Predicting Eye Fixations on Webpage With an Ensemble of Early Features and High-Level Representations from Deep Network

Chengyao Shen, Xun Huang, and Qi Zhao, *Member, IEEE*

***Abstract*—In recent decades, webpages are becoming an increasingly important visual information source. Compared with natural images, webpages are different in many ways. For example, webpages are usually rich in semantically meaningful visual media (text, pictures, logos, and animations), which make the direct application of some traditional low-level saliency models ineffective. Besides, distinct web-viewing patterns such as top-left bias and banner blindness suggest different ways for predicting attention deployment on a webpage. In this study, we utilize a new scheme of low-level feature extraction pipeline and combine it with high-level representations from deep neural networks. The proposed model is evaluated on a newly published webpage saliency dataset with three popular evaluation metrics. Results show that our model outperforms other existing saliency models by a large margin and both low- and high-level features play an important role in predicting fixations on webpage.**

***Index Terms*—Deep learning, visual attention, web viewing, webpage saliency.**

I. Introduction

With the wide spread of Internet and the prevalence of search engine and social network in recent decades, webpages are becoming an increasingly important visual input and information sources for us. According to the stats published online (*Internet Live Stats*[1]), the number of internet users across the world exceeded 3 billion in November 2014. The average time user spend online is also increasing.[2] This trend influences people's life style and companies' marketing strategy. Hence, the study of how users' attention is deployed and directed when browsing a webpage is of great research and commercial value.

The deployment of human attention on natural images has been extensively studied from computer vision and neuroscience perspectives. A commonly referred line of early computational models that predict eye fixations on images were built upon the hypothesis that the saliency of a region is the extent to which it stands out from its neighbor in terms of low-level image statistics, such as luminance, color, edge and density [1], [2]. Those bottom-up saliency models can predict fixations in natural images in an effective way, indicating the importance of low-level features in driving attention. In addition, these low-level features resemble the receptive fields of neurons in early visual path such as V1, and there is evidence [3], [4] showing that neurons in V1 may represent a bottom-up saliency map.

Recent studies, however, show that high-level features such as people and text also contribute a lot to predicting fixations [5]–[10]. Specifically, adding object detectors can dramatically improve performance of computational saliency models [5], [6], [11]. These high-level features can only be recognized in higher areas such as V4 and IT. Some recent studies show that the neuronal activities in V4 are closely correlated with gaze deployment [12]. Given the mounting computational and physiological evidence, it is arguable that both low-level and high-level features play an important role in selective visual attention, though their relative contribution is still unclear.

Compared with natural images, webpages are especially rich in visual media, such as text, pictures, logos and animations [13]. All of them are high-level features that can strongly attract attention, which presents a challenge to existing low-level saliency models (see Fig. 1). Besides, the distinct patterns in people's web-viewing behavior is also different from that on natural images. One interesting pattern is top-left bias, that is, to scan top-left region at the beginning of browsing [2], [14]. Another is banner blindness, which means people will naturally avoid to fixate on banner-like advertisement [15]–[17].

In this work, we propose a saliency framework to leverage the representational power of Deep Neural Network (DNN) on high-level concepts and combine it with low-level visual features to predict the eye fixation deployment on webpages. We first extract visual features on webpages with low-level feature maps including color contrast and orientation, and high-level feature maps from DNN. After feature extraction, we integrate

Manuscript received April 21, 2015; revised September 05, 2015; accepted September 20, 2015. Date of publication October 08, 2015; date of current version October 20, 2015. This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant R-263-000-B32-112 and by the Defense Innovative Research Programme under Grant 9014100596. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

C. Shen, and Q. Zhao are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: scyscyao@gmail.com; eleqiz@nus.edu.sg).

X. Huang is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, and also with the School of Computing, Beihang University, Beijing 100191, China (e-mail: xunhuang1995@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2483370

[1][Online]. Available: http://www.internetlivestats.com/internet-users/

[2][Online]. Available: http://www.emarketer.com/Article/Social-Usage-Involves-More-Platforms-More-Often/1010019

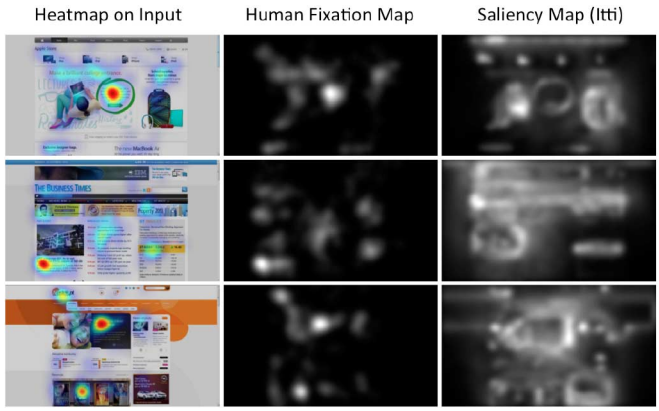


Fig. 1. Illustration on how webpages make early saliency models based on low-level image statistics [1] ineffective. It can be observed that direct integration of low-level image statistics on webpage would generate a saliency map where competition arises almost everywhere, where is quite different from the human fixation map. Heatmaps (left) corresponds to the blurred human fixation map (middle). Areas where users looked the most are colored red. The yellow areas indicate fewer views, followed by the least-viewed blue areas. Gray areas indicate no fixations.

all these feature maps using a linear SVM to generate a saliency map that predict eye fixations on webpages. Experimental results demonstrate that our model outperforms existing saliency models on a recently published webpage saliency dataset [18].

Our main contributions include the following.

Dataset: We collect and analyze an eye fixation dataset (FiWI) consisting of 149 webpages in three categories. Distinctive human-viewing patterns on webpages are confirmed.

Model: The focus of the work is saliency on webpages, and we develop a saliency model that integrates early features and high-level representations from deep network. In comparison, previous works previous works base purely on low-level features [19], purely on high-level features from DNN [9], or a combination of low-level features and hand-crafted detectors [5]. Webpages are usually rich in visual media such as text, pictures, logos and animation, which could induce noisy responses with early features. We apply thresholded center-surround filter on early features to inhibit these noise. We also introduce PCA as a redundancy reduction on high-level representations from DNN and prove that this operation improves the performance. Evaluated on three standard metrics on all the images in the FiWI dataset as well as images from each category separately, it is demonstrated that proposed model our performance all the other 11 saliency models.

