Predicting human gaze beyond pixels

Juan Xu Department of Electrical and Computer Engineering, National University of Singapore, Singapore

Ming Jiang Department of Electrical and Computer Engineering, National University of Singapore, Singapore

Shuo Wang Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA

Mohan S. Kankanhalli Department of Computer Science, School of Computing, National University of Singapore, Singapore

Qi Zhao Department of Electrical and Computer Engineering, National University of Singapore, Singapore

A large body of previous models to predict where people look in natural scenes focused on pixel-level image attributes. To bridge the semantic gap between the predictive power of computational saliency models and human behavior, we propose a new saliency architecture that incorporates information at three layers: pixel-level image attributes, object-level attributes, and semantic-level attributes. Object- and semantic-level information is frequently ignored, or only a few sample object categories are discussed where scaling to a large number of object categories is not feasible nor neurally plausible. To address this problem, this work constructs a principled vocabulary of basic attributes to describe object- and semantic-level information thus not restricting to a limited number of object categories. We build a new dataset of 700 images with eye-tracking data of 15 viewers and annotation data of 5,551 segmented objects with fine contours and 12 semantic attributes (publicly available with the paper). Experimental results demonstrate the importance of the object- and semantic-level information in the prediction of visual attention.

Introduction

Humans and other primates have a tremendous ability to rapidly direct their gaze when looking into a static or dynamic scene and to select visual information of interest. This ability enables them to deploy limited processing resources to the most relevant visual information and understand real-world scenes rapidly and accurately. Understanding and simulating this mechanism has both scientific and economic impact (Koch & Ullman, 1985; Ungerleider, 2000; Treue, 2001). A computational model predicting where humans look has general applicability in a wide range of tasks relating to human-robot interaction, surveillance, advertising, marketing, entertainment, and so on. One common approach is to take inspirations from the functionality of human visual system (Milanese, 1993; Tsotsos et al., 1995; Itti, Koch, & Niebur, 1998; Rosenholtz, 1999), while some other studies claim that visual attention is attracted to the most informative regions (Bruce & Tsotsos, 2009), the most surprising regions (Itti & Baldi, 2006), or those regions that maximize reward regarding a task (Sprague & Ballard, 2003). Existing works on saliency modeling mainly focus on pixel-level image attributes, such as contrast (Reinagel & Zador, 1999), edge content (Baddeley & Tatler, 2006), orientation (Itti et al., 1998), intensity bispectra (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000), and color (Itti et al., 1998; Jost, Ouerhani, von Wartburg, Muri, & Hugli, 2005; Engmann et al., 2009), despite various recent developments on inference (Raj, Geisler, Frazor, & Bovik, 2005; Walther, Serre, Poggio, & Koch, 2005; Gao, Mahadevan, & Vasconcelos, 2007; Harel, Koch, & Perona, 2007; Bruce & Tsotsos, 2009; Seo & Milanfar, 2009; Carbone & Pirri, 2010; Chikkerur, Serre, Tan, &

Citation: Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, *14*(1):28, 1–20, http://www.journalofvision.org/content/14/1/28, doi:10.1167/14.1.28.

doi: 10.1167/14.1.28 Received December 04, 2012; published January 28, 2014 ISSN 1534-7362

Poggio, 2010; Wang, Wang, Huang, & Gao, 2010; Hou, Harel, & Koch, 2012) to generate a saliency map.

The extent to which such bottom-up, task-independent saliency models predict fixations of free-viewers remains an active topic (Donk & van Zoest, 2008; Foulsham & Underwood, 2008; Masciocchi, Mihalas, Parkhurst, & Niebur, 2009). A more recent problem in the saliency community is the semantic gap between the predictive power of computational saliency models and human behavior. That is, pixel-level image attributes fail to encode object and/or semantic information, which is many times more important to saliency than pixel-level information. To fill the semantic gap, Krieger et al. (2000) and Einhäuser et al. (2006) suggested the incorporation of higher order statistics. Recent neurophysiological studies (Craft, Schütze, Niebur, & Von Der Heydt, 2007; Mihalas, Dong, Von Der Heydt, & Niebur, 2010) suggest that primates use a more powerful representation in which raw sensory input is perceptually grouped by dedicated neuronal circuitry. Psychophysical experiments (Einhäuser, Spain, & Perona, 2008; Nuthmann & Henderson, 2010; Foulsham & Kingstone, 2013) show that humans frequently allocate their gaze to interesting objects in a scene, and a large portion of fixations are close to the center of objects. At the object level, Gestalt psychologists have found many perceptual organization rules like convexity, surroundedness, orientation, symmetry, parallelism, and object familiarity (Palmer, 1999) that are known to play important roles in determining what we see. Before Itti et al.'s (1998) framework, Reisfeld, Wolfson, and Yeshurun (1995) already proposed a symmetry operator to guide attention. Recently, a simple bottom-up assignment model proposed by Fowlkes, Martin, and Malik (2007) suggested that a smaller, more convex, or lower region is more likely to encode midlevel (object-level) visual cues by constructing prototypical local shapes from image data. These object-level attributes have not yet been studied systematically as to how they relate to saliency, and we aim to explore their relationships in a more principled way.

On top of the object-level information that attracts attention, semantic information also contributes much to the final saliency: For example, a face tends to attract attention more than other objects (Cerf, Frady, & Koch, 2009). It is also known that survival-related attributes (e.g., food, sex, danger, pleasure, and pain) possess an innate saliency that is determined by the activity of evolutionarily selected value systems in the brain (Edelman, 1987; Friston et al., 1994). Recently several works (Cerf et al., 2009; Judd, Ehinger, Durand, & Torralba, 2009; Zhao & Koch, 2011, 2012) have added important object categories into their saliency models to improve the prediction of attentional selection. While these models consistently show improved performance, they do not scale well to many object categories in the real world, as each object requires a particular detector. Further, it is arguable that our brain is domain-specific for object processing. Thus, having an object detector for each individual possible object is not neurally plausible either. Yet is there anything (a base attribute) inherent about the object categories that make them salient? This question is largely unknown, and in this work we aim to make a first step toward this exploration. To approach this problem, we propose an attribute-based framework where each attribute captures inherent object- or semantic-level information that is important to saliency, and the combination of a limited set of attributes is able to describe a much larger set of object categories—in theory an infinite number of categories. This work is motivated to better understand how various factors contribute to saliency, e.g., what attributes are more important and how are they combined to fill the semantic gap.

In this work we propose a new three-layered architecture for saliency prediction. While most existing saliency models focus on pixel-level attributes, object- and semantic-level information has shown to be even more important than pixel-level attributes. We explicitly and principally introduce a framework that integrates object and semantic information for saliency. Instead of focusing on a few sample object categories that are difficult to scale well, this work presents a set of common attributes at object- and semantic-level to form a vocabulary that is capable of describing a much larger set of objects as well as their semantic meanings. We also analyze the relevant importance of each attribute to saliency. We construct a large eye-tracking dataset with (a) 700 images with (semantic) objects (a large portion have multiple dominant objects in the same image), (b) eye-tracking data with 15 viewers, (c) 5,551 segmented objects with fine contours, and (d) annotations of semantic attributes on all the objects.

Attributes for pixel-, object-, and semantic-levels

To accurately predict human gaze, higher-level information is important (see Figure 1). Particularly, we aim to construct a vocabulary, i.e., a relatively complete set of attributes wherein (a) each is inherent in predicting saliency, and (b) combining them covers a much larger set of object categories, as well as their semantic attributes so that the approach scales well.

Pixel-level attributes

Pixel-level image attributes, such as contrast (Reinagel & Zador, 1999), edge content (Baddeley &

Figure 1. Human fixations attracted by object-level and semantic-level attributes. The leftmost images of simple objects show the effect that most fixation points are allocated near object centers. The four columns of images to the right show that various types of semantic cues (taste, face, text, and gaze) have consistently high fixation density.

Tatler, 2006), intensity bispectra (Krieger et al., 2000), and color (Jost et al., 2005) have been well researched in saliency literature. In our model we simply include three more commonly used biologically plausible attributes (i.e., color, intensity, and orientation; Itti et al., 1998) as pixel-level attributes.

Object-level attributes

Attributes at this level describe object properties that apply to all objects and that are independent of semantics (semantic parts of objects are modeled below with the semantic-level attributes). Based on psychophysical and neurophysiological evidence (Craft et al., 2007; Einhäuser et al., 2008; Mihalas et al., 2010; Nuthmann & Henderson, 2010; Foulsham & Kingstone, 2013), we hypothesize that any object, despite its semantic meanings, attracts attention more than nonobject regions.

Particularly, we introduce five attributes at this level that are simple and effective in predicting saliency: size, complexity, convexity, solidity, and eccentricity. Before the introduction of the object-level attributes, we first define several relevant notations for objects and the convex hull of the objects (see Figure 2). Particularly we denote an object as O, and the convex hull of an object as C. Thus the area and perimeter of an object are denoted as A_O and P_O, and the area and perimeter of the convex hull of an object are denoted as A_C and P_C.

Size

Size is an important object-level attribute, yet it is not clear how it affects saliency—whether large or small objects tend to attract attention. Generally, a larger object might have more attractive details, but will probably be ignored for being a background. This attribute is denoted as $\sqrt{A_O}$, where A_O represents the object's area.

Convexity

The convexity of an object is denoted as P_C / P_O, where P_C represents the perimeter of the object's convex hull, and P_O represents the perimeter of the object's outer contour. Thus, a convex object has a convexity value of 1.

Solidity

The solidity attribute is intuitively similar to convexity, but it also measures holes in objects. Formally, solidity is denoted as A_O / A_C where A_O and A_C are the areas of the object and its convex hull, respectively. If an object is convex and without holes in it, it has a solidity value of 1.

Complexity

Complexity is denoted as $P_O/\sqrt{A_O}$. With the area of the object fixed, the complexity is higher if the contour is longer. A circle has minimum complexity.

