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of CIVR*, 2007.

[34] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of SIGIR*, 1999.

[35] W.-T. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of WWW*, 2006.