II. Related Works

A. Saliency Models on Natural Images

Previous research on saliency mainly focus on predicting saliency on natural images. Some early models are based on the assumption that the conspicuity of a region or object is encoded in low-level features [20]. These models mainly used hand-crafted features such as multi-scale luminance contrast, color contrast and edge orientations [1], [21]–[23]. In addition, there are some models that used machine learning methods to learn the features that predict saliency. For these models, independent component analysis (ICA) and sparse coding are commonly used. For example, SUN [24], ICL [25] and AIM [26] used features learned from ICA with links to information theory. Borji and Itti [27] built their model on features learned with sparse coding. These learned representations are also low-level features that resemble the Gabor filter.

Although the use of low-level features in predicting fixations has been extensively studied, the role of high-level features is far from fully explored. Most works that incorporated high-level features relied on object detectors [5], [6], [28]. However, these approaches do not scale well due to the numerous object categories in the real world. It is almost infeasible to add an object detector for each category of objects that may attract attention. Recent advances of Deep Neural Network (DNN) provide a possibility to incorporate high-level features for eye fixation prediction in a scalable and biologically plausible way. DNN can automatically learn high-level features from large-scale data and have achieved the state-of-the-art performance on a number of image recognition benchmarks [29], [30]. The features encoded in high layers of DNN were found to be highly similar with the responsive stimuli of neurons in V4 and IT [12].

A few recent models therefore employed DNNs to predict saliency. Shen and Zhao [9], [10] utilized a multi-layer sparse coding network to learn hierarchy of features and to predict saliency based on these features. Vig *et al.* [31] generated a large number of hierarchical neuromorphic networks and select features from a few networks that, when combined, gave the best predicting performance. These models are based purely on high-level features from the highest layer of DNN and they perform well on natural images, which suggest the strong capability of high-level DNN features in saliency prediction. Whether features from DNN will be effective in predicting fixations on webpage remains unclear.

B. Web-Viewing Behaviors

Rendered and displayed in a browser, webpages can be seen as a special type of images [2] that contain various visual media contents such as images, audios, text. Yet due to their specific functions of conveying information, certain designing layout, and the abundance of salient stimuli, people's web-viewing behavior is different from that on natural images, which might make saliency models on natural images ineffective. Empirical studies [32] of eye movements on webpages have also revealed several distinct patterns of web-viewing that also shed light on developing a computational model for webpage saliency.

Top-Left Bias: The most distinctive features of web-viewing behavior is the 'F-shaped pattern' of eye movement distribution on webpages [14]. This bias may result from certain design guidelines on webpage layout and the general reading habit of people. Users who have prior browsing experience of webpages would have a general expectation of important information on the top left regions of the webpage. The right bottom regions of webpages rarely attract visual attention during the first second of page viewing [2]. In addition, sheer volume of text on webpage might also result in this 'F-shaped pattern'. Recent studies showed that when viewing webpages, people tend to skim text to obtain the large amount of information [33]. Specially, people

usually spend more time reading earlier paragraphs in a page, and the beginning sentences within a paragraph receive more attention.

Banner-Blindness: Another important pattern of web-viewing is 'ad-avoidance' or usually called 'banner-blindness' [34]. Studies show that, during web-viewing, Internet users tend to consciously or subconsciously avoid looking at banner-like information abundant in salient features which is likely advertisement [15]–[17]. This phenomenon may be triggered by perceived goal impediment, perceived ad clutter during web-viewing and is likely to be caused by prior negative experience [15]. A recent study using eye tracking also revealed that most users would fixate at the banners at least once during their website visit but they usually take actions to reduce their exposure to the ads.

C. Attention Models on Webpages

In recent years, there are several conceptual models and computational models that drop into the user viewing behaviors on different webpages.

1) Conceptual Models: Faraday's visual scanning model [32] represents the first framework that gave a systematic evaluation of visual attention on webpages. This model identified six "salient visual elements"(SAE) in a hierarchy (motion, size, image, color, text-style, and position) that direct our attention in webpages and provided a description of how these elements are scanned by a user. A later research by Grier *et al.* [16] showed that Faraday's model is over-simplified for complex web-viewing behaviors (e.g., the salience order of SAE selected by the model might be inaccurate). Based on Faraday's model, Grier *et al.* described three heuristics ("top left corner of the main content area is dominant", "overly salient items do not contain information", "information of similar type will be grouped together") from their observation and they further proposed a three stage EHS (Expected Location, Heuristic Search, Systematic Search) theory that explains the viewing behavior on webpages. These conceptual models give us a good foundation on developing a computational algorithm to predict webpage saliency.

2) Computational Models Based on Non-Image Feature: The model from Buscher *et al.* [2] that utilized HTML-induced document object model (DOM) is among the most prominent. In [2], the authors first collected data when users were engaged in information foraging and page recognition tasks on 361 webpages from 4 categories (cars, diabetes, kite surfing, wind energy). They then performed a linear regression on features extracted from DOM and generated a model for predicting visual attention on webpages using decision trees. Their linear regression showed that size of the DOM is the most decisive factor and their decision tree get a precision of 75% and a recall of 53% in predicting the eye fixations on webpages. From their data, they also observed that the first few fixations (i.e., during the first second of each page view) are consistent in both tasks. Other models in this category either focus on a specific type of webpages [35] that does not generalize well, or based themselves on text semantics [36] thus quite different from the goal in this work.

3) Computational Models Based on Image Features: Our model falls in this category and there are few works that have been done. One early attempt utilizing image features to predict saliency on webpage is from Still and Masciocchi [13]. The referred work, however, simply applied the classic Itti-Koch model [1] to predict the web-viewing entry points. In our early work [18], we proposed a preliminary model that combined multi-scale low-level feature responses, face detector and positional bias and got the state-of-the-art results on their newly-built Fixations on Webpage Image (FiWI) dataset. Compared with that model, we replace specific object detector with features from DNN that could encode general higher-level concepts and get even better results on FiWI dataset.

D. Fixations on Webpage Image Dataset

In this section, we describe the Fixations on Webpage Image (FiWI) Dataset which is used to validate saliency models in predicting webpage saliency.