Eccentricity

Eccentricity is represented by the eccentricity value of an ellipse that has the same second-moments as the

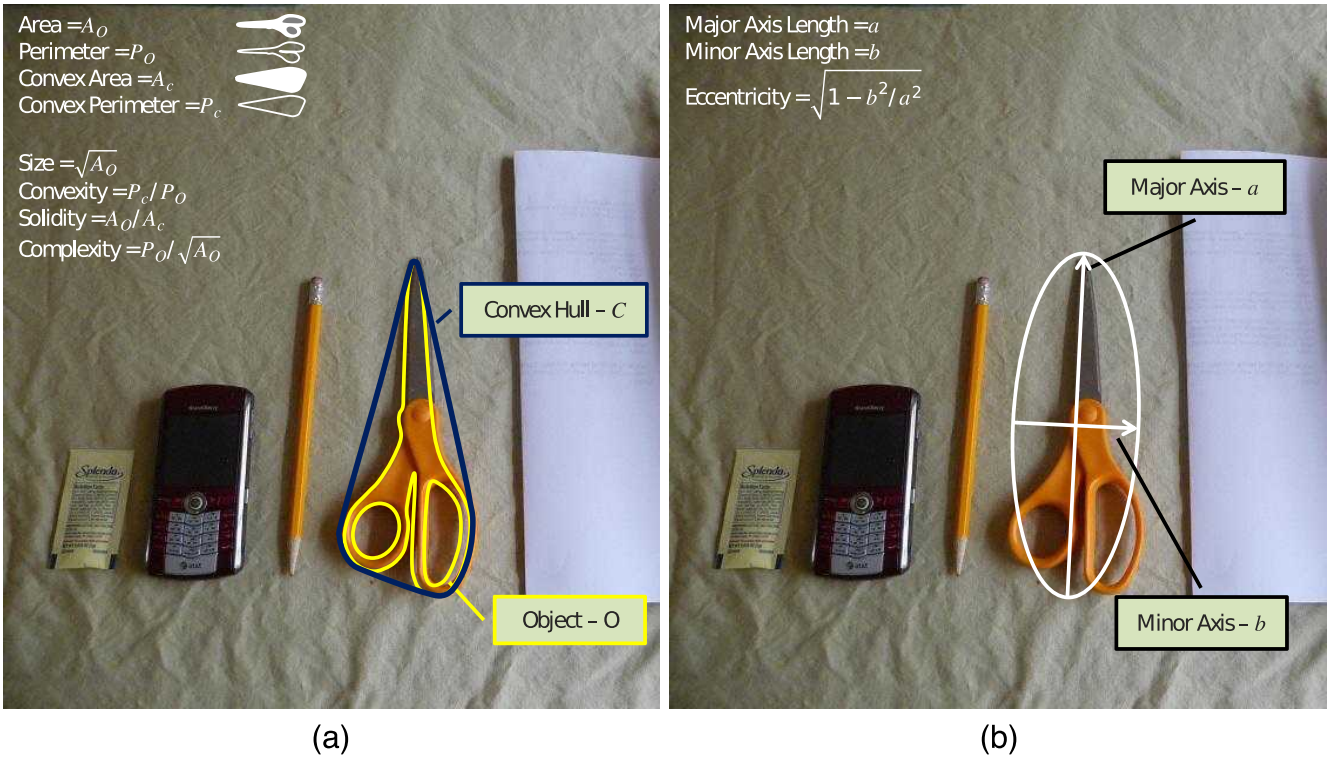

Figure 2. Illustration of object-level attributes: (a) size, convexity, solidity, complexity, and (b) eccentricity.

object region. An ellipse whose eccentricity is 0 is a circle, while an ellipse whose eccentricity is 1 is a line segment.

Semantic-level attributes

On top of the object-level attributes, humans tend to allocate attention to important semantic entities. At this semantic-level, we aim to characterize semantic information relating to saliency. It is generally accepted that "given the limited size of the human brain, it is unreasonable to expect that every one of semantic categories is represented in a distinct brain area" (Huth, Nishimoto, Vu, & Gallant, 2012). Thus to approach the problem of scalability in both the brain and in computational models, we define attributes where each of them characterizes certain inherent semantic properties and combines to describe a large class of object categories. Many cognitive psychological, neuropsychological, and computational approaches (Garrard, Ralph, Hodges, & Patterson, 2001; Cree & McRae, 2003; Farhadi, Endres, Hoiem, & Forsyth, 2009) have been proposed to organize semantic concepts in terms of their fine-grained attributes. Inspired by these works, we have constructed a semantic vocabulary that broadly covers the following four categories:

1. Directly relating to humans (i.e., face, emotion, touched, gazed). Humans and primates have dedicated systems to process faces that are represented in the fusiform face areas in humans (Kanwisher, McDermott, & Chun, 1997; Kanwisher & Yovel, 2006) and in face patches in primates (Tsao, Freiwald, Tootell, & Livingstone, 2006; Moeller, Freiwald, & Tsao, 2008). It has been demonstrated that visual attention is preferentially oriented to faces (Vuilleumier, 2000; Ro, Russell, & Lavie, 2001; Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005; Bindemann, Burton, Langton, Schweinberger, & Doherty, 2007; Cerf et al., 2009). Emotion is central to the quality and range of everyday human experience. The neurobiological substrates of human emotion are described in Dolan (2002). In particular, the human amygdala clearly contributes to processing emotionally salient and socially relevant stimuli (Kling & Brothers, 1992; Adolphs, 2010). Eyes and gazes are socially salient (Argyle, Ingham, Alkema, & McCallin, 1973; Whalen et al., 2004), and they trigger reflexive orientation of attention (Friesen & Kingstone, 1998). Gaze directions are represented in superior temporal sulcus (STS; Hoffman & Haxby, 2000; Pelphrey, Viola, & McCarthy, 2004), and Hooker et al. (2003) showed a brain network to analyze eye gaze. Tactile touch has social significance and attracts attention. The impact and neural substrates of the social touch have also been shown (Schirmer et al., 2011).
2. Objects with implied motion in the image. A number of recent studies (Kourtzi & Kanwisher, 2000;

Name	Description
Face	Back, profile, and frontal faces.
Emotion	Faces with obvious emotions.
Touched	Objects touched by a human or animal in the scene.
Gazed	Objects gazed upon by a human or animal in the scene.
Motion	Moving/flying objects, including humans/animals with meaningful gestures.
Sound	Objects producing sound (e.g., a talking person, a musical instrument).
Smell	Objects with a scent (e.g., a flower, a fish, a glass of wine).
Taste	Food, drink, and anything that can be tasted.
Touch	Objects with a strong tactile feeling (e.g., a sharp knife, a fire, a soft pillow, a cold drink).
Text	Digits, letters, words, and sentences.
Watchability	Man-made objects designed to be watched (e.g., a picture, a display screen, a traffic sign).
Operability	Natural or man-made tools used by holding or touching with hands.

Table 1. Semantic-level attributes.

Lorteije et al., 2006; Winawer, Huk, & Boroditsky, 2008; Faivre & Koch, 2013) suggest that implied motion from static stimuli and physical motion may share the same direction-selective mechanisms. Hence, objects with implied motion may also attract visual attention.

3. Relating to other (nonvisual) senses of humans (i.e., sound, smell, taste, touch). Observing whether objects relating to nonvisual senses attract visual attention allows an analysis of other sensory perceptions of humans (Onat, Libertus, & König, 2007). For example, sound, especially when sound gets emotional, elicits social orientation and activates the amygdala (Schirmer et al., 2008).
4. Designed to attract attention or for interaction with humans (i.e., text, watchability, operability). Operability is defined on tools and several reports have shown an increased response to tools in the middle temporal gyrus (MTG; Chao, Haxby, & Martin, 1999; Beauchamp, Lee, Haxby, & Martin, 2003). Text has been demonstrated to attract attention (Cerf et al., 2009), and other objects designed for people to watch potentially have similar properties. Therefore, it is of interest to explore how these attributes attract attention.

For each attribute, each object is either scored 1 to address the existence of the corresponding attribute or 0 to represent the absence of the attribute. In Table 1 we briefly list the annotation (with examples) for each attribute. Some objects may have all zero scores if none of these attributes are apparent. Figure 3 demonstrates sample objects with or without semantic attributes.

Dataset

We collected a large Object and Semantic Images and Eye-tracking (OSIE) dataset with eye-tracking data from 15 participants for a full set of 700 images. Each image was manually segmented into a collection of objects on which semantic attributes were manually labeled. The images, eye-tracking data, labels, and Matlab code for data analysis are publicly available with the paper.

Compared with several datasets that are publicly available, the main motivation of our new dataset is for object and saliency study where two major contributions are: first, while existing datasets do not have ground truth data relating to objects or semantic information, we, for the first time, provide large-scale ground truth data of 5,551 object segmentation with fine contours, and semantic attribute scores of these objects. Second, we make the image contents more suitable for statistical analysis of different object and

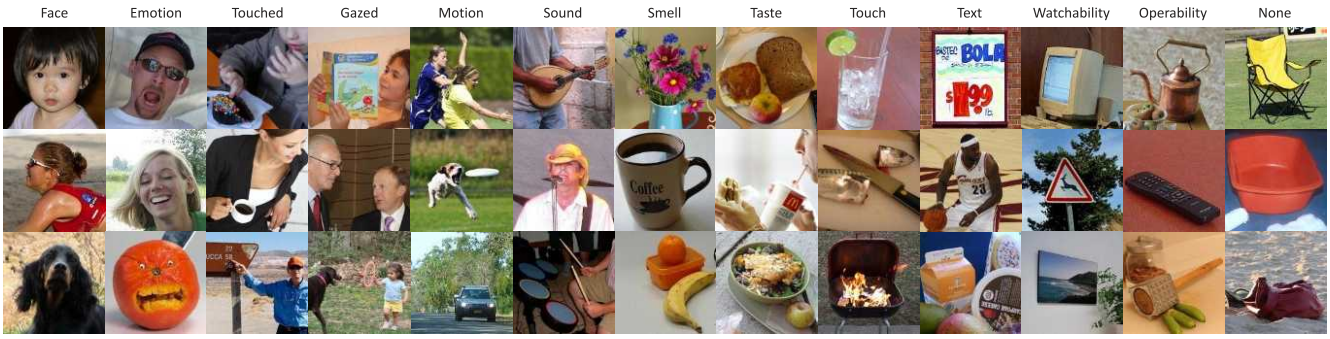

Figure 3. Example images illustrating semantic attributes. Each column is a list of objects with each semantic attribute and the last column shows sample objects without any defined semantic attributes.