E. Stimuli

A total of 149 screenshots of webpages rendered in Chrome browser in full screen mode were collected from various sources on the Internet in the resolution of 1360 by 768 pixels. These webpages were categorized as pictorial, textual and mixed according to the different composition of text and pictures. There are 50 pictorial images, 50 textual images and 49 mixed images in the dataset. Examples of webpage in each category are shown in Fig. 2 and the following criteria were used during the collection of webpage image samples.

- *Pictorial*: Webpages occupied by one dominant picture or several large thumbnail pictures and usually with less text. Examples in this category include photo sharing websites and company websites that put their products in the homepages.
- *Textual*: Webpages containing informative text with high density Examples include Wikipedia, news websites, and academic journal websites.
- *Mixed*: Webpages with a mix of thumbnail pictures and text in middle density. Examples are online shopping websites and social network sites.

The collected samples consisted of webpages from various domains. This was done to suppress the subjects' prior familiarity of the layout of the webpage as well as to prevent the subjects from developing familiarity during the experiment, so as to reduce personal bias or top-down factors.

F. Eye Tracking Data Collection

1) Subjects: A total of 11 students (4 males and 7 females) in the age range of 21 to 25 participated in data collection. All participants had normal vision or corrective visual apparatus during the experiment and all of them were experienced Internet users.

2) Apparatus and Eye Tracking: Subjects were seated in a dark room with their head positioned on a chin and forehead rest, 60 cm from the computer screen. The resolution of the screen was 1360×768 pixels. Stimuli were placed across the entire screen and were presented using MATLAB (MathWorks, Natick, Massachusetts, USA) with the Psychtoolbox 3 [37]. Eye movement data were monocularly recorded using a noninvasive

Fig. 2. Examples of webpage images in FiWI dataset. Left: pictorial webpages that are occupied by one dominant picture or several large thumbnail pictures. Middle: textual webpages containing informative text with high density. Right: webpages with a mix of thumbnail pictures and text in middle density.

Fig. 3. Fixation heat maps of the first, second, and third fixations over all the webpage images (first column) and the position distributions of the first, second, and third fixations on three example images from the dataset. (a) First fixation. (b) Second fixation. (c) Third fixation.

Eyelink 1000 system with a sampling rate of 1000 Hz. Calibration was done using the 9-point grid method.

3) Procedure: For each trial, an image was presented in random order for 5 seconds. Subjects were instructed to free-view the webpages and were informed that they had 5 seconds for each webpage. Each trial will follow by a drift correction where the subject would have to fixate at the center and initiate the next trial via a keyboard press.

G. Dataset Analysis

We analyze the eye fixation data collected from 11 subjects by visualizing their fixation heat maps. The fixation heat map was generated by convolving a 2D Gaussian filter on fixation points gathered from all the images in the dataset or in one particular category. In this work, a gaussian filter with a standard deviation of 25 pixels is used to smooth the fixation point and to generate a map. This size approximates the size of foveal region in human eye (1 visual degree approximates 50 pixels in our experimental setup).

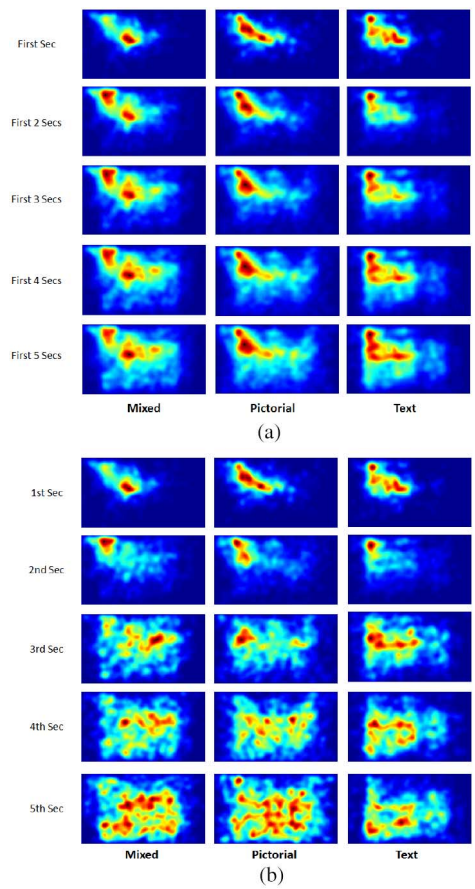


Fig. 4. Fixation heat maps on three categories of webpages during different time periods. (a) Accumulated fixation heat maps of fixations on three categories from the first second to the first five seconds. (b) Fixation heat maps on three categories with a second-by-second visualization.

Fig. 3 visualizes the distributions of the first three fixations on all the webpages and on three individual webpages. Fig. 4 illustrates category-wise fixation heat maps in first five seconds.

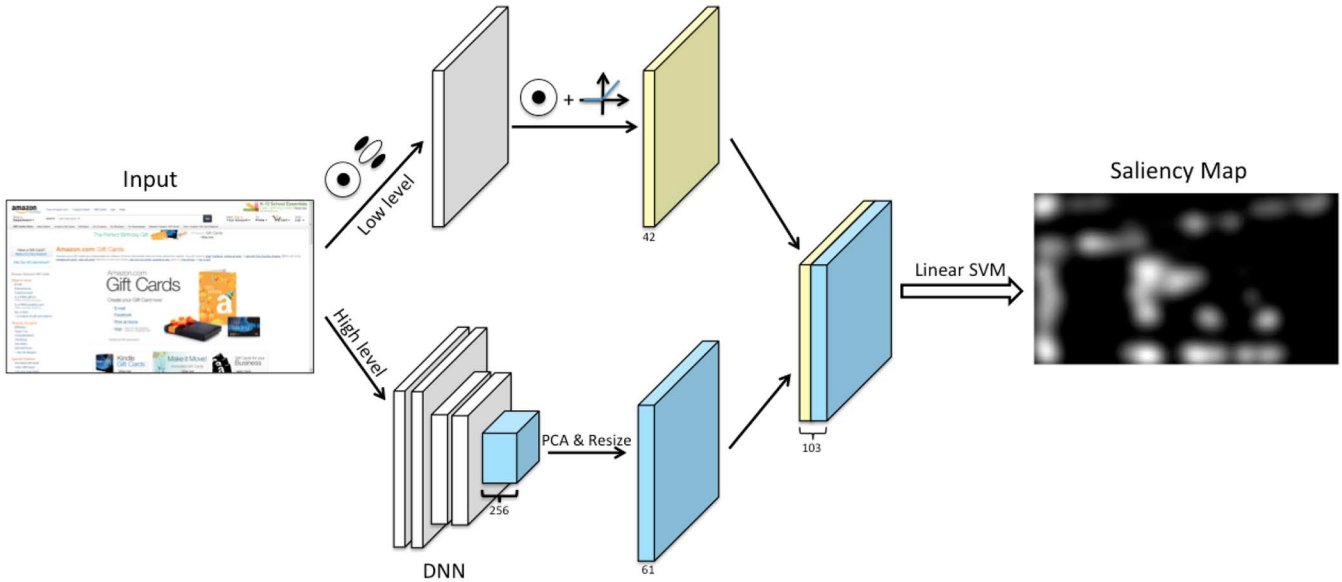


Fig. 5. Structure and the pipeline of our model that combines multi-scale low-level feature maps and representations from DNN to generate a saliency map.