Database	MIT (Judd et al., 2009)	FIFA (Cerf et al., 2009)	Toronto (Bruce & Tsotsos, 2009)	NUSEF (Ramanathan et al., 2010)	OSIE
Images	1,003	200	120	758	700
Resolution	1024 × (405 − 1024)	1024 × 768	681 × 511	1024 × 728	800 × 600
Viewers per image	15	8	11	25.3 (75 subjects each viewing a random set of 400 images)	15
Viewing time per image	3 s	2 s	4 s	5 s	3 s
Theme / distinguishing features	Every day scenes	Images with faces	Indoor and outdoor scenes	Affective objects, e.g., expressive faces, nudes, unpleasant concepts, and interactive actions	Every day scenes, many object categories with semantic meanings, multiple dominant objects per image
Ground truth annotation	None	Location of faces	None	ROIs, foreground segmentation for some objects (one object per image and 54 images), valence and arousal scores, text captions	Object segmentation with fine contours for all objects (5,551) and semantic attribute labels for all objects

Table 2. Comparison with other eye-tracking datasets.

semantic attributes by including multiple dominant objects in each image. This way by analyzing where fixations landed, statistical conclusions can be derived as to which objects/attributes attract attention. In comparison, a considerable number of images in existing datasets contain one dominant object in the center (such bias is common in photos, as human photographers place objects of interest in the center), which does not allow a direct comparison of different objects/attributes. Further, our new dataset contains a large number of object categories, including a sufficient number of objects with semantic meanings. The image contents and the labels allow quantitative analysis of object- and semantic-level attributes in driving gaze deployment. Examples of the image stimuli and eye-tracking data are illustrated in Figure 1, and Table 2 summarizes a comparison between several recent eye-tracking datasets and ours.

Experimental procedures

Fifteen subjects (undergraduate and graduate students ages 18–30 with uncorrected normal eyesight) free-viewed 700 images that comprised everyday indoor and outdoor scenes, as well as aesthetic photographs from Flickr and Google Images. These images were presented on a 22-in. LCD monitor. As subjects viewed the images, we used an Eyelink 1000 (SR Research, Osgoode, Canada) eye-tracking device to record eye movements at a sample rate of 2000 Hz. The eye-tracker system consisted of an infrared sensing camera placed alongside the computer monitor at a distance of about 26 in. from the subjects. The screen size was 47.39 × 29.62 cm (40.5° × 25.3°), with a pixel density of 90.1 ppi. The screen resolution was set to 1680 × 1050, and the 800 × 600 images were scaled to occupy the full screen height when presented on the display. Therefore, the visual angle of the stimuli was about 33.7° × 25.3°, and each degree of visual angle contained about 24 pixels in the 800 × 600 image. A chin-rest and a forehead-rest were used to stabilize the subject's head. All data were acquired from the right eyes.

In the experiments, each image was presented for 3 s and followed by a drift correction, which required subjects to fixate in the center and press the *space* key to continue. We divided the viewing into two sessions, with 300 and 400 randomly ordered images respectively, and each session was completed within 1 hr, on average two days apart. The 700 images were separated into seven blocks. Before each block, a nine-point target display was used for calibration and a second one was used for validation. After each block subjects took a 5-min. break and did a memory test: 10 images from the last 100 images and 10 new images were presented to the subjects in random order, and they

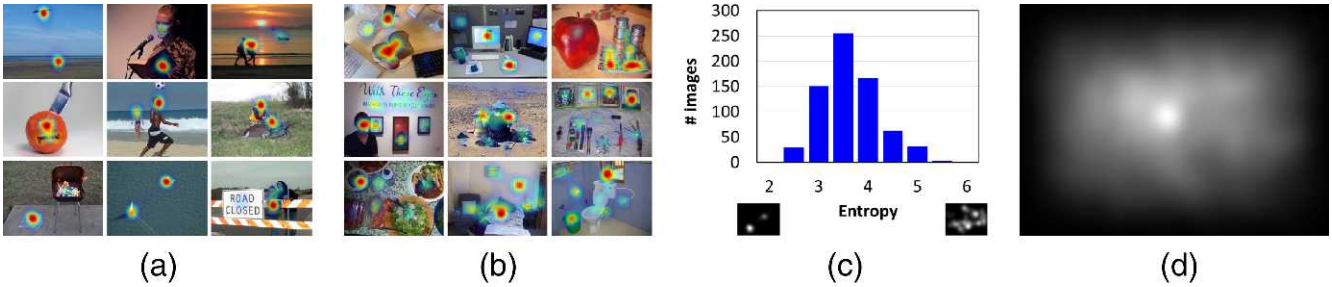

Figure 4. Human fixations with (a) lowest and (b) highest entropies in the form of heat map overlapped to the original images. Images with lower entropies tend to have fewer objects while images with higher entropies often contain several different types of objects. (c) Histogram of the fixation map entropies. (d) Average saliency map combining all fixation data, which indicates a strong bias to the center of the image.

were asked to indicate which ones they had seen before. The purpose of this memory test was to motivate subjects to pay attention to the images. To avoid task-based priming of visual attention, we did not require the subjects to memorize the contents of the presented image stimuli, but only instructed them to free-view the images. Yet there might have been a memory component in later blocks when subjects explicitly knew the subsequent memory tests. Since the test was simple enough to pass and the subjects were not motivated to pursue a high score, we believe that the memory component did not likely play a role in altering subjects' gaze patterns when viewing the images.

Statistics and analysis of the dataset

Most images in the OSIE dataset include multiple dominant objects in each image, allowing statistical comparisons of relevant importance of the attributes. In particular, among the 700 images, 682 include multiple (i.e., $\geq$2) dominant objects (i.e., dominant objects are defined to have more than 15 fixations in it).

In the experimental setup, for a saccade to be detected, the velocity threshold is 22°/s by default, which is slightly sensitive to eye-tracking noises and therefore resulted in a few short fixations (less than 100 ms in duration). These unstable fixations were discarded to reduce the noises, so the minimum duration was limited to 100 ms, while the maximum lasted about 2 s.

Consistent with previous findings (Tatler, 2007; Cerf et al., 2009; Judd et al., 2009; Zhao & Koch, 2013), our data display a center bias. Figure 4 shows the average human fixation map from all 700 images. Thirty-three percent of the fixations lie within the center 11% of the image, and 62% of fixations lie within the center 25% of the image. Compared with recent datasets where a large portion of images have one dominant object, which is commonly in the center of the image, center bias in our dataset is smaller (e.g., for the MIT dataset, 40% of fixations lie within the center 11% of the image, and 70% of fixations lie within the center 25% of the image; Judd et al., 2009). To confirm this, for both datasets, we then computed in each image the average distance (in visual angle) from all fixations to the image center, and compared them using a t test. It is shown that the distance of our dataset (7.83° $\pm$ 1.50°, mean $\pm$ SD) is significantly larger ($p < 0.01$) than that of the MIT dataset (5.76° $\pm$ 1.23°).

Psychophysical fixation maps were constructed by convolving a fovea-sized (i.e., 24 pixels in the 800 $\times$ 600 image) Gaussian kernel over the successive fixation locations of all subjects viewing the images. The entropies of the fixation maps were measured to analyze the consistency/commonality of the viewing and calculated from fixation maps resized to 200 $\times$ 150. The entropy, which is higher if the corresponding image contains more objects, is a statistical measure of randomness to characterize the fixation map of each image, defined as $S = \sum_{i=1}^{n}(-p_i log_2^{p_i})$ where the vector p represents a histogram of $n = 256$ bins. Figure 4c shows the distribution of all entropies (3.37 $\pm$ 0.57). These entropies in our dataset are significantly smaller ($p < 0.01$) than those of the MIT dataset (Judd et al., 2009; 4.00 $\pm$ 0.75), as most of the images in our dataset contain distinct objects that consistently attract human attention.

Methodology for manual object segmentation and semantic attribute labeling

Each image can be viewed as a collection of objects. In this dataset, we provided ground truth segmentation with fine object contours (5,551 objects on 700 images). In several recent eye-tracking datasets (Cerf et al., 2009; Ramanathan, Katti, Sebe, Kankanhalli, & Chua, 2010), bounding boxes around objects were labeled, but there were very few large-scale contoured object segmentations provided. The advantage of contours over bounding boxes is that it allows more accurate

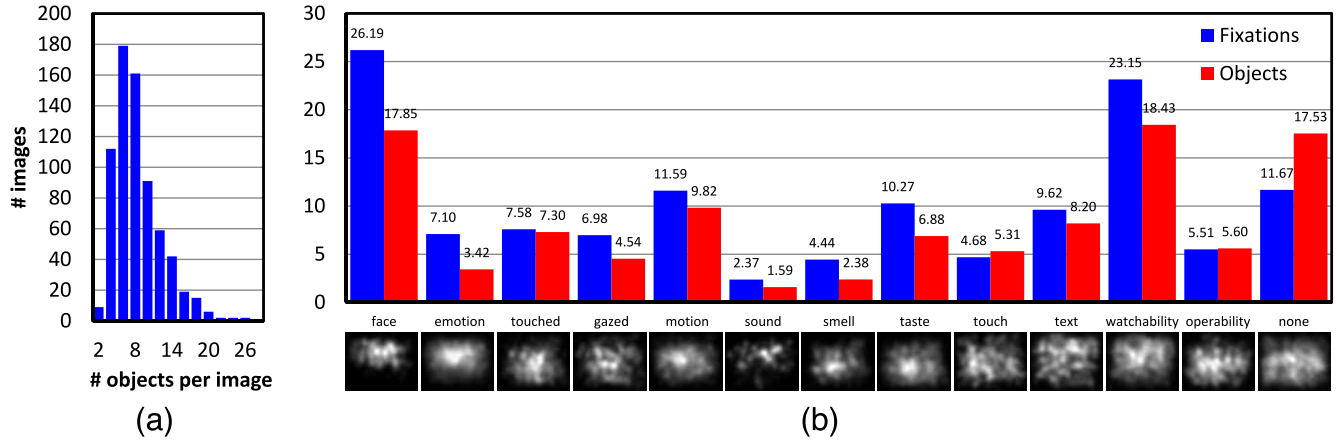

Figure 5. (a) Histogram of object numbers per image. (b) The percentages of fixations and objects labelled with each semantic attribute, along with those without any attribute (none). Below are the aggregated fixation maps for each semantic attribute (or none).

quantitative analysis. For example, fixations falling into the bounding box of the objects but not the real objects can be eliminated. Object centers that are often the focus of attention can also be more accurately estimated with fine contours. Another advantage is that some important information about saliency can only be measured by contour segmentation. For example, convexity is an important object-level attribute that describes the shape of the objects, and objects with low convexity values may indicate occluded objects. With bounding box labeling, such information would be lost.

In this work, objects in the images are first segmented with a graph cuts algorithm, using the Interactive Segmentation Tool developed by McGuinness and O'Connor (2010). Image regions without any segmented objects are regarded as the background. Since there are a large number and variety of objects in natural scenes, to make the ground truth data least dependent on subjective judgments, we followed several guidelines for the segmentation: (a) objects that are either too small or too blurry to recognize are not segmented because of their loss of semantic meaning. (b) Objects that cover a large area or hide behind the main objects in the scene (e.g., sky, ocean, ground, wall, etc.) are regarded as background and are not extracted, as humans tend to ignore the background objects. (c) Objects of the same type that are piled or clustered are grouped as one object, but similar objects at different spatial locations are not grouped. (d) All objects relating to faces (frontal, profile, and back views of human, animal, and artificial faces, etc.) and text have been shown to be salient (Cerf et al., 2009), and are explicitly defined as objects. These guidelines provided a baseline for a more objective labeling process, and they generally worked well in practice.

The distribution of the numbers of segmented objects per image are shown in Figure 5a. Semantic attributes are labeled on the objects with scores, as introduced in the above sections. The segmentation and labeling was done by paid subjects. We recruited 10 subjects who had experience in image editing to label the images. Each subject was randomly assigned a subset of the images (70 out of a total 700). The subjects were instructed to extract all foreground objects by labeling the fine object contours. We did not make assumptions as to which factors are more important to saliency to make sure the labeling was not biased. To increase cross-subject consistency, before labeling, we showed subjects several examples including humans, animals, vehicles, text, and tools as guidelines for labeling, and trained them to use the segmentation tools to label the contours. The ways to handle special cases like composite objects, occluded objects, and grouped objects were also demonstrated to the subjects.

Figure 5b summarizes the percentages of objects and their corresponding fixations with each of the semantic attributes. Note that all pixel- and object-level attributes can be automatically calculated for each object, but each object only has some (or even none, like a piece of stone or an empty table) of the semantic attributes. In total, there are 86,768 fixations on the labeled objects. As seen in Figure 5b, 17.53% of them are on objects without semantic labels, while more than a quarter of these fixations are on faces. We have also plotted in the same figure the fixation map for each attribute (including one for no attribute). It can be observed that the center bias effects in these maps are slightly different; for example, fixations on faces are highly centered in the upper region of the screen.

Experimental results

This section reports statistical analysis and computational experimental results on features, fixation

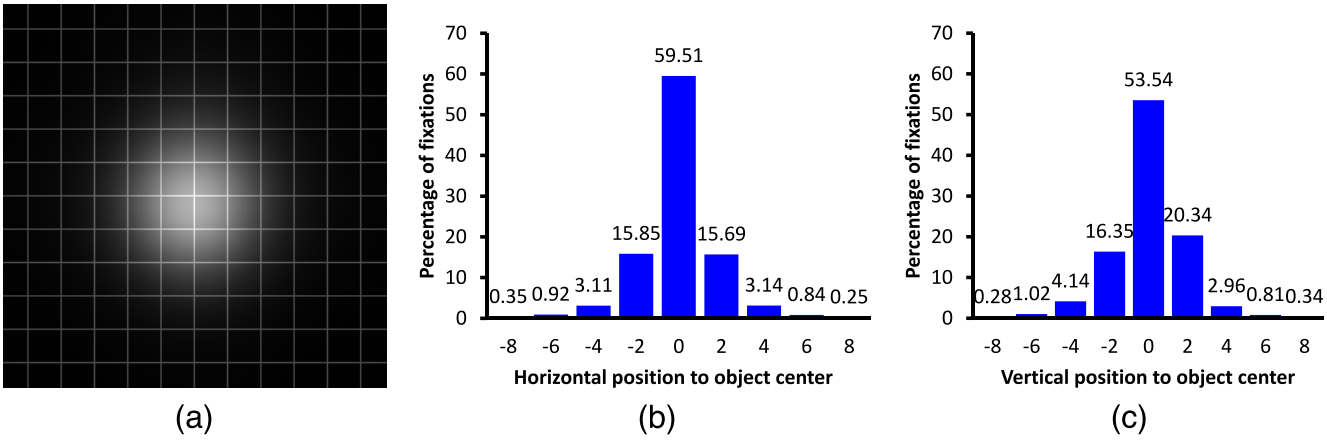

Figure 6. (a) Fixations are object-centered. The grid interval is 1°. (b) Horizontal and (c) vertical distribution of fixations.

distributions, and saliency models. We first discuss an observed "object center bias," that is, humans tend to look at the centers of the objects, despite their semantic meanings. This bias is unique to objects and thus coupled with object- and semantic-level attributes. Secondly, statistical analysis of the proposed semantic attributes is carried out to quantitatively show the validity of each one. Third, across the three layers, we learn their relative importance in driving gaze allocation. Further analysis is performed on semantic attributes to investigate how fast they attract attention. Lastly, to demonstrate the importance of such object- and semantic- level information, we construct computational models and perform comparisons with different combinations of attributes in predicting saliency. Comparisons with several other recent saliency models are also included.

Object center bias

For statistical analysis on how an object attracts attention, we first matched each fixation to a single object or the background by comparing its location against each object. If a fixation was inside an object, or its distance to the object boundary was less than a threshold, it was identified as a possible match. If a fixation had multiple possible matches, the nearest object (i.e., the one whose center location was the closest to the fixation) was chosen. The rest of the fixations were matched to the background.

To analyze how the fixations are biased towards the object centers, we plotted all fixations in an object-centered coordinate system, where all object centers are translated to the origin. All fixations added together to form a summed fixation map centered in the origin. As shown in Figure 6, the spatial distribution of the fixations in the object-centered coordinate system can be approximated as a two-dimensional (2-D) normal distribution $N(\mu, \Sigma)$, where μ is the average fixation location in the object-centered coordinate system, and $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$. Particularly, in our dataset, $\mu = (-0.02, 0.05)$ and $\sigma = (1.86, 1.90)$, which means 82.32% fixations were within a 2° visual field in the horizontal direction to the object center, while 79.45% were in the vertical direction. These statistics agree with the finding that most fixations tend to fall around the centers of objects (Nuthmann & Henderson, 2010).

While the bias toward the image centers is attributed to a variety of reasons like the experimental setup and strategic factors (Tatler, Baddeley, & Gilchrist, 2005; Zhao & Koch, 2011), the bias toward object centers relates largely to strategic advantages (i.e., center regions of objects generally contain more information about the objects).

Analysis on each semantic attribute in saliency

Is each defined semantic attribute valid and reasonable? To answer this question, we next quantified how fixations are attracted to objects with defined semantic attributes compared to those without any defined attributes. We expected that objects with defined semantic attributes attract significantly more fixations than those without defined attributes, thus indicating that the defined semantic attributes are reasonable and valid.

We categorized the semantic attribute of each fixation as it mapped onto an object. To analyze the validity of a particular attribute, we constrained the analysis to fixations from objects with only one attribute. For example, to analyze the impact of the "taste" attribute, all fixations were collected from objects that only had the label of "taste." Note that one exception of this procedure was for the "face" and "emotion" attributes due to their tight correlation—

Semantic attribute	Number of fixations	*t*(*df*)	*P*
Face without emotion	16,591	125.8673	<0.0001
Face with emotion	5,148	126.0089	<0.0001
Touched	2,170	26.1691	<0.0001
Gazed	528	37.6065	<0.0001
Motion	8,047	25.9506	<0.0001
Sound	63	−0.8475	0.8016
Smell	288	−0.3652	0.6425
Taste	5,046	15.5250	<0.0001
Touch	2,592	0.9458	0.1721
Text	10,375	81.8678	<0.0001
Watchability	6,858	45.0752	<0.0001
Operability	1,998	−1.5488	0.9393
None (control group)	10,815		

Table 3. The *t* test results on the fixation densities of each semantic attribute.

each "emotion" label is on a "face." To make each attribute in this analysis independent, the "face" group is split into "face with emotion" and "face without emotion." We subsequently compared these fixations to a control group of fixations that are from objects that have no defined semantic attributes. Fixations were randomly and independently sampled and their saliency values from the corresponding saliency maps (i.e., ground truth fixation density maps from human data) were compared using a one-tailed *t* test (see Table 3). The false positive rate was set at 0.05/12 (Bonferroni correction for 12 comparisons in total; Bland & Altman, 1995). We found that the mean saliency for most semantic attributes was significantly larger than that of the control group, with the exception of "sound," "smell," "touch," and "operability." Our data suggest that our defined semantic attributes are valid and reasonable and have positive impacts on objects' saliency.

Analysis on the relative attribute importance in saliency

We used a support vector machine (SVM) classification to analyze the proposed attributes and train the saliency model directly from human eye-tracking data (see Figure 7). For each image, we precomputed the feature maps for every pixel of the image resized to 200 × 150 and used the maps to train our model. Figure 8 shows the feature maps computed for a sample image. The pixel-level feature maps were generated with Itti et al.'s (1998) algorithm, while the object- and semantic-level feature maps were generated by placing a 2-D Gaussian kernel at each object's center, which models the object center bias effect that we discussed above. The Gaussian bandwidth approximates the standard deviation of the object center bias discussed above, which is 2° visual angle and 48 pixels in the images. The Gaussian kernel generally falls within the object region, and the magnitude of the Gaussian is the calculated object-level or manually labeled semantic-level feature value.