From the figures we made the following observations: From the figures we made the following observations.

- *Positional Bias*: The positional bias on top-left region is evident in the visualization. From Fig. 3, we observe that most of the first, second and third fixations fall in this region. More specifically, the first fixations tend to locate in the center position that is slightly toward top-left corner and the second and third fixations usually fall on the trajectory from center to top-left corner. From Fig. 4, we can further observe that this top-left bias is common in all the three categories at first three seconds. These findings are in line with the F-shaped pattern described in [2], [14], and [16].
- *Object and Text Preference*: By looking into the eye fixation distributions on each individual webpage, we found that the first several fixations usually fall on large texts, logos, faces and objects that near the center or the top-left regions (Fig. 3, 2rd to 4th columns).
- *Category Difference*: From Fig. 4, we observe that, in all categories, fixations tend to cluster at the center and top-left region in the first two seconds and start to diversify after the 3rd second. Webpages from the 'Textual' category display a preference of the middle left and bottom left regions in 4th and 5th second while the fixations on the other two categories are more evenly distributed across all the locations.

III. The Saliency Model

The classical Itti-Koch saliency model [1], [38] computes multi-scale intensity, color, and orientation conspicuity maps from an image using multi-scale center-surround filters and Gabor filters and then combine these conspicuity maps into one saliency map after normalization. In our model, we further improve this low-level representations to a more compact and segregated one in DKL color space. In addition, we extract high-level features from a Deep Neural Network. We then combine low- and high-level features using a linear SVM trained on FiWI dataset. The structure of the model is shown in Fig. 5 and we describe it in detail below.

A. DKL Color Space and Low-Level Feature Extraction

In our implementation, we extract intensity, color and orientation features in different scales on the Derrington-Krauskopf-Lennie (DKL) color space [39] of a webpage image. The DKL color space is defined physiologically using the relative excitations of the three types of retinal cones (L, M, S, named after their sensitivity on light at long, medium and short wavelengths). The three channels of DKL color space, which are denoted as luminance (Lu, L+M), red and green opponency (RG, L-M) and yellow and blue opponency (BY, L+M-S), are orthogonal to each other [40].

For the computation of low-level feature maps, we convert the input image from RGB space to DKL space and rescale it to a six-level image pyramid. The scaling factor of neighboring levels in the parameter is 0.5. We then apply center-surround filters and Gabor filters with orientations of 0°, 45°, 90°, 135° on this image pyramid as below

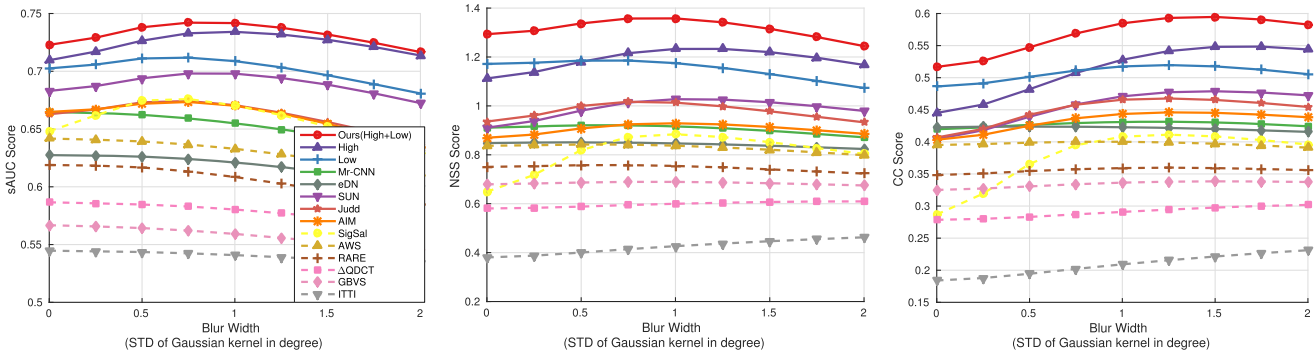
$$R_{(c,s)} = \|F * I_{(c,s)}\| \tag{1}$$

where F denotes the center-surround filters or Gabor filters, I_c denotes each channel c on scale s of the image pyramid, and R denotes the resulting feature response maps.

In this way, a total number of 42 multi-scale low-level feature maps are yielded on the 7 channels in 6 scales (The DKL color space generates 3 center-surround channels. 4 Gabor filters applied on the intensity maps result in 4 orientation channels).

B. Thresholded Center-Surround Filter

For the feature maps generated by center-surround and Gabor filters mentioned above, we observe that there exists edge artifacts which is false alarms around the boundary of an object. The edge artifacts are mainly caused by one single layer of



(a) (b) (c)

Fig. 6. Performance of different saliency models on the FiWI dataset under different similarity metrics and blurring width. Low indicates the model. High indicates the model purely based on high-level features.

center-surround or Gabor filters, especially for those of low spatial frequency (as illustrated in Fig. 5) In our model, in order to eliminate such edge artifacts, we further process the contrast and orientation representations by adding one more thresholded center-surround filter

$$M_{(c,s)} = \max(G * R_{(c,s)}, 0). \tag{2}$$

Here G is a center surround filter whose size is two times larger than that described in F and $M_{(c,s)}$ is the final conspicuity map to be integrated into a saliency map. After this operation, these false alarms would be largely inhibited and the responses on feature maps would be more concentrated on the center of the cluster (as illustrated in Fig. 5).