To train and test this model, we divided our dataset into 500 training images and 200 testing images. From the ground truth fixation map of each image, 20 pixels were randomly sampled from the top 20% salient locations, and 60 pixels were sampled from the bottom 60% salient locations, yielding a training set of 10,000 positive samples and 30,000 negative samples. The use of a small coverage of salient regions and a relatively larger nonsalient area is the consideration of the interobserver congruency. That is, we chose only

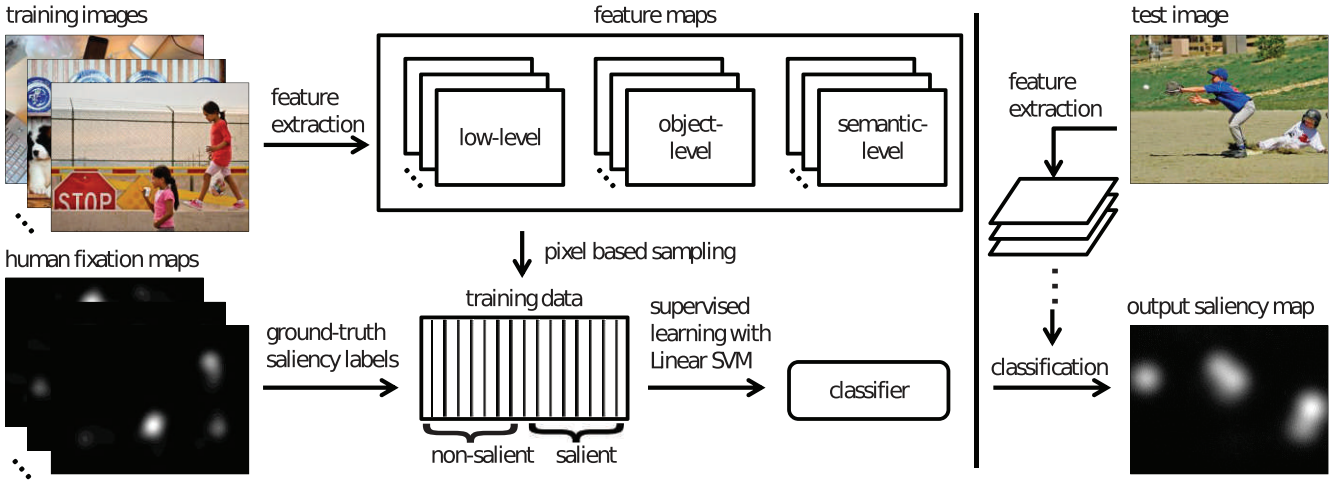

Figure 7. An overview of the computational saliency model. The three levels of features are extracted from the input images. We use a pixel-based random sampling to collect the training data and train a linear SVM classifier with the relative attribute importance. Given a test image, the feature maps are linearly combined using the trained classifier to generate the saliency map.

Figure 8. An example of the pixel-, object-, and semantic-level feature maps. The fixation map of the image is shown in the top-left corner.

regions fixated by multiple subjects as salient regions, while leaving a large portion of the image as background where fewer fixations occur. This method is also consistent with the implementation in the MIT model (Judd et al., 2009). The purpose of choosing a 1 : 3 sampling ratio is to balance the distributions of positive and negative sample pixels in the same image, since a large portion of the less salient region is the background where no object or semantic attributes are sampled. The training samples were normalized to have zero mean and unit variance. The same parameters were used to normalize the test set.

A linear SVM (Fan, Chang, Hsieh, Wang, & Lin, 2008) was first used to learn the weight of each pixel-, object-, and semantic-level attribute in determining its importance in attention allocation. The use of a linear integration method is motivated by the neuronal process mechanism of visual information. Linear SVM is also faster to compute, and the resulting weights of attributes are intuitive to understand—we also have tested logistic and LASSO type algorithms for the same purpose but have not found advantages in our specific tasks; therefore, an L2-regularized L2-loss SVM classification was applied and the misclassification cost c was set to 1. The learned weight of each attribute is shown in Figure 9a. For semantic attributes, consistent with previous findings (Cerf et al., 2009), face and text outweighed other attributes, followed by gazed, taste,

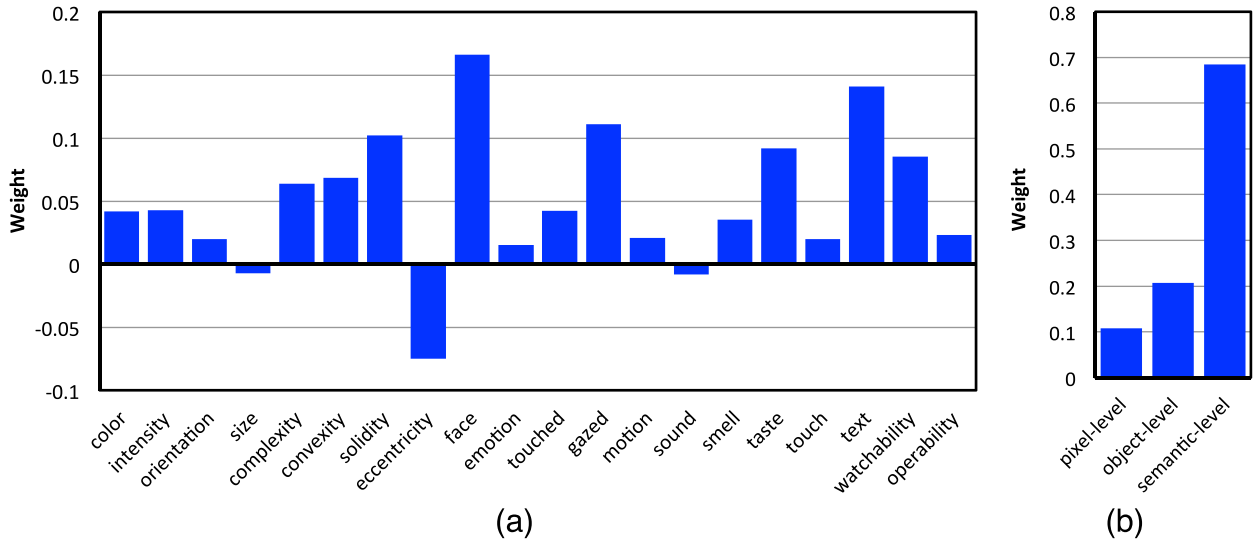

Figure 9. (a) The learned weights of all attributes. Face far outweighs other semantic attributes, followed by text, gaze, and taste. (b) The importance of three levels of attributes.

and watchability. The face channel weighed the highest, largely attributed to the dedicated pathways on the human and primate visual systems to process faces. The high weight of the "gazed" channel shows the effect of a joint attention. Viewers readily detect the focus of attention from other people's eye gaze, and orient their own to the same location (Friesen & Kingstone, 1998; Nummenmaa & Calder, 2009). The weights of object-level attributes also agree with previous finding in figure-ground perception: that smaller, more convex regions tend to be in the foreground (Fowlkes et al., 2007). A complex shape contains more information, so it is also more salient than a simple one. The weight of eccentricity shows that longer shapes are less salient than round blob-like ones.

We further compared the overall weights of the pixel-, object- and semantic-levels, by combining feature maps within each level into an intermediate saliency map of that particular level using the previously learned weights, and performed a second pass learning using the three intermediate maps. The learned weights of each level were 0.11, 0.21, and 0.68 for pixel, object, and semantic information, respectively, suggesting that semantic-level attributes attract attention most strongly, followed by object-level attributes.

To further investigate the nature of the pixel-, object-, and semantic-level attributes in driving gaze, consistent with the time-dependent model of Gautier and Le Meur (2012), we calculated attributed weights as a function of fixation (i.e., computed weights using the first N fixations from all subjects) and compared the weights over time.

For a number of attributes, a clear decreasing/increasing trend can be observed, suggesting that some attract attention faster than others. Specifically, three types of trends can be seen: (a) the weight decreases over time—when the training data include only the first fixations from all subjects, the weights of all pixel-level attributes, two object-level attributes (size and eccentricity), and three semantic-level attributes (face, emotion, and motion) are the largest, and they decrease monotonically as more fixations per image per subject are used (as shown in Figure 10a, 10b and 10c). It suggests that these attributes attract attention rapidly, especially for the face and emotion channels—which may be due to the fact that humans have a dedicated face region and pathway to process face-related information. (b) As shown in Figure 10e, the weights of text, sound, touch, touched, and gazed increase as viewing proceeds, indicating that although some of the attributes attract attention, they are not as rapid. (c) The weights of other semantic attributes including smell, taste, operability, and watchability do not show apparent trend over time, as illustrated in Figure 10f. The fact that attribute weights are time-dependent seems quite interesting, which enables us to predict the fixation order and the scanpath across the viewing time. In this work, the saliency prediction results are mostly computed based on all fixations in the viewing time (i.e., 3 s), to be directly comparable with other models in the state-of-the-art, time-dependent model similar to that of Zhao and Koch, 2011, with the proposed attributes considered as future work.

Quantitative and qualitative comparisons of computational saliency models

We performed quantitative and qualitative comparisons of our models with different combinations of attributes, as well as comparisons with several other recent saliency models. Particularly the comparison models included the MIT model (Judd et al., 2009), the Graph-Based Visual Saliency (GBVS) model (Harel, Koch, & Perona, 2007), the GBVS combined with a face detector (GBVS+VJ; Cerf et al., 2009), the Image Signature model by Hou et al. (2012), the Attention based on Information Maximization (AIM) model (Bruce & Tsotsos, 2009), the SUN bottom-up model (Zhang, Tong, Marks, Shan, & Cottrell, 2008), and the Itti et al. (1998) model. An ROC analysis is shown in Figure 11a. Our saliency models were generated by a weighted linear combination of the feature maps using the learned weights of each attribute. We also evaluated the performance of linear combination with uniform weights (UW), where all attributes were assumed to equally contribute to the saliency prediction. The ROC curve was plotted by varying the saliency percentage to cover all possible ranges of values the saliency map predicts.

Figure 11b shows the area under the ROC curve (AUC) for each model. We normalized the AUC values by an "ideal AUC" (Cerf et al., 2009), which measures how well the fixations of each subject can be predicted by those of the other $n - 1$ subjects. The computation was done by iterating over all n subjects and averaging the AUC scores of all the predictions. It reflects the performance of humans and serves as an upper bound to the performance of a computational model. In the comparison we use the same parameter for blurring for all models in this experiment, which approximates 1° of the visual field. In addition, the MIT model is trained on the same training set as our method, without the original "distance to center" channel, for a fair comparison.