C. Representations From DNN

Webpages usually contain rich high-level features (e.g., faces, texts) that strongly attract attention. In this work, we replace specific object detectors with features from a DNN that could represent general high-level concepts. Since the current size of eye fixation datasets is too small to train a high-capacity DNN, we use features from AlexNet [30] trained on the large-scale ImageNet dataset [41], which has shown excellent generalization ability in many other visual tasks [42], [43]. To utilize AlexNet in our specific task, we did two modifications, as follows.

1) We remove all fully-connected layers in the original network. With fully-connected layers, the network receives fix-sized input image and produce a one-dimensional feature vector. Removing them allows the network to receive arbitrary-sized images and produce spatially dense feature maps of corresponding sizes once at a time. This is equivalent to applying the network in a sliding-window fashion, but is much more efficient than a naive sliding window since computation is shared in overlapping regions during convolution. The 256 top-layer feature maps from the AlexNet [30] are used in the next stage.
2) We apply Principle Component Analysis (PCA) to the 256 top-layer DNN features in order to reduce the dimensionality of high-level features since we observe high correlation between the 256 DNN features. The large number of correlated high-level features may overshadow other low-level features during SVM training. In addition, considering our limited sample size, high dimensionality of features may lead to overfitting. We retain 80% of variance after PCA, resulting in 61 uncorrelated high-level features. The 61 high-level features are then concatenated with 42 low-level features, to yield a final feature vector of length 103 for each pixel.

TABLE I
PERFORMANCE OF DIFFERENT MODELS ON THE WHOLE *FIWI* DATASET WITH OPTIMAL BLURRING

	sAUC	NSS	CC
Ours (Low+High)	0.7421	1.3578	0.5947
High	0.7341	1.2331	0.5484
Low	0.7118	1.1855	0.5197
SUN [24]	0.6981	1.0274	0.4789
Judd [5]	0.6742	1.0168	0.4676
AIM [26]	0.6734	0.9290	0.4460
SigSal [45]	0.6759	0.8840	0.4112
MrCNN [46]	0.6648	0.9207	0.4313
AWS [47]	0.6417	0.8421	0.4002
eDN [31]	0.6274	0.8507	0.4238
RARE [48]	0.6189	0.7580	0.3598
ΔQDCT [49]	0.5866	0.6095	0.3018
GBVS [50]	0.5667	0.6893	0.3385
Itti [1]	0.5447	0.4626	0.2312

D. Feature Integration

For saliency map generation, Support Vector Machine (SVM) is used to learn the weights to integrate feature maps into a saliency map. The response vectors extracted from training set are used as training samples for SVM, each labeled as positive (salient) or negative (non-salient). The hyperplane learned by a linear SVM is represented by weights w and bias b of the hyperplane, and all the feature maps M are combined into one saliency map S as

$$S = g * \max\left(\sum_{c,s} w_{(c,s)} M_{(c,s)} + b, 0\right) \tag{3}$$

where g is a Gaussian mask to smooth the saliency map. We apply rectified linear operation after integration to inhibit the noise generated in non-salient regions.

IV. EXPERIMENTS AND RESULTS

This section reports experimental results to validate the performance of our model on webpage saliency. We first train a

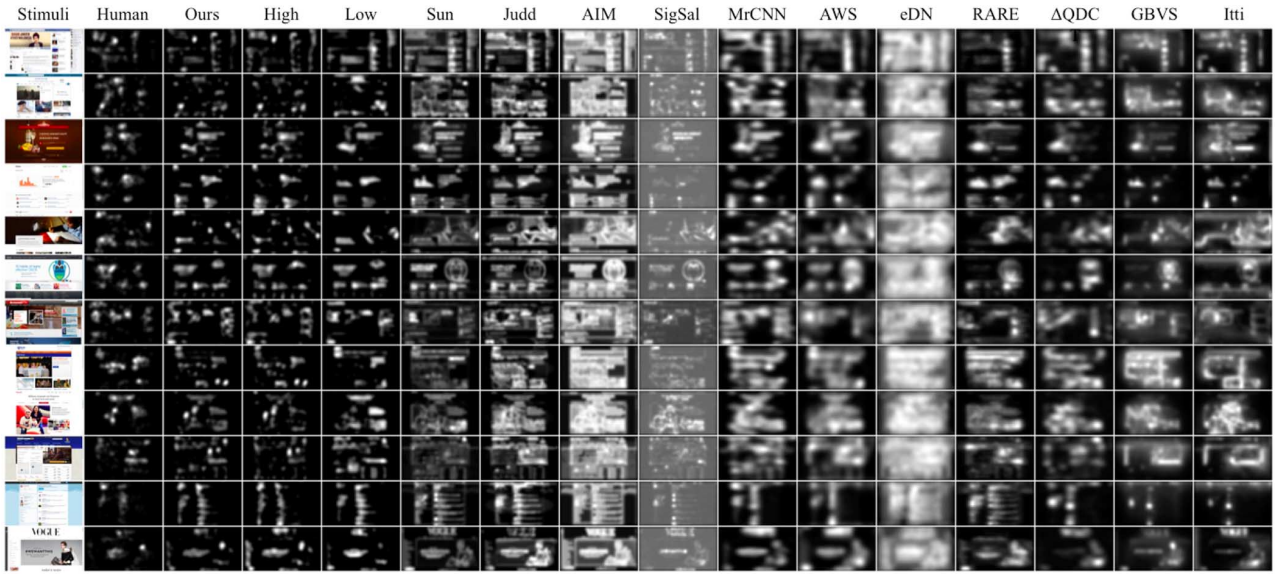


Fig. 7. Qualitative comparisons of the proposed models and other saliency models on the FiWI dataset.

SVM model to integrate both low- and high-level features. We then compare the proposed model with other saliency detection algorithms on the whole webpage saliency dataset as well as images in each category. We then discuss and analyze our results.

A. Training and Validation on Webpage Saliency Dataset

We trained and validated our model on the Fixations in Webpage Images (FiWI) dataset [18]. The FiWI dataset contains 149 webpage screenshots (1360 by 768 pixels) with eye movement data from 11 observers during free-viewing. 10-fold cross validation over the dataset was carried out. In each trial, we trained a SVM with 134 images (135 in the last time). We collected positive samples and negative samples from the training images in the dataset. For each image, we randomly extracted 10 positively labeled feature vectors from top 20% salient regions and 10 negatively labeled feature vectors from bottom 50% salient regions to yield a training set of 2680 training samples. We then generated saliency maps for remaining images with the learned parameters.

B. Saliency Evaluation Metrics

The saliency evaluation metrics we use include shuffled Area Under Curve (sAUC), linear Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS) whose codes and descriptions are all available online [11], [44].

AUC is the most widely used score for saliency model evaluation. In the computation of AUC, the estimated saliency map is used as a binary classifier to separate the positive samples (human fixations) from the negatives (random points). By varying the threshold on the saliency map, a Receiver Operating Characteristics (ROC) curve can then be plotted as the true positive rate vs. false negative rate. AUC is then calculated as the area under this curve. For the AUC score, 1 means perfect prediction while 0.5 indicates chance level. However, AUC can easily be influenced by center-bias, which makes a fair model comparison difficult. *sAUC* (shuffled AUC) is the same as AUC except using fixations of other images in the same dataset as negatives and can eliminate the effect of center-bias.

CC measures the linear correlations between the estimated saliency map and the ground truth fixation map. The closer CC to 1, the better the performance of the saliency algorithm.

NSS measures the average of the response values at fixation locations along the scanpath in the normalized saliency map. The larger the NSS score, the more corresponding between predictions and ground truths.

All these metrics have their advantages and limitations and a model that performs well should have relatively high score in all these metrics.

C. Performance

To measure the performance of our model, we compare it with other nine state-of-the-art saliency models, including Judd [5], AWS [23], Δ QDCT [49], RARE [48], AIM [26], GBVS [50], SUN [24], Image Signature [45], Itti [1], eDN [31], and MrCNN [46]. For Judd's Model [5], we retrain it on the webpage dataset using the exactly same training paradigm as our model to ensure a fair comparison. For MrCNN [46], we request the maps from the authors. For other models we use the default parameters provided by the authors. In addition, we train two sub-models using the same settings except that the sub-models only use a subset of features. The Low model only uses 42 low-level features based on color contrast and orientation. The High model only uses 61 high-level features from DNN. Since blurring can also significantly affect model performance, we smoothed the saliency maps of each model using a Gaussian kernel with different standard deviation from 0 to 2 Degree of Visual Angle (DVA). One DVA corresponds to approximately 20 image pixels. The effect of blurring on performance is illustrated in Fig. 6. We also list the final evaluation scores, which are obtained as the highest scores with optimal blurring, in Table I. Results show that all our three models outperform other saliency models under all the three evaluation metrics. The High model performs better than the Low model, possibly due to the rich semantic contents contained in webpages. Our final model, combining both low-level

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON EACH IMAGE CATEGORY WITH OPTIMAL BLURRING

	Pictorial			Textual			Mixed		
	sAUC	NSS	CC	sAUC	NSS	CC	sAUC	NSS	CC
Ours (Low+High)	0.7522	1.4374	0.6173	0.7293	1.2887	0.5563	0.7458	1.3527	0.6108
High	0.7484	1.3096	0.5705	0.7127	1.1134	0.4883	0.7423	1.2793	0.5878
Low	0.7215	1.2946	0.5577	0.7039	1.1339	0.4911	0.7115	1.1321	0.5100
SUN [24]	0.7074	1.0715	0.4875	0.6870	0.9932	0.4590	0.7001	1.0173	0.4903
Judd [5]	0.6670	0.9765	0.4684	0.6812	1.0395	0.4728	0.6745	1.0349	0.4615
AIM [26]	0.6628	0.9070	0.4535	0.6796	0.9366	0.4452	0.6780	0.9436	0.4391
SigSal [45]	0.6640	0.8475	0.4148	0.6818	0.8849	0.4083	0.6822	0.9203	0.4106
MrCNN [46]	0.6896	1.0098	0.4624	0.6290	0.8197	0.3802	0.6761	0.9343	0.4524
AWS [47]	0.6329	0.8059	0.4010	0.6478	0.8669	0.4061	0.6446	0.8540	0.3933
eDN [31]	0.6404	0.8825	0.4269	0.6123	0.8335	0.4128	0.6296	0.8383	0.4334
RARE [48]	0.6300	0.7827	0.3590	0.5974	0.6887	0.3303	0.6294	0.8046	0.3930
ΔQDCT [49]	0.6065	0.6557	0.3110	0.5519	0.5144	0.2529	0.6017	0.6656	0.3428
GBVS [50]	0.5637	0.6850	0.3508	0.5614	0.6808	0.3291	0.5753	0.7079	0.3355
Itti [1]	0.5384	0.4427	0.2333	0.5437	0.4829	0.2337	0.5521	0.4622	0.2265

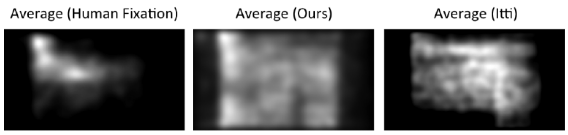


Fig. 8. Average of human fixation maps and saliency maps from our model and Itti's classical model [1] across the whole FiWI dataset. It can be observed that the our model displays top-left bias in average map.

TABLE III
PERFORMANCE UNDER DIFFERENT AMOUNT OF VARIANCE RETAINED

	sAUC	NSS	CC
70%(36)	0.7247	1.1731	0.5236
80%(61)	0.7421	1.3578	0.5947
90%(103)	0.7381	1.3399	0.5861
100%(256)	0.7166	1.0990	0.4986

TABLE IV
PERFORMANCE UNDER DIFFERENT EARLIER LAYERS INCOPERATED

	sAUC	NSS	CC
conv1-5	0.6912	1.0293	0.4518
conv2-5	0.6854	1.0242	0.4475
conv3-5	0.6780	0.9596	0.4245
conv4-5	0.7020	1.1288	0.4850
conv5	0.7166	1.0990	0.4986

TABLE V
PERFORMANCE OF DIFFERENT MODELS ON THE MIT1003 DATASET

Model	sAUC	NSS	CC
Ours	0.7193	0.5134	1.4445
MrCNN	0.7096	0.4992	1.3053
eDN	0.6749	0.4579	1.0628
SUN	0.6496	0.4319	0.9922
Judd	0.6651	0.4563	1.0952
AIM	0.6804	0.4691	1.0824
SigSal	0.6659	0.4650	1.0848
AWS	0.6858	0.4452	1.1073
RARE	0.6769	0.5015	1.2846
ΔQDCT	0.6561	0.4296	1.0524
GBVS	0.6434	0.5024	1.2540
Itti	0.6446	0.4679	1.1267

and high-level features, performs the best among all compared models, suggesting the importance of both low- and high-level features. Fig. 6. demonstrates that our final model consistently achieve the best performance under almost all blurring factors. Although Judd's model uses the same supervised training procedure, it performs much less well, indicating the superiority of our multi-level features in terms of predicting fixations on webpage. Among other saliency models, SUN model [24] performs the best over all metrics. In Fig. 7 we show some typical images and their corresponding fixation maps and saliency maps generated by different algorithms. Our model can accurately predict human fixations and is more selective than compared models.