From Figure 11, we make the following key observations: (a) To obtain a better performance, we can add semantic-level information to models with pixel-level information only. Further, the richer and the more complete the semantic contents, the better the performance—our model with 12 base semantic attri-

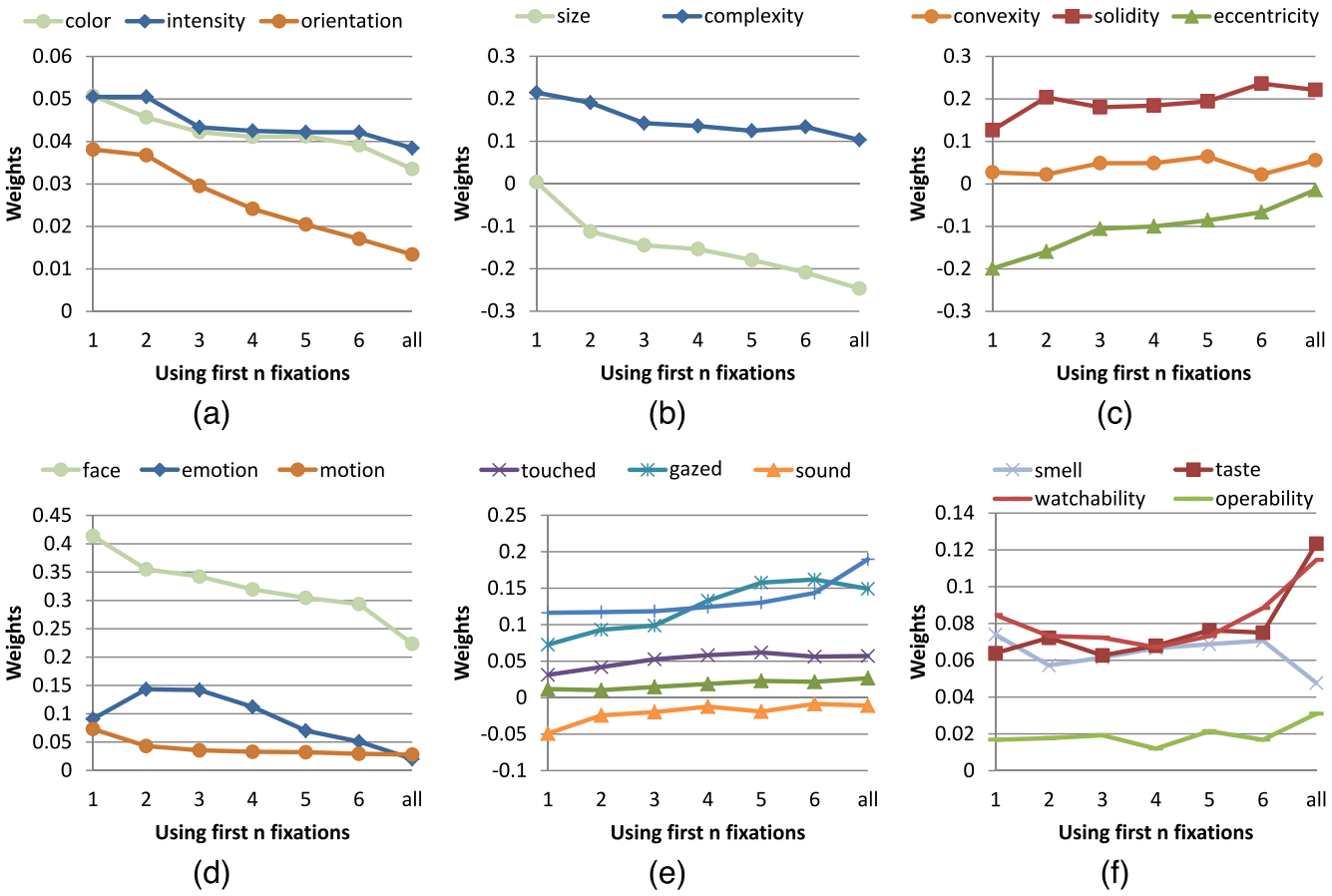

Figure 10. Optimal weights with respect to viewing time for pixel-, object-, and semantic- level attributes. (a) The weights of the pixel-level attributes decrease consistently over time. (b) Object-level attributes whose weights decrease over time. (c) Object-level attribute whose weights increase over time. (d) Semantic-level attributes whose weights decrease over time attract attention rapidly. This is particular to face related information, in consistent with the fact that face has its dedicated processing region and pathway in human brains. (e) Semantic-level attributes whose weights increase over time attract attention not as rapidly. (f) Semantic-level attributes whose weights do not show an obvious trend over time.

butes performs better than the GBVS+VJ (Cerf et al., 2009) and MIT (Judd et al., 2009) models that include only one to three sample object categories. (b) Object-level information is also important in saliency. Without semantic attributes, our model with pixel- and object-level attributes performs better than other models (Harel et al., 2007; Zhang et al., 2008; Bruce & Tsotsos, 2009; Cerf et al., 2009; Judd et al., 2009; Hou et al., 2012). (c) Our model with pixel-level information outperforms the classic Itti et al. (1998) model, despite the same attributes used, indicating that different attributes contribute differently to saliency, and taking it into account improves saliency prediction.

For a qualitative assessment, maps of our object saliency model and the compared models are demonstrated in Figure 12. First, our model predicts semantically meaningful objects (e.g., faces, texts) to be more salient than other objects and the background. These examples show that compared to the uniform weighting, the weights learned from eye data lead to more accurate predictions that differentiate the most salient objects from the least salient ones. Second, the proposed method scales well to a large number of categories in real life. While other models, including a couple of detectors, accurately predict the encoded categories as salient (e.g., face detection in GBVS+VJ; Cerf et al., 2009), our model predicts general objects (e.g., the black cat in Figure 12e) reasonably well without the incorporation of any object detectors. Third, within an object, the center regions are highlighted in our saliency maps consistent with human behaviors. In comparison, in saliency maps based on pixel-level attributes only, object boundaries are usually predicted to be more salient due to higher pixel-level contrast. One limitation of the current model is its degenerated performance on crowded scenes with multiple objects of the same category (e.g., the keyboard in Figure 12g and the text in Figure 12h). It is partially due to the difficulty in deciding whether to group objects together or consider them as individual

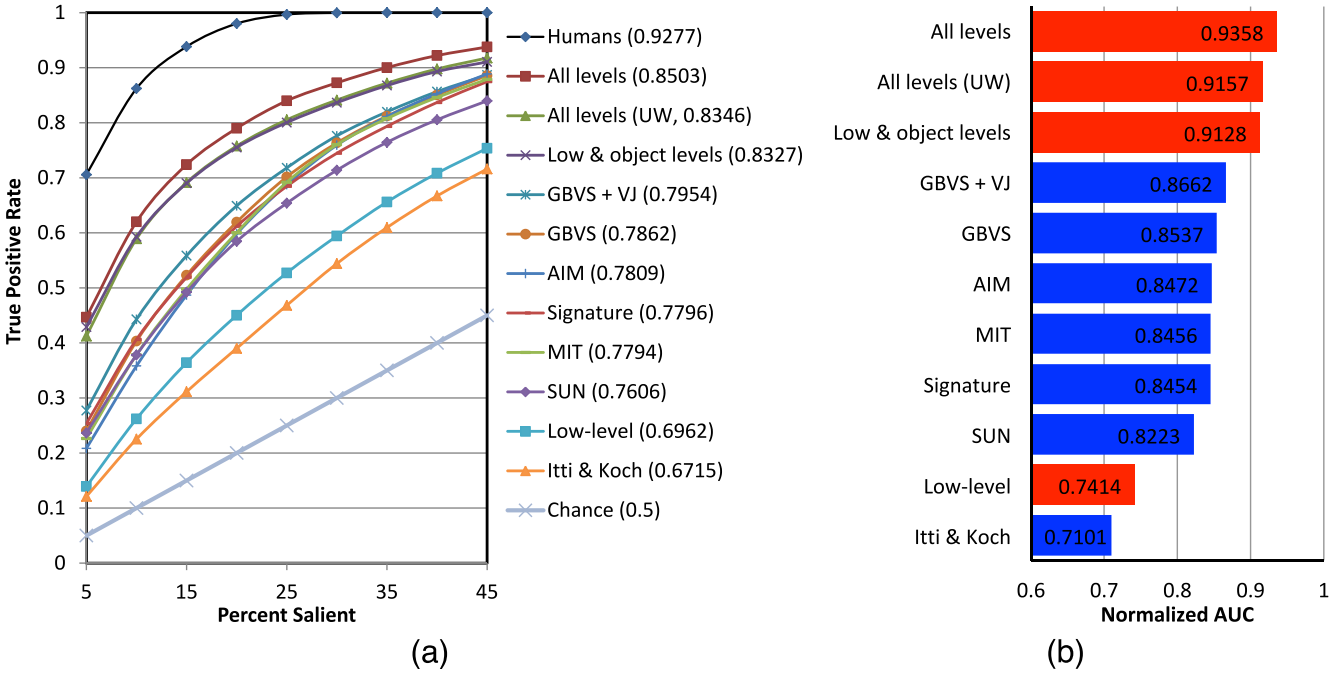

Figure 11. (a) The ROC curves and the raw AUC values (in parentheses) of models trained with different sets of attributes compared with other saliency models, as well as human and chance. For a fair comparison, the MIT model is trained on the same training set as our method, without the original "distance to center" channel. (b) The normalized AUC values of each model. Note that the normalized AUC of a model is not obtained by a direct division of its raw AUC by that of human performance; instead it is calculated on each single test image first and then averaged to get the normalized AUC value of the model.

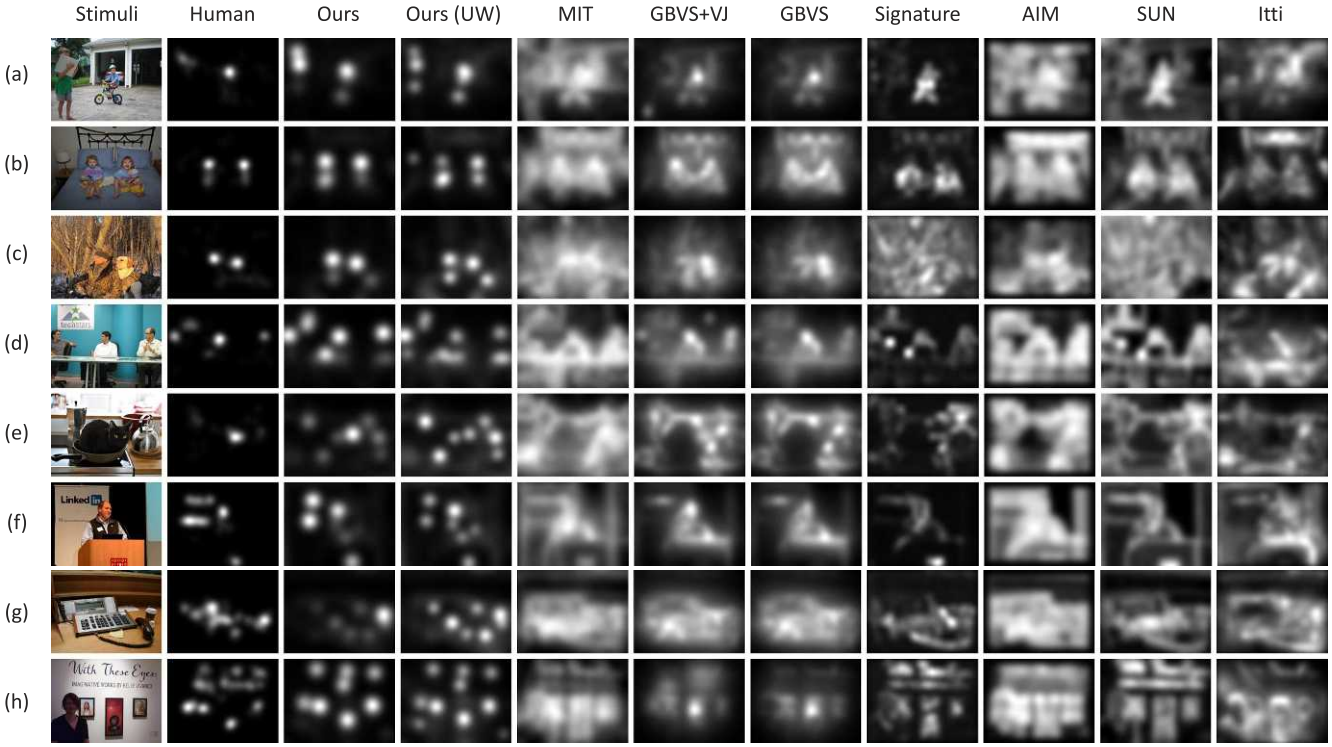

Figure 12. The qualitative results generated by the proposed saliency models in comparison with the state-of-the-art. UW = Uniform Weighting of all attributes.

objects, and possibly due to the more sophisticated strategies of humans to fixate on some of the objects instead of others, despite the objects being of the same type. Saliency in crowded environments is also an interesting topic worthy of future investigations.

Discussions

From the analysis results, although the proposed saliency model has been built upon the common and natural free-viewing task to avoid top-down biases, semantic attributes (e.g., face, text, gazed, etc.) still contribute more than lower-level ones to the allocation of visual attention, which agrees with previous studies in various aspects (Friesen & Kingstone, 1998; Vuilleumier, 2000; Ro et al., 2001; Bindemann et al., 2005; Bindemann et al., 2007; Onat et al., 2007; Cerf et al., 2009; Schirmer et al., 2011). The object-level attributes proposed in this work are also shown to be strongly correlated with attention selection, consistent with several related works (Craft et al., 2007; Einhäuser et al., 2008; Mihalas et al., 2010; Nuthmann & Henderson, 2010).

The use of a task-free paradigm with a 3-s viewing period is in line with various studies modeling saliency in the allocation of visual attention (e.g., Parkhurst, Law, & Niebur, 2002; Açik, Onat, Schumann, Einhäuser, & König, 2009, in which a 5-s free-viewing paradigm is applied). Several recent datasets (Bruce & Tsotsos, 2009; Cerf et al., 2009; Judd et al., 2009; Ramanathan et al., 2010) all set the free-viewing time to 2–5 s per image. In our paradigm, the 3-s design is mostly motivated by the following factors: The duration provided sufficient time to sample various locations and objects in a natural image. If the viewing duration is too short, subjects might not have enough time to sample locations or objects that are also important, especially with the presence of a center bias. On the other hand, if the viewing duration is too long, as the viewing proceeded, top-down or other factors (for example, subjects feel bored, tired, or distracted) come into play and fixations become noisier. Further, to view 700 images, this viewing duration makes the total experimental time feasible in practice. It has been suggested to use a task-dependent paradigm with variable viewing durations to minimize psychological expectations and reduce unwanted top-down strategies (Tatler, Baddeley, & Gilchrist, 2005). However, there might be an interactive effect of the number of interesting objects in a scene and the viewing duration. The viewing strategy might be influenced by the top-down instruction and thus the viewing might become unnatural to reveal pure bottom-up saliency.

There has been a debate on picture-viewing paradigms and saliency-based schemes in modeling gaze allocation in scene viewing. Tatler, Hayhoe, Land, and Ballard (2011) argued that models built from the simple free-viewing paradigm (i.e., subjects view static scenes for a few seconds in laboratory settings) are difficult to be generalized to natural behavior. We agree that one major issue of the purely bottom-up saliency model is the lack of real-world tasks. Indeed, top-down influences like experience, reward, and contextual priors should be taken into account for a more complete model in complex scene viewing like in the natural settings. In this work, modeling object and semantic attributes in the data-driven framework is an attempt to learn the task-free object viewing experiences of humans. The framework can also adapt to accommodate other top-down influences by including a set of task-relevant attribute weights. In other words, even when subjects follow the task instruction to search for targets, our saliency model is still able to predict fixations. In comparison with previous top-down models—for example, the computational model proposed by Wischnewski, Belardinelli, Schneider, and Steil, 2010, which combines proto-objects and top-down tasks with bottom-up saliency—our model focuses more on the common and task-free attributes of the objects, for example, their semantic meanings. As suggested by the pedestrian searching model (Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009) and the SUN top-down model (Kanan, Tong, Zhang, & Cottrell, 2009), the target-related context guidance (Torralba, Oliva, Castelhano, & Henderson, 2006) that guides attention to the locations that an object is likely to appear could be an useful extension in visual search tasks. The weighted linear combination could also be replaced with a weighted product method, which seems to be more adequate at predicting the overall fixation distribution in visual search tasks (Hwang, Higgins, & Pomplun, 2009). Recently, Hwang, Wang, and Pomplun (2011) investigated the influence of semantic similarity among scene objects on eye movements in visual search. At the core of their work is a high dimensional "semantic space" from the text corpus, and thus the similarity of each pair of words can be calculated as the cosine value of the angle between the two corresponding vectors in the space. Their semantic relations are formed at a conceptual level rather than a visual level, which has been pointed out by the authors as a limitation of the work, as the latter is a practically difficult problem. The proposed work naturally approaches the problem as the modeling of the small set of semantic attributes at a visual level is much more feasible than the original intractable set of semantic entities. The attribute-based framework is thus able to scale well and characterize a wide variety of semantic

objects, without the requirement of text labels, as did in Hwang et al. (2011).

Conclusions

Recent neurophysiological and psychophysics experiments have suggested the importance of object- and semantic-level information in visual perception. To fill the semantic gap between the saliency models and human behavioral data, we propose a three-layered architecture for saliency modeling, and for the first time, we have explicitly and principally modeled saliency at the object and semantic levels. We have constructed a vocabulary at three levels to capture inherent mechanisms in gaze allocation and learn their relevant importance in saliency. By combining the set of attributes we are able to describe any object categories, therefore overcoming the current problem with adding limited number (usually $\leq$3) of object detectors into saliency models, which does not scale well in the real world. To validate our proposed framework and for future research on object and semantic saliency in the community, a large eye-tracking dataset with 700 images and eye-tracking data with 15 viewers has been constructed and is publicly available with the paper. In the dataset we have also for the first time provided large-scale object segmentation with fine contours (5,551 objects) and annotation of 12 semantic attributes for all the objects. Experiments demonstrate the importance of object and semantic information in predicting human gaze.

Keywords: visual saliency, saliency attribute, object saliency, semantic saliency, dataset, computational model

Acknowledgments

The authors would like to thank Dr. Christof Koch and members of the Koch Lab at Caltech for valuable comments. This research was partially supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO, and the Singapore Ministry of Education Academic Research Fund Tier 1 (No.R-263-000-648-133).

Commercial relationships: none.
Corresponding author: Qi Zhao.
Email: eleqiz@nus.edu.sg.
Address: Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

References

Açik, A., Onat, S., Schumann, F., Einhäuser, W., & König, P. (2009). Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories. *Vision Research*, *49*(12), 1541–1553.

Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences*, *1191*(1), 42–61.

Argyle, M., Ingham, R., Alkema, F., & McCallin, M. (1973). The different functions of gaze. *Semiotica*, *7*(1), 19–32.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. *Vision Research*, *46*(18), 2824–2833.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2003). Fmri responses to video and point-light displays of moving humans and manipulable objects. *Journal of Cognitive Neuroscience*, *15*(7), 991–1001.

Bindemann, M., Burton, A. M., Hooge, I. T., Jenkins, R., & de Haan, E. H. (2005). Faces retain attention. *Psychonomic Bulletin & Review*, *12*(6), 1048–1053.

Bindemann, M., Burton, A. M., Langton, S. R., Schweinberger, S. R., & Doherty, M. J. (2007). The control of attention to faces. *Journal of Vision*, *7*(10):15, 1–8, http://www.journalofvision.org/content/7/10/15, doi:10.1167/7.10.15. [PubMed] [Article]

Bland, J., & Altman, D. (1995). Multiple significance tests: The bonferroni method. *BMJ*, *310*(6973), 170.

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3):5, 1–24, http://www.journalofvision.org/content/9/3/5, doi:10.1167/9.3.5. [PubMed] [Article]

Carbone, A., & Pirri, F. (2010). Learning saliency. An ICA based model using Bernoulli mixtures. In *Proceedings of brain inspired cognitive systems*.

Cerf, M., Frady, E., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12):10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article]

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913–919.