In Table II we compare model performance in each image category. It can be seen that all our models perform better over pictorial webpages than textual webpages. Further, we found that SUN, eDN, RARE and Δ QDCT perform better over pictorial webpages while Judd, SigSal, AIM, AWS performs better over textual webpages. MrCNN performs very well on pictorial webpages while its performance on textual webpages is not very good. For GBVS and ITTI, the overall performance over the two categories is similar.

D. Results Analysis

We then analyzed our results to see whether there exists top-left bias and banner blindness in our saliency maps.

Top-Left Bias: To investigate into the top-left bias existing in the predicted saliency maps, we compute the average of all the saliency maps across the whole FiWI dataset. The scores of using average human fixation to predict webpage saliency across the whole dataset are sAUC: 0.4994, NSS: 0.9220, and CC:0.4578. We also compute the average map for human eye fixation and the results of Itti's model across the dataset. In Fig. 8, these three average maps are normalized and illustrated. It can be seen that the average map of our algorithm does display some extent of top-left bias. A further calculation of correlation between average maps of human fixation and our model yield a correlation score of 0.82, while the correlation score between average maps of human fixation and Itti's model is 0.69, which fits the qualitative illustration well.

Banner Blindness: To study whether there is banner blindness in our model, we selected webpage images that contain large banners and visualize their fixation maps and saliency maps. From Fig. 9, it can be seen that the responses of our saliency maps have no or small responses in banner regions, while the saliency maps of Itti's model usually have large responses on this region. This qualitative difference may result from two reasons: 1) The concept of banner is implicitly represented in our features; 2) with supervised training on webpage eye tracking dataset, our model learns a small or negative weights for features representing the banner.

TABLE VI
RUNNING TIME FOR ALL THE MODELS

Ours	SigSal	ΔQDCT	AWS	Itti	RARE	GBVS	SUN	eDN	MrCNN	Judd	AIM
0.53	0.04	0.11	0.26	0.44	0.53	0.65	4.70	10.76	14	35.40	88.66

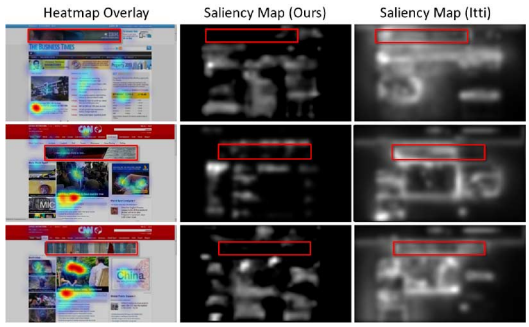


Fig. 9. Banner-like patterns from selected images in FiWI dataset and banner blindness on human fixation map, our saliency map, and Itti's saliency map [1]. Banner regions are highlighted in a red bounding box and it can be observed that our saliency maps have no or small responses in banner regions.

E. Model Structure Analysis

We further analyze the influence of model structure on the final performance. the optimal parameter is obtained by training models with different amount of variance kept and comparing the scores of each candidate model (as illustrated in Table III). It can be seen that 80% of variance achieve the best performance.

Besides, we also incorporate different amounts of earlier layers and the results are illustrated in Table IV. It can be seen that the use of final convoution layer from AlexNet for high-level feature extraction is the most optimal one. For a fair comparison, all the candidate models here are trained without PCA.

F. Performance on Traditional Fixation Dataset

To investigate whether our model still work well on traditional dataset, we compare the performance of our models and all the other state-of-the-art models on the MIT1003 dataset. The result is presented in Table V. It can be seen that our model is still competitive over other models on traditional eye fixation dataset.

G. Running Time Analysis

Finally, we then analyze the running time for each model compared and present them in Table VI. Their running time are got mainly by running their code on our machine except that the statistics of MrCNN is extracted from their paper (Section III-C of [46]).

V. CONCLUSION

Despite the large amount of existing saliency models that predict where humans look at in natural images, there are few studies on saliency in webpages. Considering the important role webpages play in our daily life and the significantly different human viewing patterns between webpages and natural images, a model that can accurately predict saliency in webpages is of high commercial and research value. In this work, we integrate multi-level representations to predict saliency on webpage. Traditional low-level features (color, intensity and orientations), as well as high-level features from deep neural networks, are integrated using a linear SVM to construct the saliency map. Experiments show that our model outperforms existing saliency models by a large margin in predicting webpage saliency.