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010).

What and where: A bayesian inference theory of attention. *Vision Research*, *50*(22), 2233–2247.

Craft, E., Schütze, H., Niebur, E., & Von Der Heydt, R. (2007). A neural model of figure–ground organization. *Journal of Neurophysiology*, *97*(6), 4310–4326.

Cree, G., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of *chipmunk, cherry, chisel, cheese,* and *cello* (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, *298*(5596), 1191–1194, doi:10.1126/science.1076358.

Donk, M., & van Zoest, W. (2008). Effects of salience are short-lived. *Psychological Science*, *19*(7), 733–739.

Edelman, G. (1987). *Neural darwinism: The theory of neuronal group selection*. New York: Basic Books.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6/7), 945–978.

Einhäuser, W., Rutishauser, U., Frady, E., Nadler, S., König, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision*, *6*(11):1, 1–11, http://www.journalofvision.org/content/6/11/1, doi:10.1167/6.11.1. [PubMed] [Article]

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14):18, 1–26, http://www.journalofvision.org/content/8/14/18, doi:10.1167/8.14.18. [PubMed] [Article]

Engmann, S., Hart, B., Sieren, T., Onat, S., König, P., & Einhäuser, W. (2009). Saliency on a natural scene background: Effects of color and luminance contrast add linearly. *Attention, Perception, & Psychophysics*, *71*(6), 1337–1352.

Faivre, N., & Koch, C. (2013). Integrating information from invisible signals: The case of implied motion. *Journal of Vision*, *13*(9):962, http://www.journalofvision.org/content/13.9.962, doi:10.1167/13.9.962. [Abstract]

Fan, R., Chang, K., Hsieh, C., Wang, X., & Lin, C. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874.

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1778–1785). Washington, DC: IEEE Computer Society.

Foulsham, T., & Kingstone, A. (2013). Optimal and preferred eye landing positions in objects and scenes. *Quarterly Journal of Experimental Psychology*, *66*(9), 1707–1728.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2):6, 1–17, http://www.journalofvision.org/content/8/2/6, doi:10.1167/8.2.6. [PubMed] [Article]

Fowlkes, C., Martin, D., & Malik, J. (2007). Local figure–ground cues are valid for natural images. *Journal of Vision*, *7*(8):2, 1–9, http://www.journalofvision.org/content/7/8/2, doi:10.1167/7.8.2. [PubMed] [Article]

Friesen, C., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*(3), 490–495.

Friston, K., Tononi, G., Reeke, G., Sporns, O., & Edelman, G. (1994). Value-dependent selection in the brain: Simulation in a synthetic neural model. *Neuroscience*, *59*(2), 229–243.

Gao, D., Mahadevan, V., & Vasconcelos, N. (2007). The discriminant center-surround hypothesis for bottom-up saliency. *Advances in Neural Information Processing Systems*, *19*, 497–504.

Garrard, P., Ralph, M., Hodges, J., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, *18*(2), 125–174.

Gautier, J., & Le Meur, O. (2012). A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions. *Cognitive Computation*, *4*(2), 141–156.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, *19*, 545–552.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, *3*(1), 80–84.

Hooker, C. L., Palier, K. A., Gitelman, D. R., Parrish, T. B., Mesulam, M.-M., & Reber, P. J. (2003). Brain networks for analyzing eye gaze. *Cognitive Brain Research*, *17*(2), 406–418.

Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transac-*

tions on Pattern Analysis and Machine Intelligence, *34*(1), 194–201.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, *9*(5):25, 1–18, http://www.journalofvision.org/content/9/5/25, doi:10.1167/9.5.25. [PubMed] [Article]

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*(10), 1192–1205.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, *18*, 547–554.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jost, T., Ouerhani, N., von Wartburg, R., Muri, R., & Hugli, H. (2005). Assessing the contribution of color in visual attention. *Computer Vision & Image Understanding*, *100*(1–2), 107–123.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th International Conference on Computer Vision,* 2009 (pp. 2106–2113). Washington, DC: IEEE Computer Society.

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). Sun: Top-down saliency using natural statistics. *Visual Cognition*, *17*(6–7), 979–1003.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1476), 2109–2128.

Kling, A. S., & Brothers, L. A. (1992). The amygdala and social behavior. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion* (pp. 353–377). New York: Wiley.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiololgy*, *4*(4), 219–27.

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human mt/mst by static images with implied motion. *Journal of Cognitive Neuroscience*, *12*(1), 48–55.

Krieger, G., Rentschler, L., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, *73*(2–3), 201–214.

Lorteije, J. A., Kenemans, J. L., Jellema, T., Van Der Lubbe, R. H., De Heer, F., & Van Wezel, R. J. (2006). Delayed response to animate implied motion in human motion processing areas. *Journal of Cognitive Neuroscience*, *18*(2), 158–168.

Masciocchi, C., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision*, *9*(11):25, 1–22, http://www.journalofvision.org/content/9/11/25, doi:10.1167/9.11.25. [PubMed] [Article]

McGuinness, K., & O'Connor, N. (2010). A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, *43*(2), 433–444.

Mihalas, S., Dong, Y., Von Der Heydt, R., & Niebur, E. (2010). Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Journal of Vision*, *10*(7):979, http://www.journalofvision.org/content/10/7/979, doi:10.1167/10.7.979. [Abstract]

Milanese, R. (1993). *Detecting salient regions in an image: from biological evidence to computer implementation*. (Unpublished doctoral dissertation) University of Geneva, Switzerland.

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: A unified system for processing faces in the macaque temporal lobe. *Science*, *320*(5881), 1355–1359.

Nummenmaa, L., & Calder, A. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, *13*(3), 135–143.

Nuthmann, A., & Henderson, J. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8):20, 1–19, http://www.journalofvision.org/content/10/8/20, doi:10.1167/10.8.20. [PubMed] [Article]

Onat, S., Libertus, K., & König, P. (2007). Integrating audiovisual information for the control of overt attention. *Journal of Vision*, *7*(10):11, 1–16, http://www.journalofvision.org/content/7/10/11, doi:10.1167/7.10.11. [PubMed] [Article]

Palmer, S. (1999). *Vision science: Photons to phenomenology* (Vol. 1). Cambridge, MA: MIT Press.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling

the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Pelphrey, K. A., Viola, R. J., & McCarthy, G. (2004). When strangers pass processing of mutual and averted social gaze in the superior temporal sulcus. *Psychological Science*, *15*(9), 598–603.

Raj, R., Geisler, W., Frazor, R., & Bovik, A. (2005). Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *Journal of the Optical Society of America A*, *22*(10), 2039–2049.

Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T. (2010). An eye fixation database for saliency detection in images. *Computer Vision-ECCV*, *6314*, 30–43.

Reinagel, P., & Zador, A. (1999). Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, *10*(4), 341–350.

Reisfeld, D., Wolfson, H., & Yeshurun, Y. (1995). Context-free attentional operators: The generalized symmetry transform. *International Journal of Computer Vision*, *14*(2), 119–130.

Ro, T., Russell, C., & Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. *Psychological Science*, *12*(1), 94–99.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, *39*(19), 3157–3163.

Schirmer, A., Escoffier, N., Zysset, S., Koester, D., Striano, T., & Friedend, A. D. (2008). When vocal processing gets emotional: On the role of social orientation in relevance detection by the human amygdala. *Neuroimage*, *40*(3), 1402–1410.

Schirmer, A., Teh, K. S., Wang, S., Vijayakumar, R., Ching, A., Nithianantham, D., ... Cheok, A. D. (2011). Squeeze me, but don't tease me: Human and mechanical touch enhance visual attention and emotion discrimination. *Social Neuroscience*, *6*(3), 219–230.

Seo, H., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, *9*(12):15, 1–27, http://www.journalofvision.org/content/9.12.15, doi:10.1167/9.12.15. [PubMed] [Article]

Sprague, N., & Ballard, D. (2003). Eye movements for reward maximization. *Advances in Neural Information Processing Systems*, *16*, 1467–1474.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5):5, 1–23, http://www.journalofvision.org/content/11/5/5, doi:10.1167/11.5.5. [PubMed] [Article]

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, *24*(5), 295–300.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670–674.

Tsotsos, J., Culhane, S., Kei Wai, W., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, *78*(1), 507–545.

Ungerleider, S. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23*(1), 315–341.

Vuilleumier, P. (2000). Faces call for attention: Evidence from patients with visual extinction. *Neuropsychologia*, *38*(5), 693–700.

Walther, D., Serre, T., Poggio, T., & Koch, C. (2005). Modeling feature sharing between object detection and top-down attention. *Journal of Vision*, *5*(8):1041, http://www.journalofvision.org/content/5/8/1041, doi:10.1167/5.8.1041. [Abstract]

Wang, W., Wang, Y., Huang, Q., & Gao, W. (2010). Measuring visual saliency by site entropy rate. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2368–2375). Washington, DC: IEEE Computer Society.

Whalen, P. J., Kagan, J., Cook, R. G., Davis, F. C., Kim, H., Polis, S., ... Johnstone, T. (2004). Human amygdala responsivity to masked fearful eye whites. *Science*, *306*(5704), 2061–2061.

Winawer, J., Huk, A. C., & Boroditsky, L. (2008). A motion aftereffect from still photographs depicting motion. *Psychological Science*, *19*(3), 276–283.

Wischnewski, M., Belardinelli, A., Schneider, W., & Steil, J. J. (2010). Where to look next? Combining static and dynamic proto-objects in a tva-based

model of visual attention. *Cognitive Computation*, *2*(4), 326–343.

Zhang, L., Tong, M., Marks, T., Shan, H., & Cottrell, G. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7):32, 1–20, http://www.journalofvision.org/content/8/7/32, doi:10.1167/8.7.32. [PubMed] [Article]

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*(3):9, 1–15, http://www.journalofvision.org/content/11/3/9, doi:10.1167/11.3.9. [PubMed] [Article]

Zhao, Q., & Koch, C. (2012). Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. *Journal of Vision*, *12*(6):22, 1–15, http://www.journalofvision.org/content/12/6/22, doi:10.1167/12.6.22. [PubMed] [Article]

Zhao, Q., & Koch, C. (2013). Learning saliency-based visual attention: A review. *Signal Processing*, *93*(6), 1401–1407.