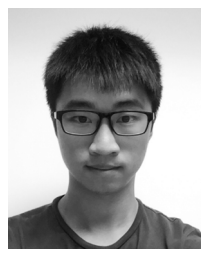
REFERENCES

[1] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, no. 10–12, pp. 1489–1506, 2000.
[2] G. Buscher, E. Cutrell, and M. R. Morris, "What do you see when you're surfing?: Using eye tracking to predict salient regions of web pages," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2009, pp. 21–30.
[3] L. Zhaoping, "Attention capture by eye of origin singletons even without awareness? A hallmark of a bottom-up saliency map in the primary visual cortex," *J. Vision*, vol. 8, no. 5, p. 1, 2008.
[4] X. Zhang, L. Zhaoping, T. Zhou, and F. Fang, "Neural activities in V1 create a bottom-up saliency map," *Neuron*, vol. 73, no. 1, pp. 183–192, 2012.
[5] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Computer Vis.*, Sep.–Oct. 2009, pp. 2106–2113.
[6] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Adv. Neural Inf. Process. Syst.*, vol. 20, pp. 241–248, 2008.
[7] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vision*, vol. 8, no. 14, 2008.
[8] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vision*, vol. 11, no. 3, 2011.
[9] C. Shen, M. Song, and Q. Zhao, "Learning high-level concepts by training a deep network on eye fixations," in *Proc. NIPS Deep Learn. Unsupervised Feature Learn. Workshop*, 2012, vol. 2.
[10] C. Shen and Q. Zhao, "Learning to predict eye fixations for semantic contents using multi-layer sparse network," *Neurocomput.*, vol. 138, pp. 61–68, 2014.
[11] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 438–445.
[12] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate it cortex for core visual object recognition," *PLoS Comput. Biol.*, vol. 10, no. 12, p. e1003963, 2014.
[13] J. D. Still and C. M. Masciocchi, "A saliency model predicts fixations in web interfaces," in *Proc. 5th Int. Workshop Model Driven Develop. Adv. User Interfaces*, 2010, pp. 25–28.
[14] J. Nielsen, Nielsen Norman Group. Fremont, CA, USA, "F-shaped pattern for reading web content," Apr. 2006 [Online]. Available: http://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/, Accessed on: Sep. 1, 2015
[15] C.-H. Cho and H. J. Cheon, "Why do people avoid advertising on the Internet?," *J. Advertising*, vol. 33, no. 4, pp. 89–97, 2004.
[16] R. Grier, P. Kortum, and J. Miller, "How users view web pages: An exploration of cognitive and perceptual mechanisms," in *Human Computer Interaction Research in Web Design and Evaluation*. Hershey, PA, USA: IGI Global, 2007, pp. 22–41.
[17] G. Hervet, K. Guérard, S. Tremblay, and M. S. Chtourou, "Is banner blindness genuine? Eye tracking Internet text advertising," *Appl. Cognitive Psychol.*, vol. 25, no. 5, pp. 708–716, 2011.
[18] C. Shen and Q. Zhao, "Webpage saliency," in *Proc. ECCV*, 2014, pp. 33–46.
[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[20] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–27, 1985.
[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[22] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, 2009.
[23] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vis. Comput.*, vol. 30, no. 1, pp. 51–64, 2012.
[24] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *J. Vision*, vol. 8, no. 7, 2008.
[25] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Adv. Neural Inf. Process. Syst.*, vol. 21, pp. 681–688, 2008.
[26] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vision*, vol. 9, no. 3, 2009.
[27] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, Jun. 2012, pp. 478–485.
[28] M. Liang and X. Hu, "Predicting eye fixations with higher-level visual features," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1178–1189, Mar. 2015.
[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1701–1708.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.
[31] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2798–2805.
[32] P. Faraday, "Visually critiquing web pages," in *Multimedia '89*. New York, NY, USA: Springer, 2000, pp. 155–166.
[33] G. B. Duggan and S. J. Payne, "Skim reading by satisficing: Evidence from eye tracking," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2011, pp. 1141–1150.
[34] J. Nielsen, Nielsen Norman Group. Fremont, CA, USA, "Banner blindness: Old and new findings," Aug. 2007 [Online]. Available: http://www.nngroup.com/articles/banner-blindness-old-and-new-findings/, Accessed on: Sep. 1, 2015
[35] E. Cutrell and Z. Guan, "What are you looking for? An eye-tracking study of information usage in web search," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 407–416.
[36] B. Stone and S. Dennis, "Using LSA semantic fields to predict eye movement on web pages," in *Proc. 29th Cogn. Sci. Soc. Conf.*, 2007, pp. 665–670.
[37] D. H. Brainard, "The psychophysics toolbox," *Spatial Vis.*, vol. 10, no. 4, pp. 433–436, 1997.
[38] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.
[39] A. M. Derrington, J. Krauskopf, and P. Lennie, "Chromatic mechanisms in lateral geniculate nucleus of macaque," *J. Physiol.*, vol. 357, no. 1, pp. 241–265, 1984.
[40] H.-P. Frey, C. Honey, and P. König, "What's color got to do with it? the influence of color on visual attention in different categories," *J. Vision*, vol. 8, no. 14, p. 6, 2008.
[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 248–255.
[42] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," *CoRR*, 2013 [Online]. Available: http://arxiv.org/abs/1310.1531
[43] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2014, pp. 512–519.
[44] Z. Bylinskii, T. Judd, F. Durand, A. Oliva, and A. Torralba, "MIT saliency benchmark," Jan. 2012 [Online]. Available: http://saliency.mit.edu/, Accessed on: Sep. 1, 2015
[45] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.
[46] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 362–370.
[47] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *J. Vision*, vol. 12, no. 6, p. 17, 2012.
[48] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Process.: Image Commun.*, vol. 28, no. 6, pp. 642–658, 2013.
[49] B. Schauerte and R. Stiefelhagen, "Predicting human gaze using quaternion DCT image signature saliency and face detection," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2012, pp. 137–144.
[50] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 545–552, 2007.

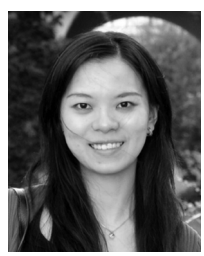
Chengyao Shen received the B.S. degree in microelectronics from Shanghai Jiaotong University, Shanghai, China, in 2010, and is currently working toward the Ph.D. degree at the National University of Singapore, Singapore.

His research interests included computer vision, machine learning, and natural image statistics.



Xun Huang is currently working toward the B.S. degree in computer science at Beihang University, Beijing, China.

He is currently a Visiting Scholar with the National University of Singapore, Singapore. His research interests include deep learning, computer vision, and cognitive science.



Qi Zhao (S'04–M'09) received the M.Sc. and Ph.D. degrees in computer engineering from the University of California at Santa Cruz, Santa Cruz, CA, USA, in 2007 and 2009, respectively.

From 2009 to 2011, she was a Postdoctoral Researcher with the Computation & Neural Systems Group and the Division of Biology, California Institute of Technology, Pasadena, CA, USA. She is currently an Assistant Professor with the Electrical and Computer Engineering Department, National University of Singapore (NUS), Singapore, and the Principal Investigator with the Visual Information Processing Lab, NUS. She also holds an appointment with the Ophthalmology Department and the Interactive and the Digital Media Institute, NUS. She has authored or coauthored more than 30 journal and conference papers, and is currently editing the book *Computational and Cognitive Neuroscience of Vision* (Springer). Her main research interests include computational vision, machine learning, computational cognition, and neuroscience.